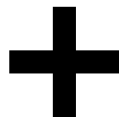

Estadística

Métodología de Análisis en
Opinión Pública – Olego 2024



Hoja de
ruta

+

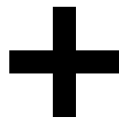
—

Módulos

1. Estadística
2. Inferencia Estadística
3. Probabilidad
4. Modelos Avanzados y sus Apps

+

¿Cuál podrían decir que es la diferencia entre la estadística, la probabilidad y la inferencia?

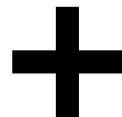


¿Por qué E., P. e I. en Opinión Pública y en ciencias sociales?

E) Te ayuda a **organizar y describir datos** existentes. Es el presente, lo que podés observar y medir.

P) Permite **medir la incertidumbre y anticipar resultados**. Es el futuro, lo que podría pasar

I) Es el puente entre **lo que sabes y lo que querés** saber sobre una **población más grande**



¿Por qué E., P. e I. en Opinión Pública y en ciencias sociales?

E) Ayuda a **procesar las respuestas** de miles de personas y **resumirlas en indicadores** clave, como el porcentaje de personas satisfechas, el promedio de satisfacción en una escala de 1 a 10, y la distribución de opiniones entre diferentes grupos sociales

P) Permite **anticipar resultados** y **tomar decisiones** bajo **incertidumbre**, como predecir la probabilidad de ciertos resultados electorales o de cambios en la opinión pública.

I) Permite hacer **generalizaciones sobre la población total** a partir de los datos de **una muestra**, y es esencial para convertir los datos de encuestas en conclusiones útiles sobre la opinión pública en general



0. Preguntas Introductorias | ¿Qué hacemos acá?

Aplicaciones



ESTADÍSTICA:

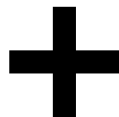
El uso de un índice de riesgo laboral permite abordar distintas problemáticas propias del mundo del trabajo a partir de la política pública

PROBABILIDAD:

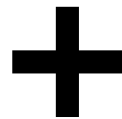
A un banco le es útil calcular la probabilidad de repago de una persona al otorgar un crédito

INFERENCIA:

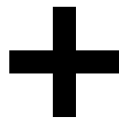
Cálculo de diferencias significativas en la intención de voto a dos o más candidatos



- A) Terminologías y definiciones
- B) Distribuciones de Frecuencia
- C) Medidas de Tendencia Central
- D) Variabilidad
- E) Distribución Bivariada
- F) Dependencia de la media
- G) Regresión Lineal Simple
- H) Correlación



- A) Terminologías y definiciones
- B) Distribuciones de Frecuencia
- C) Medidas de Tendencia Central
- D) Variabilidad
- E) Distribución Bivariada
- F) Dependencia de la media
- G) Regresión Lineal Simple
- H) Correlación



POBLACION

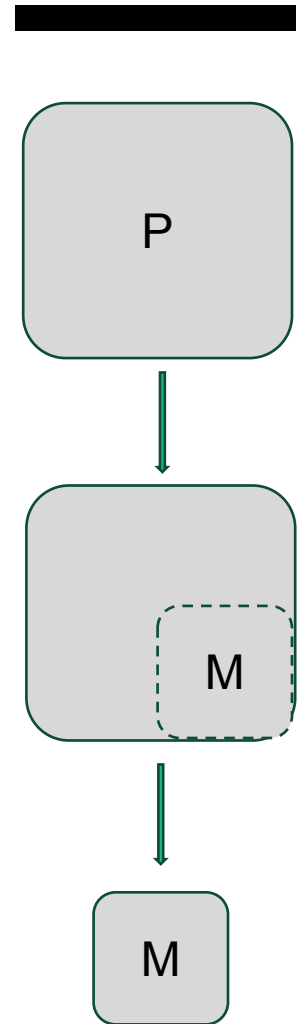
La Población es el conjunto de todas las entidades que tenemos interés en estudiar. Esto incluye desde una colección de objetos, transacciones, eventos, personas, etc.

MUESTRA

La Muestra es el subconjunto de las unidades de una población obtenida a través de un proceso de selección con el propósito de investigar las características de la población.

¿Cuál es la población en...

- a) Una encuesta sobre las opiniones de los argentinos respecto al fin del sistema SUBE, basada en entrevistas a una muestra de 1500 personas en el AMBA?
- b) Una encuesta a todas las empresas del sector mecánico en una determinada área geográfica para estudiar sus características y, específicamente, determinar el número total de empleados?



PARÁMETRO

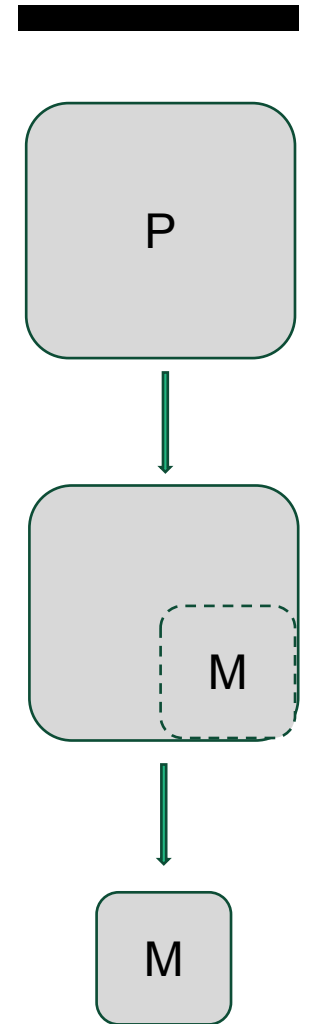
Es una cantidad descriptiva de la población total

ESTADÍSTICO

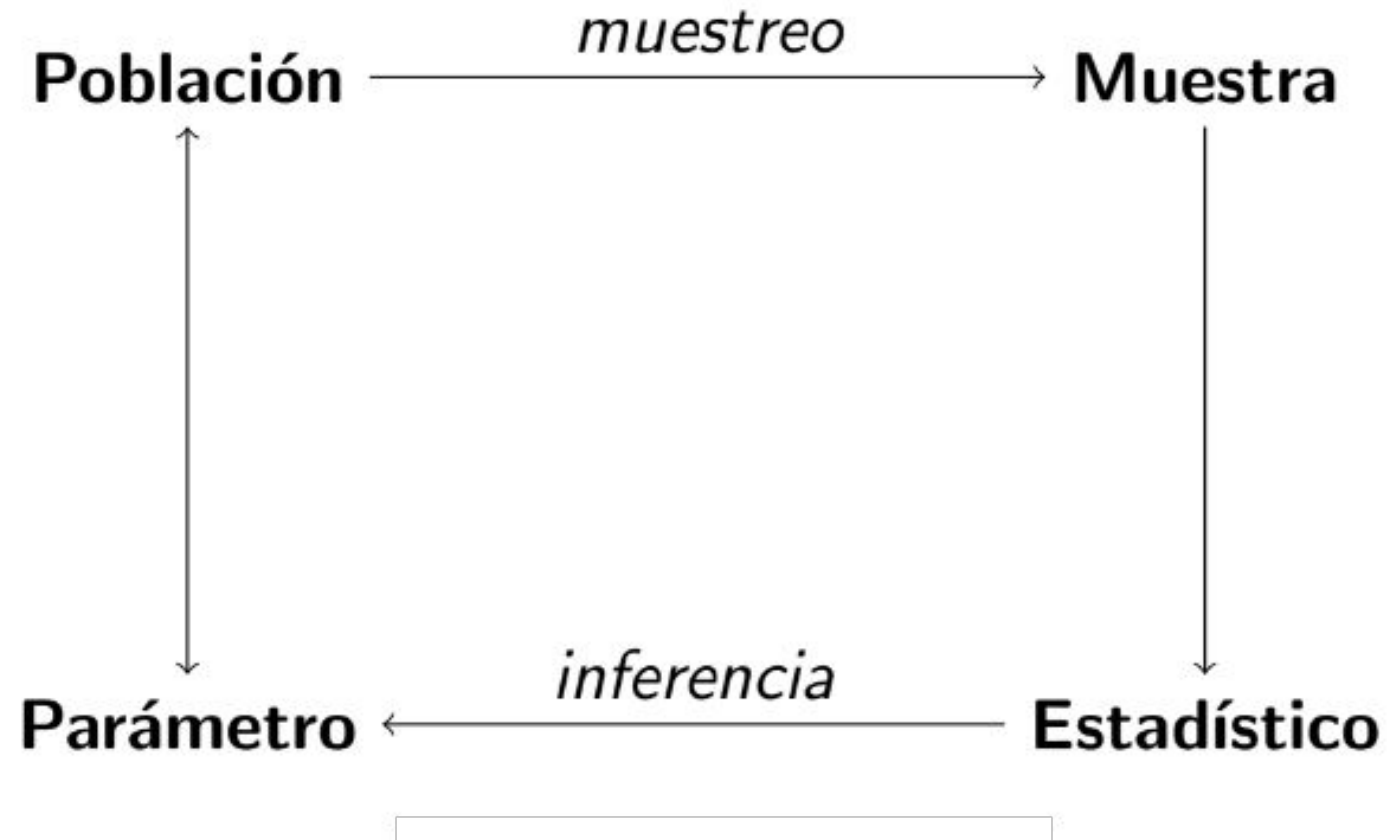
Elemento que describe una muestra, es función de éstos. Algunos de ellos sirven como estimadores de parámetros poblacionales.

Por ejemplo:

Se quiere conocer la edad promedio de los alumnos de ciencia política de Fsoc actualmente (**parámetro**). Para ello se toma una muestra de 400 alumnos que cursan la carrera actualmente y se calcula la media (**estadístico/estimador**) de las edades de los 400 encuestados.



1. Estadística I A) Terminologías y Definiciones: Población y Muestra



UNIDAD ESTADÍSTICA

Las Unidades Estadísticas son los miembros individuales de la población. Son las entidades desde las cuales se recolecta la data.

VARIABLE

La Variable es cualquier característica o propiedad de la unidad estadística que queremos analizar. Puede variar tomando diferentes valores de una unidad a otra.

¿Cuál es la variable en los ejemplos posteriores?

- a) Una encuesta sobre las opiniones de los argentinos respecto al fin del sistema SUBE, basada en entrevistas a una muestra de 1,500 personas en el AMBA.
- b) Una encuesta a todas las empresas del sector mecánico en una determinada área geográfica para estudiar sus características y, específicamente, determinar el número total de empleados.

UNIDAD ESTADÍSTICA

Las Unidades Estadísticas son los miembros individuales de la población. Son las entidades desde las cuales se recolecta la data.

VARIABLE

La Variable es cualquier característica o propiedad de la unidad estadística que queremos analizar. Puede variar tomando diferentes valores de una unidad a otra.

¿Cuál es la variable en los ejemplos anteriores?

- a) Una encuesta sobre **las opiniones** de los argentinos respecto al fin del sistema SUBE, basada en entrevistas a una muestra de 1,500 personas en el AMBA.
- b) Una encuesta a todas las empresas del sector mecánico en una determinada área geográfica para estudiar sus **características** y, específicamente, determinar **el número total de empleados**.

UNIDAD ESTADÍSTICA

Las Unidades Estadísticas son los miembros individuales de la población. Son las entidades desde las cuales se recolecta la data.

VARIABLE

La Variable es cualquier característica o propiedad de la unidad estadística que queremos analizar. Puede variar tomando diferentes valores de una unidad a otra.

¿Cuál es la unidad estadística en los ejemplos anteriores?

- a) Una encuesta sobre las opiniones de los argentinos respecto al fin del sistema SUBE, basada en entrevistas a una muestra de 1,500 personas en el AMBA.
- b) Una encuesta a todas las empresas del sector mecánico en una determinada área geográfica para estudiar sus características y, específicamente, determinar el número total de empleados.

UNIDAD ESTADÍSTICA

Las Unidades Estadísticas son los miembros individuales de la población. Son las entidades desde las cuales se recolecta la data.

VARIABLE

La Variable es cualquier característica o propiedad de la unidad estadística que queremos analizar. Puede variar tomando diferentes valores de una unidad a otra.

¿Cuál es la unidad estadística en los ejemplos anteriores?

- a) Una encuesta sobre las opiniones de **los argentinos** respecto al fin del sistema SUBE, basada en entrevistas a una muestra de 1,500 personas **en el AMBA. Cada uno de los porteños encuestados.**
- b)
- c) Una encuesta a todas **las empresas del sector mecánico** en una determinada área geográfica para estudiar sus características y, específicamente, determinar el número total de empleados. **Cada empresa del sector mecánico.**
- d)

¿Cómo se obtienen los datos?



ENCUESTAS Subgrupo de la población. Importancia de las distintas técnicas de muestreo

CENSOS Recolección de información de toda la población

DATOS ADMINISTRATIVOS Bases de datos de empresas, organismos nacionales e internacionales, etc.

DATOS DE ESTUDIOS EXPERIMENTALES

INTERNET Web scrapping de redes sociales, sitios web, sitios de noticias, etc.



3 tipos de datos



DATOS DE CORTE TRANSVERSAL

DATOS DE SERIE TEMPORAL

DATOS DE PANEL



Corte transversal:

| i | X_1 | X_2 | | x_p |
|---|----------|----------|-------|----------|
| 1 | X_{11} | X_{12} | | x_{1p} |
| 2 | X_{21} | X_{22} | | x_{2p} |
| . | . | . | | . |
| . | . | . | | . |
| . | . | . | | . |
| n | X_{n1} | X_{n2} | | x_{np} |



Datos tomados en un punto de tiempo determinado. Múltiples unidades de análisis: transversalidad

Series temporales

| t | X_1 | X_2 | | x_p |
|---|----------|----------|-------|----------|
| 1 | X_{11} | X_{12} | | x_{1p} |
| 2 | X_{21} | X_{22} | | x_{2p} |
| · | · | · | | · |
| · | · | · | | · |
| · | · | · | | · |
| T | X_{T1} | X_{T2} | | x_{Tp} |



Datos tomados para una unidad de análisis a lo largo del tiempo

1. Estadística | A) Terminologías y Definiciones: Tipos de datos. Un ejemplo



: P16ST

6

Visible: 275 de 275 variables

| | NUMINVESTES | IDENPA | NUMENTRE | REG | CIUDAD | TAMCIUD | COMDIST | EDAD | SEXO | CODIGO | DIAREAL | MESREAL | INTENCION_VOTO | P16ST | INI |
|----|-------------|-----------|----------|---------------|---------------|-----------|---------|------|--------|--------|---------|---------|--------------------------------|-------|------|
| 1 | 2023 | Argentina | 1 | AR: Capita... | AR: Capita... | (Capital) | 1 | 67 | Mujer | 1 | 22 | Febrero | No sabe | 6 | 842 |
| 2 | 2023 | Argentina | 2 | AR: Capita... | AR: Capita... | (Capital) | 1 | 29 | Mujer | 1 | 22 | Febrero | Peronismo | 7 | 1434 |
| 3 | 2023 | Argentina | 3 | AR: Capita... | AR: Capita... | (Capital) | 1 | 71 | Hombre | 1 | 24 | Febrero | Peronismo | 5 | 925 |
| 4 | 2023 | Argentina | 4 | AR: Capita... | AR: Capita... | (Capital) | 1 | 31 | Hombre | 1 | 24 | Febrero | Peronismo | 5 | 1438 |
| 5 | 2023 | Argentina | 5 | AR: Capita... | AR: Capita... | (Capital) | 1 | 28 | Hombre | 1 | 26 | Febrero | Peronismo | 5 | 1348 |
| 6 | 2023 | Argentina | 6 | AR: Capita... | AR: Capita... | (Capital) | 1 | 51 | Mujer | 1 | 26 | Febrero | AR: Union Civica Radical (UCR) | 5 | 1941 |
| 7 | 2023 | Argentina | 7 | AR: Capita... | AR: Capita... | (Capital) | 1 | 77 | Mujer | 1 | 27 | Febrero | Peronismo | 3 | 947 |
| 8 | 2023 | Argentina | 8 | AR: Capita... | AR: Capita... | (Capital) | 1 | 35 | Hombre | 1 | 27 | Febrero | No vota/Ninguno | 6 | 1447 |
| 9 | 2023 | Argentina | 9 | AR: Capita... | AR: Capita... | (Capital) | 1 | 26 | Hombre | 1 | 28 | Febrero | AR: Avanza Libertad | 7 | 838 |
| 10 | 2023 | Argentina | 10 | AR: Capita... | AR: Capita... | (Capital) | 1 | 25 | Hombre | 1 | 28 | Febrero | No vota/Ninguno | 5 | 1421 |

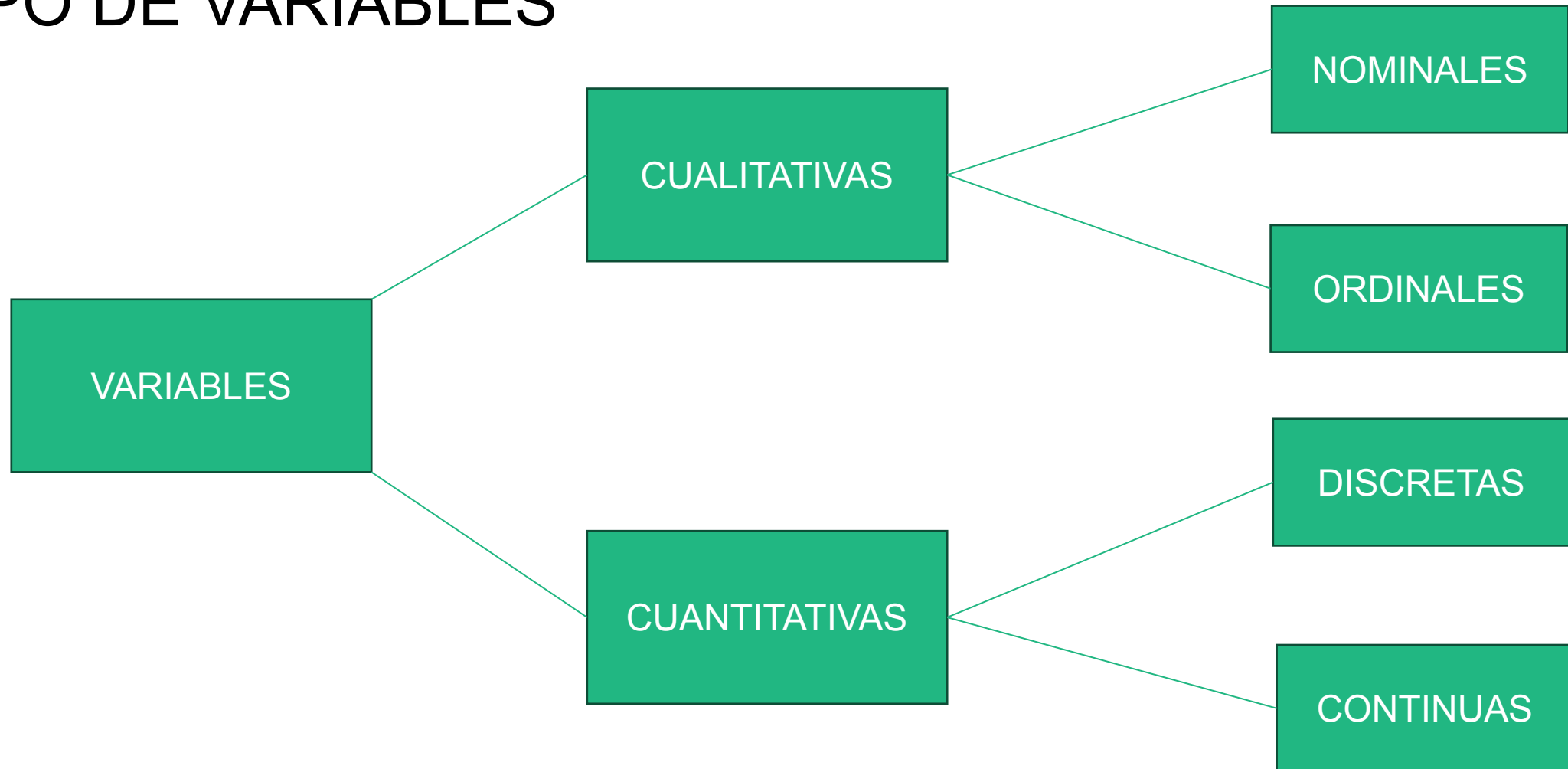


MATRIZ

La Matriz es la manera en que se organizan los datos estadísticos con el fin de procesarlos o diseminarlos para otros estudios y análisis. La Matriz de Datos ordena los datos a través de columnas y filas, donde cada columna se corresponde con una variable y cada fila con una unidad estadística.

| MATRIZ | VARIABLE 1 | VARIABLE 2 | VARIABLE n |
|----------------------|------------|------------|------------|
| UNIDAD ESTADÍSTICA 1 | V1UE1 | V2UE1 | VnUE1 |
| UNIDAD ESTADÍSTICA 2 | V1UE2 | V2UE2 | VnUE2 |
| UNIDAD ESTADÍSTICA N | V1UEN | V2UEN | VnUEN |

TIPO DE VARIABLES



VARIABLE CUALITATIVA

Se expresan mediante CATEGORÍAS (atributos, nombres o etiquetas) que deben ser:

- Mutuamente excluyentes (ninguna unidad estadística puede ser ubicada en más de una categoría)
- Exhaustiva (hay una categoría para cada unidad estadística)

Las categorías pueden o no estar organizadas en algún orden:

- Si el orden es indistinto, entonces tenemos una variable NOMINAL (Ej.: sexo, partido político, religión, etc.)
- Si es posible establecer un orden, tenemos una variable ORDINAL (Ej.: nivel de satisfacción). Estas categorías pueden ser etiquetadas con números u otros símbolos. En cualquier caso, cada categoría puede ser considerada mayor a ($>$) o menor a ($<$) su vecina

VARIABLE CUANTITATIVA

Se expresan mediante NÚMEROS, por lo tanto, pueden ser:

- DISCRETAS cuando se trata de un número contable de valores posibles (Ej.: número de estudiantes que asisten a clase en un curso, o cantidad de personas que viven en un hogar, etc.). Su medición es el resultado de un proceso de recuento de cada unidad estadística en consideración (sea de la población o de la muestra).
- CONTINUAS cuando puede tomar un valor cualquiera en un intervalo de números reales (Ej. Altura, edad, temperatura). Su medición no es un recuento como el caso de las variables discretas. Más bien consiste en observar las marcas con un instrumento de medición apropiado. Dada la limitada precisión de los instrumentos de medición, aun para las variables continuas, las mediciones son necesariamente discretas.

UN EJEMPLO SOBRE VARIABLES

El jefe de recursos humanos de una empresa textil provee un cuestionario breve a 10 trabajadore/as preguntando:

- grado de dificultad de su trabajo;
- número de hijos en su familia;
- el promedio de la hora salarial (en pesos); y
- si posee un auto propio

UN EJEMPLO SOBRE VARIABLES

| Nº TRABAJADOR | DIFICULTAD | Nº DE HIJO/AS | PROMEDIO HORA (\$) | AUTO PROPIO |
|---------------|------------|---------------|--------------------|-------------|
| 1 | 3 | 0 | 2.250 | 0 |
| 2 | 4 | 1 | 2.300 | 1 |
| 3 | 1 | 3 | 1.850 | 1 |
| 4 | 3 | 2 | 2.400 | 1 |
| 5 | 1 | 0 | 2.600 | 0 |
| 6 | 3 | 1 | 1.930 | 0 |
| 7 | 1 | 0 | 2.150 | 1 |
| 8 | 2 | 2 | 1.670 | 1 |
| 9 | 3 | 2 | 1.790 | 1 |
| 10 | 4 | 1 | 2.300 | 0 |

Dificultad

- 1 = No muy agotador
- 2 = Moderadamente agotador
- 3 = Agotador
- 4 = Muy agotador
- 5 = Extremadamente agotador

Auto propio

- 0 = No
- 1 = Si

... ALGO OBVIO

- a) Un valor particular puede repetirse varias veces en la misma columna.
- b) Al seleccionar una variable específica, es posible identificar en la tabla el valor observado para cada unidad estadística.
- c) A la inversa, al seleccionar una unidad estadística específica, es posible leer todo su conjunto de valores observados (para cada variable).

UN EJEMPLO SOBRE VARIABLES

- a) ¿Cuáles y cuántas son las unidades estadísticas?
- b) Para cada variable, indica si es cualitativa (nominal/ordinal) o cuantitativa (discreta/continua).
- c) ¿El codificado utilizado para "grado de pesadez del trabajo" y "disponibilidad de un auto propio" cambia la naturaleza de las dos variables?
- d) ¿Podemos afirmar que la diferencia entre 'No muy cansador' y 'Moderadamente cansador' es igual a la diferencia entre 'Muy cansador' y 'Extremadamente cansador'?

UN EJEMPLO SOBRE VARIABLES

- a) Los trabajadores de la fábrica textil, 10 unidades estadísticas
- b) Nro de trabajadores, nro de hijos, promedio \$/h: v. continuas, dificultad: v. ordinal, auto propio: v. nominal
- c) No, no confundir etiquetas o códigos con valores reales
- d) No, no tenemos forma de estimar la distancia entre categorías de variables ordinales

1. Estadística | A) Terminologías y Definiciones: ej. de OP: Latinobarómetro

P6STGBS- ¿Considera que la situación económica actual del país es mejor, un poco mejor, igual, un poco peor o mucho peor que hace 12 meses?

| | TOTAL | Sexo entrevistado | | Edad | | | |
|---------------|------------|-------------------|-------|-------|-------|-------|----------|
| | | Hombre | Mujer | 15-25 | 26-40 | 41-60 | 61 y más |
| Mucho mejor | 0,3 (4) | 0,6 | 0,1 | 0,4 | 0,3 | 0,4 | 0,3 |
| Un poco mejor | 9,3 (111) | 9,2 | 9,4 | 7,6 | 7,8 | 11,2 | 10,2 |
| Igual | 21,0 (252) | 21,8 | 20,2 | 20,4 | 24,0 | 19,8 | 19,2 |
| Un poco peor | 33,2 (398) | 32,0 | 34,3 | 39,6 | 32,7 | 31,1 | 30,8 |
| Mucho peor | 35,6 (427) | 36,2 | 34,9 | 31,0 | 34,2 | 37,5 | 38,9 |
| No sabe | 0,6 (7) | 0,2 | 1,0 | 1,1 | 0,7 | - | 0,7 |
| No contesta | 0,1 (1) | - | 0,1 | - | 0,3 | - | - |
| (N) | (1.200) | (580) | (620) | (238) | (344) | (361) | (257) |

P10STGBS- ¿Con cuál de las siguientes frases está Ud más de acuerdo?

| | TOTAL | Sexo entrevistado | | Edad | | | |
|--|------------|-------------------|-------|-------|-------|-------|----------|
| | | Hombre | Mujer | 15-25 | 26-40 | 41-60 | 61 y más |
| La democracia es preferible a cualquier otra forma de gobierno | 61,6 (740) | 62,7 | 60,7 | 60,3 | 59,7 | 61,9 | 65,1 |
| En algunas circunstancias, un gobierno autoritario puede ser preferible a uno democrático. | 18,4 (221) | 18,8 | 18,0 | 16,2 | 18,2 | 19,8 | 18,9 |
| A la gente como uno, nos da lo mismo un régimen democrático que uno no democrático | 15,1 (181) | 14,3 | 15,9 | 15,6 | 16,8 | 14,8 | 12,8 |
| No sabe | 4,1 (50) | 3,8 | 4,5 | 8,0 | 4,1 | 3,0 | 2,2 |
| No contesta | 0,7 (8) | 0,5 | 0,9 | - | 1,2 | 0,5 | 1,0 |
| (N) | (1.200) | (580) | (620) | (238) | (344) | (361) | (257) |

1. Estadística | A) Terminologías y Definiciones: ej. de OP: Latinobarómetro

P16ST- En política, la gente normalmente habla de “izquierda” y “derecha”. En una escala donde 0 es la izquierda y 10 es la derecha, ¿dónde te ubicarías?

| | TOTAL | Sexo entrevistado | | Edad | | | |
|-------------------|------------|-------------------|-------|-------|-------|-------|----------|
| | | Hombre | Mujer | 15-25 | 26-40 | 41-60 | 61 y más |
| 00 Izquierda | 3,3 (40) | 2,8 | 3,8 | 2,4 | 4,0 | 2,7 | 4,2 |
| 1 | 1,5 (18) | 0,3 | 2,6 | 1,0 | 2,1 | 1,6 | 0,9 |
| 2 | 3,0 (37) | 2,5 | 3,6 | 4,0 | 4,7 | 2,0 | 1,5 |
| 3 | 5,4 (65) | 6,2 | 4,7 | 5,6 | 7,2 | 3,3 | 6,0 |
| 4 | 6,0 (72) | 6,9 | 5,3 | 6,0 | 6,3 | 5,6 | 6,4 |
| 5 | 31,8 (381) | 33,7 | 30,0 | 30,4 | 32,5 | 35,8 | 26,5 |
| 6 | 8,1 (97) | 8,7 | 7,6 | 8,4 | 8,6 | 8,9 | 6,1 |
| 7 | 7,7 (93) | 7,9 | 7,6 | 7,9 | 7,2 | 10,3 | 4,6 |
| 8 | 6,3 (76) | 6,1 | 6,6 | 4,4 | 4,8 | 5,6 | 11,3 |
| 9 | 0,7 (9) | 1,1 | 0,4 | 0,6 | 1,4 | 0,4 | 0,3 |
| 10 Derecha | 6,6 (79) | 7,0 | 6,2 | 5,6 | 5,3 | 7,4 | 8,2 |
| Ninguno | 12,2 (146) | 10,6 | 13,7 | 13,3 | 9,6 | 11,1 | 16,0 |
| No sabe | 5,1 (62) | 4,8 | 5,5 | 10,2 | 4,1 | 3,5 | 4,1 |
| No contesta | 2,1 (25) | 1,6 | 2,6 | 0,4 | 2,1 | 2,0 | 3,8 |
| (N) | (1.200) | (580) | (620) | (238) | (344) | (361) | (257) |
| Media | 17,4 | 15,8 | 18,9 | 18,9 | 14,5 | 16,3 | 21,5 |
| Desviación típica | 31,0 | 29,1 | 32,7 | 32,8 | 28,1 | 29,5 | 34,8 |
| N | (1.113) | (543) | (571) | (213) | (322) | (341) | (237) |

1. Estadística

- A) Terminologías y definiciones
- B) Distribuciones de Frecuencia
- C) Medidas de Tendencia Central
- D) Variabilidad
- E) Distribución Bivariada
- F) Dependencia de la media
- G) Regresión Lineal Simple
- H) Correlación



1) **Estadística descriptiva univariada:**

Tablas de frecuencias

Medidas de tendencia central

Medidas de dispersión y de posición

Visualización de los datos

2) **Estadística descriptiva bivariada:**

Tablas de contingencia

Medidas de asociación según tipos de variables



ANÁLISIS UNIVARIADO

DISTRIBUCIÓN DE FRECUENCIAS

El primer paso en el análisis estadístico es la distribución de frecuencias: esta resume la información del conjunto de valores individuales.

La distribución de frecuencia se construye al agrupar en k clases el N de unidades estadísticas de acuerdo con los posibles valores k de la variable observada

La palabra valor es utilizada para indicar los posibles resultados de la variable X , la categoría o número depende de si la variable es cualitativa o cuantitativa

SERIE DE OBSERVACIONES

En el conjunto de las unidades estadísticas $U = \{1, 2, \dots, N\}$ donde la variable X es observada, la secuencia x_1, x_2, \dots, x_k es la serie de categorías de la variable X

Digamos que X es una variable (indistintamente cualitativa o cuantitativa) y x_1, x_2, \dots, x_k son los valores que puede tomar.

La distribución de frecuencias es una tabla que enlista los valores de X junto con la cantidad de veces que asume ese valor, es decir, la frecuencia:

| VALOR DE X | FRECUENCIA |
|--------------|------------|
| x_1 | n_1 |
| x_2 | n_2 |
| x_k | n_k |
| Total | N |

Donde n_1, n_2, \dots, n_k es la frecuencia de x_1, x_2, \dots, x_k respectivamente



Hacemos la distribución de frecuencia de la variable *Auto Propio* del ejemplo anterior

| Nº TRABAJADOR | AUTO PROPIO |
|---------------|-------------|
| 1 | 0 |
| 2 | 1 |
| 3 | 1 |
| 4 | 1 |
| 5 | 0 |
| 6 | 0 |
| 7 | 1 |
| 8 | 1 |
| 9 | 1 |
| 10 | 0 |



| AUTO PROPIO (X) | FRECUENCIA |
|-----------------|------------|
| 0 | 4 |
| 1 | 6 |
| Total | 10 |

Nótese que la matriz contiene N filas, mientras que la distribución de frecuencias contiene k filas (y el total es la cuenta de cada fila)



DISTRIBUCIÓN DE FRECUENCIAS RELATIVAS

Las frecuencias relativas o proporciones se obtienen al dividir las frecuencias absolutas por N y se denota como f_1, f_2, \dots, f_k donde

$$f_i = \frac{\text{frecuencia de } x_i}{\text{número total de unidades}} = \frac{n_i}{N}, \quad i = 1, 2, \dots, k$$

Las frecuencias relativas son útiles para comparar la importancia (el peso) de una categoría individual dentro de la distribución de frecuencias.

Las siguientes son **propiedades** de las frecuencias relativas

- $f_1 + f_2 + \dots + f_k = 1$;
- $0 \leq f_i \leq 1, i = 1, 2, \dots, k$.

DISTRIBUCIÓN DE FRECUENCIAS PORCENTUALES

Las frecuencias porcentuales se utilizan y pueden ser obtenidas al multiplicar las frecuencias relativas por 100:

$$p_i = f_i * 100, i = 1, 2, \dots, k$$

Las siguientes son **propiedades** de las frecuencias porcentuales

- $p_1 + p_2 + \dots + p_k = 100$;
- $0 \leq p_i \leq 100, i = 1, 2, \dots, k$.

DISTRIBUCIÓN DE FRECUENCIAS RELATIVAS Y PORCENTUALES

Las tablas donde las frecuencias absolutas son reemplazadas por frecuencias relativas o por frecuencias porcentuales se llaman **distribución de frecuencia relativas** y **distribución de frecuencias porcentuales**

| VALOR DE X | FREQ. RELATIVA |
|------------|-------------------|
| x_1 | $f_1 = n_1/N$ |
| x_2 | $f_2 = n_2/N$ |
| ... | ... |
| x_k | $f_k = n_k/N$ |
| TOTAL | 1 |

| VALOR DE X | FREQ. PORCENTUAL |
|------------|---------------------|
| x_1 | $p_1 = f_1 * 100$ |
| x_2 | $p_2 = f_2 * 100$ |
| ... | ... |
| x_k | $p_k = f_k * 100$ |
| TOTAL | 100 |

EJEMPLO DE DISTRIBUCIÓN ABSOLUTA, RELATIVA Y PORCENTUAL:

| Nº TRABAJADOR | AUTO PROPIO | AUTO PROPIO (X) | FREQ. ABSOLUTA | FREQ. RELATIVA | FREQ. PORCENTUAL |
|---------------|-------------|-----------------|----------------|----------------|------------------|
| 1 | 0 | 0 | | | |
| 2 | 1 | 1 | | | |
| 3 | 1 | 1 | | | |
| 4 | 1 | Total | 10 | 1 | 100 |
| 5 | 0 | | | | |
| 6 | 0 | | | | |
| 7 | 1 | | | | |
| 8 | 1 | | | | |
| 9 | 1 | | | | |
| 10 | 0 | | | | |

TABLA DE FRECUENCIAS

MATRIZ

EJEMPLO DE DISTRIBUCIÓN ABSOLUTA, RELATIVA Y PORCENTUAL:

| Nº TRABAJADOR | AUTO PROPIO | AUTO PROPIO (X) | FREQ. ABSOLUTA | FREQ. RELATIVA | FREQ. PORCENTUAL |
|---------------|-------------|-----------------|----------------|----------------|------------------|
| 1 | 0 | 0 | 4 | 0,4 | 40% |
| 2 | 1 | 1 | 6 | 0,6 | 60% |
| 3 | 1 | | | | |
| 4 | 1 | Total | 10 | 1 | 100 |
| 5 | 0 | | | | |
| 6 | 0 | | | | |
| 7 | 1 | | | | |
| 8 | 1 | | | | |
| 9 | 1 | | | | |
| 10 | 0 | | | | |

TABLA DE FRECUENCIAS

MATRIZ

FRECUENCIAS ACUMULADAS

Considerando una variable ordinal o cuantitativa, la frecuencia acumulada de un valor particular de una variable se define como el número de observaciones que son menores o iguales al valor. Nos dice cuántos casos o respuestas se han acumulado hasta cierto valor.

Atención: no aplicable a variables nominales.

DISTRIBUCIÓN DE FRECUENCIAS ACUMULADAS

La distribución de frecuencias acumuladas se define como el enlistamiento de los valores de la variable junto con su frecuencia acumulada correspondiente. Nos permite entender cómo se distribuyen las respuestas a lo largo de los valores de la variable, viendo el acumulado progresivo.

| VALOR DE X | FREQ. ACUMULADA |
|------------|-------------------|
| x_1 | $N_1 = n_1$ |
| x_2 | $N_2 = n_1 + n_2$ |
| ... | ... |
| x_k | |



$$\sum_{i=1}^k n_i$$

Este símbolo es el de la sumatoria y puede leerse como “la suma de n_i desde $i = 1$ a k ” donde i es el índice de sumatoria y k es el límite de la sumatoria.

DISTRIBUCIÓN DE FRECUENCIAS RELATIVAS ACUMULADAS

La siguiente ratio es llamada frecuencia relativa acumulada

$$F_i = \frac{N_i}{N}, i = 1, 2, \dots, k.$$

Similarmente, las siguientes cantidades se llaman frecuencias porcentuales acumuladas

$$P_i = F_i * 100, i = 1, 2, \dots, k.$$

DISTRIBUCIÓN DE FRECUENCIAS RELATIVAS Y PORCENTUALES ACUMULADAS

Las tablas donde las frecuencias absolutas son reemplazadas por frecuencias relativas acumuladas o por frecuencias porcentuales acumuladas se llaman **distribución de frecuencia relativas acumuladas** y **distribución de frecuencia porcentual acumuladas**

| VALOR DE X | FREQ. RELATIVA ACUMULADA |
|------------|--------------------------------|
| x_1 | $F_1 = f_1$ |
| x_2 | $F_2 = f_1 + f_2$ |
| ... | ... |
| x_k | |

| VALOR DE X | FREQ. PORCENTUAL ACUMULADA |
|------------|----------------------------------|
| x_1 | $P_1 = p_1$ |
| x_2 | $P_2 = p_1 + p_2$ |
| ... | ... |
| x_k | |

DISTRIBUCIÓN DE FRECUENCIAS RELATIVAS Y PORCENTUALES ACUMULADAS

Ejemplo 1 (variable intervalar o numérica)

| CANTIDAD DE HIJOS | FRECUENCIA | FRECUENCIA ACUMULADA | FRECUENCIA RELATIVA | FRECUENCIA ACUMULADA RELATIVA |
|-------------------|------------|----------------------|---------------------|-------------------------------|
| 0 | 3 | 3 | 0.11 | 0.11 |
| 1 | 7 | 10 | 0.25 | 0.36 |
| 2 | 10 | 20 | 0.36 | 0.72 |
| 3 | 3 | 23 | 0.11 | 0.83 |
| 4 | 4 | 27 | 0.14 | 0.95 |
| 5 | 0 | 27 | 0 | 0.95 |
| 6 | 1 | 28 | 0.04 | 1 |

DISTRIBUCIÓN DE FRECUENCIAS RELATIVAS Y PORCENTUALES ACUMULADAS

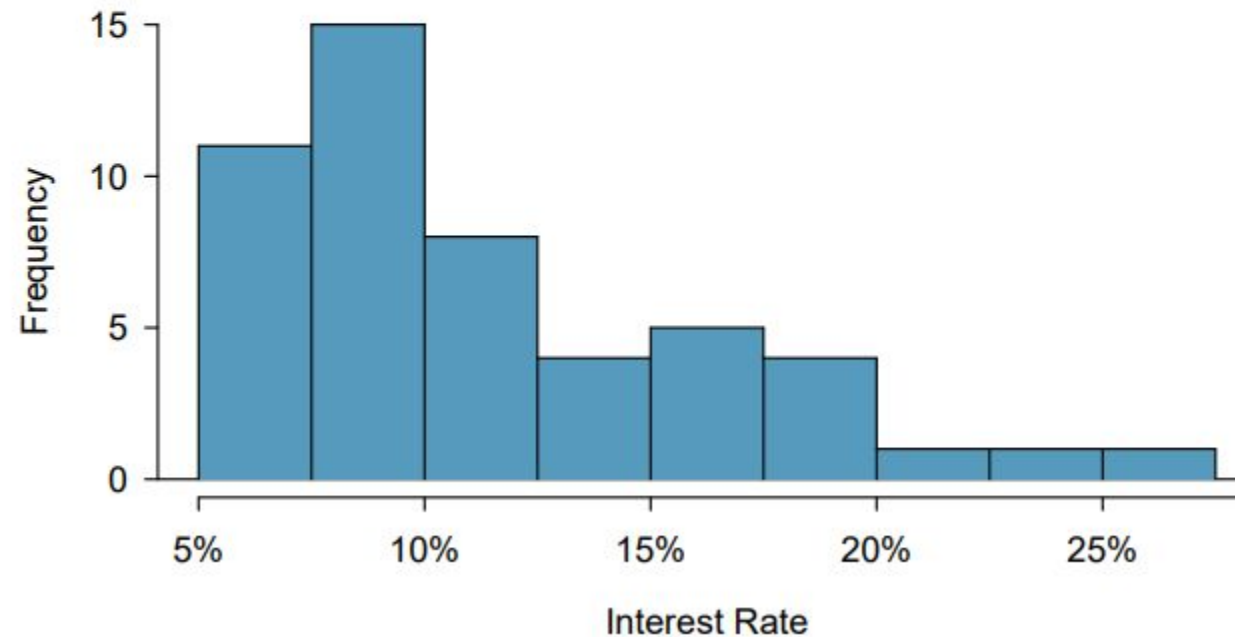
Ejemplo 2 (variable ordinal)

| OPINIÓN SOBRE GESTIÓN NACIONAL | FRECUENCIA | FRECUENCIA ACUMULADA | FRECUENCIA RELATIVA | FRECUENCIA ACUMULADA RELATIVA |
|--------------------------------|------------|----------------------|---------------------|-------------------------------|
| Muy mala | 11 | 11 | 0.11 | 0.11 |
| Mala | 22 | 33 | 0.22 | 0.33 |
| Ni buena mala | 40 | 73 | 0.40 | 0.73 |
| Buena | 20 | 93 | 0.20 | 0.93 |
| Muy buena | 7 | 100 | 0.07 | 1 |

Graficando una distribución univariada: histograma de la tasa de interés

| Interest Rate | 5.0% - 7.5% | 7.5% - 10.0% | 10.0% - 12.5% | 12.5% - 15.0% | ... | 25.0% - 27.5% |
|---------------|-------------|--------------|---------------|---------------|-----|---------------|
| Count | 11 | 15 | 8 | 4 | ... | 1 |

Figure 2.5: Counts for the binned `interest_rate` data.




Se trata de una distribución sesgada hacia la derecha



MEDIDAS DE RESUMEN



MEDIDAS DE TENDENCIA CENTRAL



Una medida de resumen o estadístico descriptivo es un número o valor capaz de resumir una gran cantidad de datos. Tenemos:

- 1) **Medidas de tendencia central:** resumen en un “valor típico” o representativo el conjunto de datos
- 2) **Medidas de dispersión:** resumen cuánto se alejan los datos respecto de ese centro
- 3) **Medidas de posición:** indican en qué lugar o punto de la distribución se encuentra un valor o un grupo de valores, dividiendo los datos en partes iguales





- La **moda** es el valor que más veces se repite en la distribución de datos

Distribución de frecuencia para la variable N° de hijos de cada trabajador

| N° TRABAJADOR | N° DE HIJO/AS | N° DE HIJOS | ni |
|---------------|---------------|-------------|----|
| 1 | 0 | 0 | 5 |
| 2 | 0 | 1 | 2 |
| 3 | 3 | 2 | 2 |
| 4 | 0 | 3 | 1 |
| 5 | 0 | | |
| 6 | 1 | | |
| 7 | 0 | | |
| 8 | 2 | | |
| 9 | 2 | | |
| 10 | 1 | | |

→ MODA



- La **mediana** es el valor que ocupa la posición central de los datos cuando éstos están ordenados de menor a mayor

Ejemplo, en una secuencia ordenada de valores de cantidad de notas:

Caso de mediana impar:

2, 4, 5, 6, 8, 9, 10

La mediana es aquel valor que divide el conjunto en iguales porciones de un lado y el otro

$$(N + 1) / 2$$

$$7 + 1 / 2 = 4$$

Caso de mediana par:

4, 5, 6, 8, 9, 10

La mediana es la media aritmética de los dos valores centrales que dividen el conjunto en iguales porciones de un lado y el otro, en este caso es 7 (promedio de 6 y 8).

- La **mediana** es el valor que ocupa la posición central de los datos cuando éstos están ordenados de menor a mayor

Ejemplo, en una secuencia ordenada ascendentemente de valores de notas:

- MEDIANA IMPAR

5,6,7,8,9

La mediana se calcula como $N+1 / 2$

Siendo que $N=5 \rightarrow 6 / 2 = 3$

La mediana de este conjunto de valores es aquel que está en la posición 3, es decir 7.

- MEDIANA PAR

4,5,6,7,8,9

La mediana se calcula como $N/2 + N/2 + 1$

Siendo que $N=6$:

$6/2$ $6/2 + 1 = 3$ y 4 son las posiciones de los dos valores centrales, y al promediarlos se obtiene la mediana:

$6 + 7 = 6,5$



En la práctica la mediana se suele aplicar sobre las frecuencias de los datos. Es el valor que divide a la muestra en dos partes iguales (en términos del porcentaje acumulado). Ejemplo:

| Rango de salario (en \$) | Porcentaje de la muestra |
|--------------------------|--------------------------|
| 0-60000 | 10% |
| 61000-100000 | 15% |
| 101000-200000 | 25% |
| 20100-300000 | 20% |
| 301000-400000 | 15% |
| 401000 o más | 10% |



- La **media** es un valor representativo de un conjunto de datos que resume, en un solo número, la “tendencia central” del conjunto, reflejando de manera típica los valores que lo componen. La media más conocida y utilizada es la aritmética, pero también existen otras:
- ❖ **Media aritmética:** es la suma de todos los valores observados dividida por la cantidad de datos
- ❖ **Media geométrica:** la raíz n -ésima del producto de todos los valores, útil cuando los datos se multiplican o representan tasas de crecimiento
- ❖ **Media ponderada:** similar a la aritmética, pero cada valor tiene un peso diferente que refleja su importancia relativa.

- Media aritmética

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Ejemplo: cálculo de la tasa de interés promedio

$$\bar{x} = \frac{10.90\% + 9.92\% + 26.30\% + \cdots + 6.08\%}{50} = 11.57\%$$

Nota: pese a ser una de las medidas más utilizadas y simples de calcular, la media presenta algunas limitaciones en términos de describir los datos reales que a veces podrían justificar el uso de otros estadísticos, pensemos por ejemplo en datos sobre ingresos en pesos de las personas.

- Media geométrica

$$\overline{X}_G = \sqrt[n]{\prod_{i=1}^n X_i} = \sqrt[n]{X_1 \cdot X_2 \cdots X_n}$$

La media geométrica suele utilizarse para calcular promedios de valores acumulados
Ejemplo de aplicación: cálculo de inflación en Argentina

Según cálculos del INDEC, la tasa de inflación en Argentina fue de 51% en 2021, 95% en 2022, y 211% en 2023.

Pregunta: ¿cuál fue la tasa promedio de inflación durante esos 3 años?

- $(51\% + 95\% + 211\%) / 3 = 119\%$



ERROR

- La inflación es un factor de crecimiento, es decir que la inflación de un año no es independiente de la del siguiente por ejemplo:

Si un precio sube un 51% en un año, un peso vale ahora:

$$1 \cdot (1 + 0,51) = 1,51$$

Si al año siguiente la inflación es de 95%, el mismo peso vale ahora:

$$1,51 \cdot (1 + 0,95) = 1,51 \cdot 1,95 = 2,94$$

Y al siguiente, con inflación de 211%:

$$2,94 \cdot (1 + 2,11) = 2,94 \cdot 3,11 = 9,14$$

$$1,51 * 1,95 * 3,11 = 9,16$$



Inflación bruta total entre esos 3 años= 916%
Es actor de crecimiento o inflación en porcentaje, luego se debe obtener la tasa de inflación.

¿Cuál es la inflación promedio a lo largo de ese período? Para distribuir la inflación bruta entre los tres años se toma raíz, cúbica en este caso porque tenemos 3 observaciones:

$$= \sqrt[3]{9,16}$$

$$= 2,1 - 1 = 1,1 = 110\%$$

- Media ponderada:

Siendo x_1, x_2, \dots, x_n las observaciones de una variable continua y w_1, w_2, \dots, w_n sus respectivos pesos, la media ponderada de estas observaciones está dada por la división entre la suma de las observaciones multiplicadas por sus pesos y la suma de sus pesos:

$$\bar{x} = \frac{(x_1 \cdot w_1) + (x_2 \cdot w_2) + \dots + (x_n \cdot w_n)}{w_1 + w_2 + \dots + w_n}$$

Ejemplo de media ponderada

Supongamos que se hizo una encuesta de opinión pública pero la misma quedó descompensada en cuanto a los datos de nivel socioeconómico:

NSE en la población general:

- 5% ABC1
- 45% C2C3
- 50% D1D2E

Distribución de NSE en la muestra:

- 20% ABC1
 - 30% C2C3
 - 50% D1D2E
- $5/20 = 0,25 = w_1$
 $45/30 = 1,5 = w_2$
 $50/50 = 1 = w_3$

Calculamos el promedio ponderado de la variable opinión sobre un candidato:

- Opiniones promedio para cada grupo:

ABC1 8 - C2C3 6,5 - D1D2E 5,5

$$\text{Promedio ponderado} = \frac{(8 * 0,25) + (6,5 * 1,5) + (5,5 * 1)}{0,25 + 1,5 + 1} = \frac{17,25}{2,75} = 6,27$$



- Media ponderada
- ❖ Este ejemplo ilustrativo vale para cualquier grupo en función del cual sea relevante ponderar en una investigación y para estimar el estadístico de cualquier variable (como las opiniones en este caso). En la práctica se aplican ponderadores en los distintos software/lenguajes de programación.
- ❖ Aplicaciones: Ponderación de encuestas a partir de datos de perfil (NSE, género. Targets difíciles de alcanzar o subrepresentados en la muestra por alguna razón). Encuesta sobre un tema al público general y a expertos en un tema de interés: se le atribuye mayor peso a las respuestas de expertos por sobre las del resto de los encuestados.



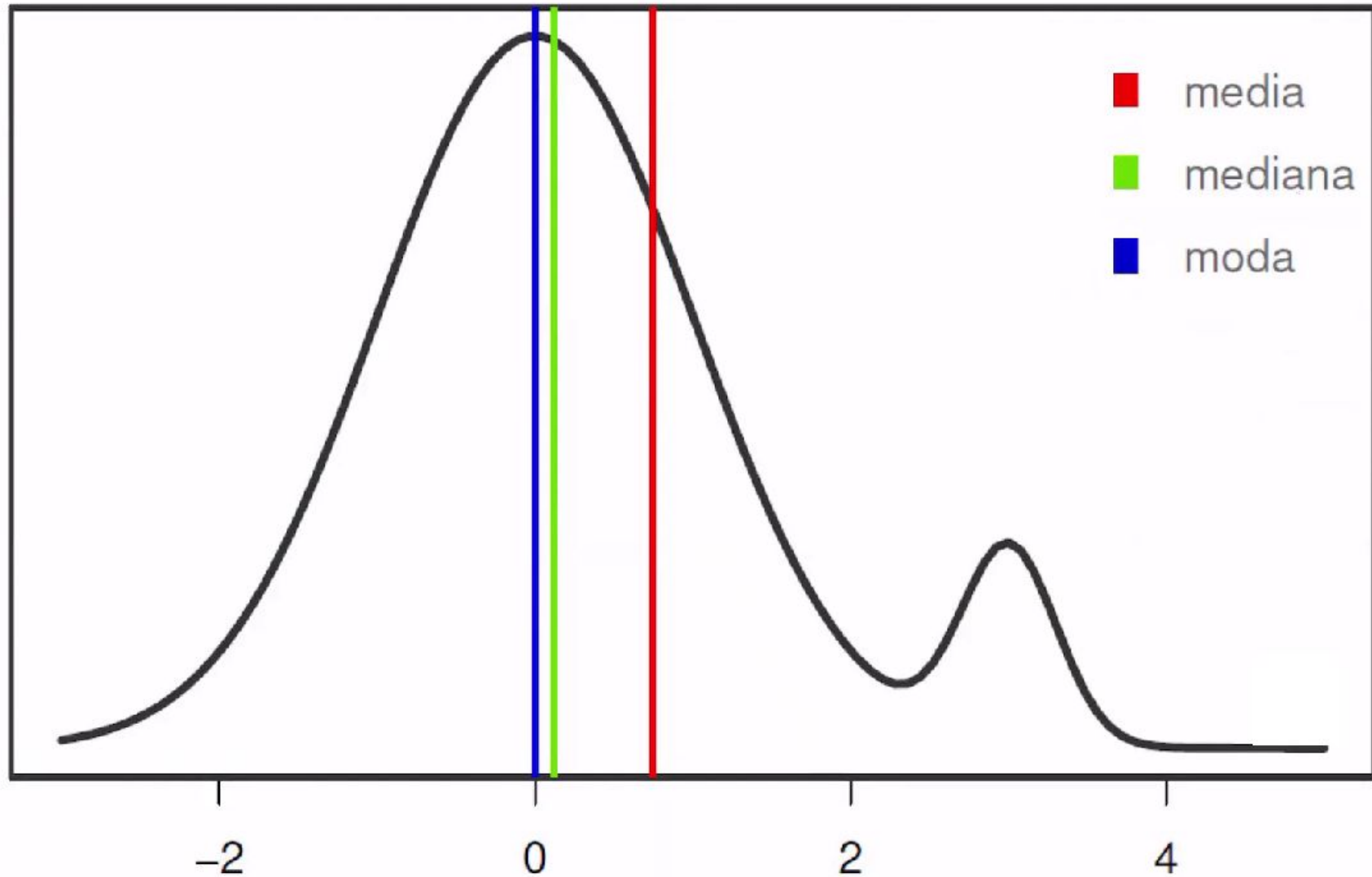


Pros y contras de cada medida

| Medida | Pro | Contra |
|---------|--|--|
| Moda | <ul style="list-style-type: none">-Cálculo sencillo-Interpretación clara-Puede utilizarse en variables cualitativas | <ul style="list-style-type: none">-Sensible a n (varía mucho según cuántos datos se coleccionan, y cómo)-Puede haber más de una moda-No siempre está en el centro-Poco sentido para variables continuas |
| Mediana | <ul style="list-style-type: none">-Cálculo sencillo-Robusto a outliers (por ej valores extremos, errores)-No es sensible a n (frente a algún cambio en la muestra la mediana no varía mucho) | <ul style="list-style-type: none">-Para variables ordenables-Puede ser un concepto poco familiar |
| Media | <ul style="list-style-type: none">-Fácil de entender-Poco sensible a n | <ul style="list-style-type: none">-Afectado por outliers |





Interpretación gráfica de lo que hace cada medida





A light gray vertical bar is located on the left side of the slide.A thick black horizontal bar is located in the top right corner of the slide.

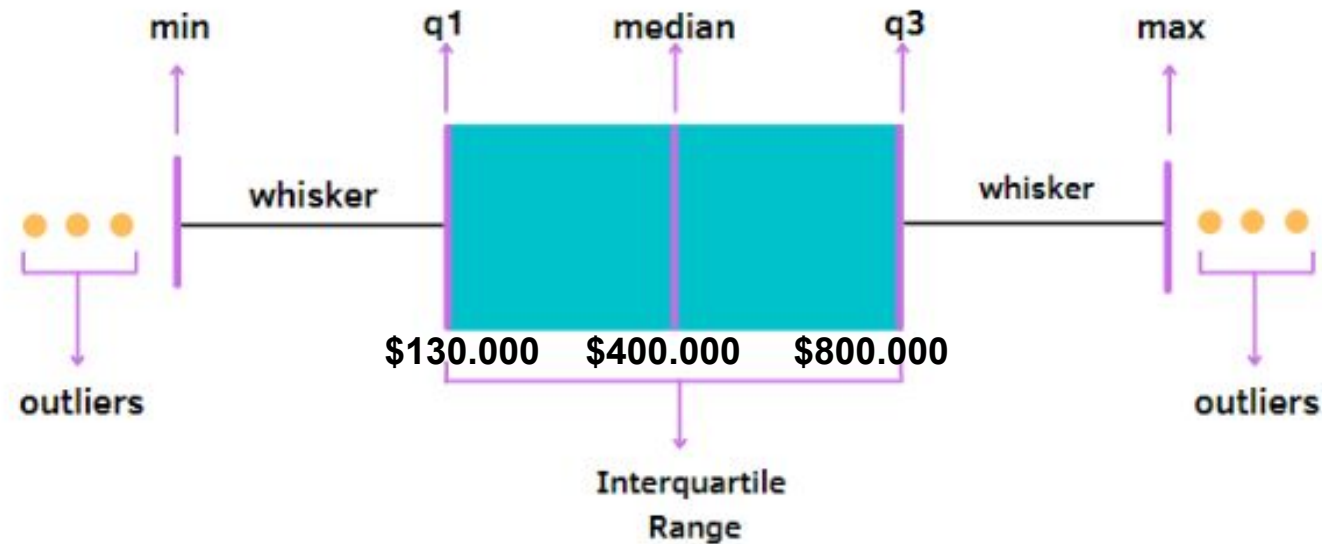
MEDIDAS DE POSICIÓN



- 
- 
- ❖ Las medidas de posición son indicadores estadísticos que nos ayudan a situar o ubicar un valor dentro de un conjunto de datos, mostrando dónde se encuentra respecto a los demás datos. A diferencia de las medidas de tendencia central (media, mediana, moda), las de posición no resumen toda la información, sino que permiten comparar y clasificar valores.
 - ❖ Algunas medidas de posición comunes:
 - Cuartiles (4 partes iguales)
 - Deciles (10 partes iguales)
 - Percentiles (100 partes iguales)
 - ❖ Muchas aplicaciones:
 - Dividir una población según percentiles (p.ej., los 25% con menor ingreso, los 25% con mayor ingreso) para análisis socioeconómico o marketing
 - Comparar ingresos o puntajes entre diferentes regiones o poblaciones, usando Q1, Q2 y Q3.
 - Permiten entender la dispersión y la concentración sin depender de la media

- 
- ❖ Los cuartiles dividen un conjunto de datos ordenados en cuatro partes iguales, de manera que cada parte contiene aproximadamente el 25% de los datos. Son tres valores clave:
 - Q1 (primer cuartil): separa el 25% inferior de los datos del 75% superior.
 - Q2 (segundo cuartil = mediana): separa la mitad inferior de la mitad superior, es decir, el 50% de los datos está por debajo y el 50% por encima.
 - Q3 (tercer cuartil): separa el 75% inferior del 25% superior de los datos.
- 

Boxplot Ejemplo 1: ingresos

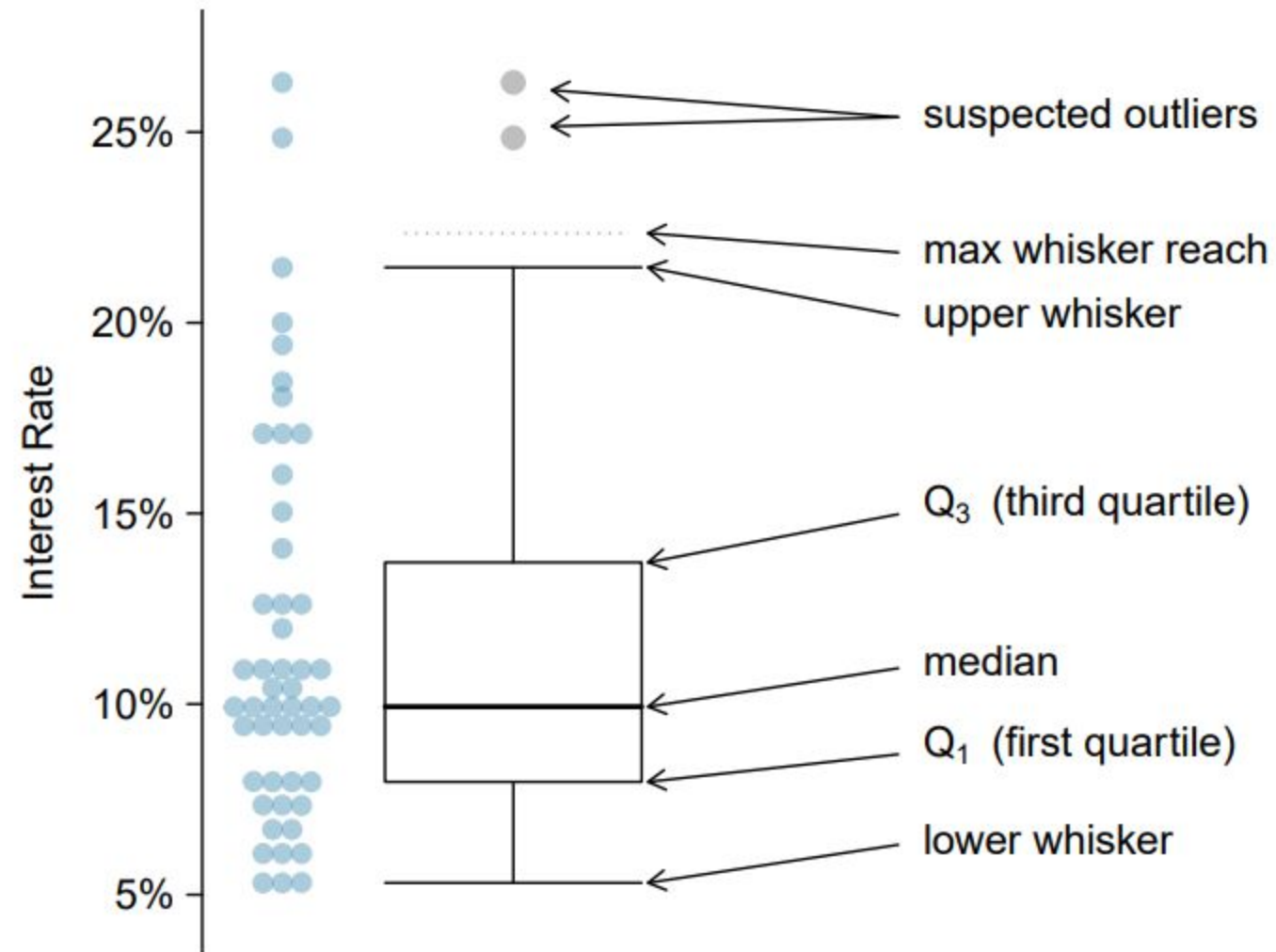


Frecuentemente se utilizan gráficos de cajas o boxplots para representar los cuartiles. Permiten ver gráficamente la distribución de una variable.

En este caso, por ejemplo, podemos decir que el 25% de los casos tiene un salario menor a \$130.000, o que el 25% superior gana más de \$800.000

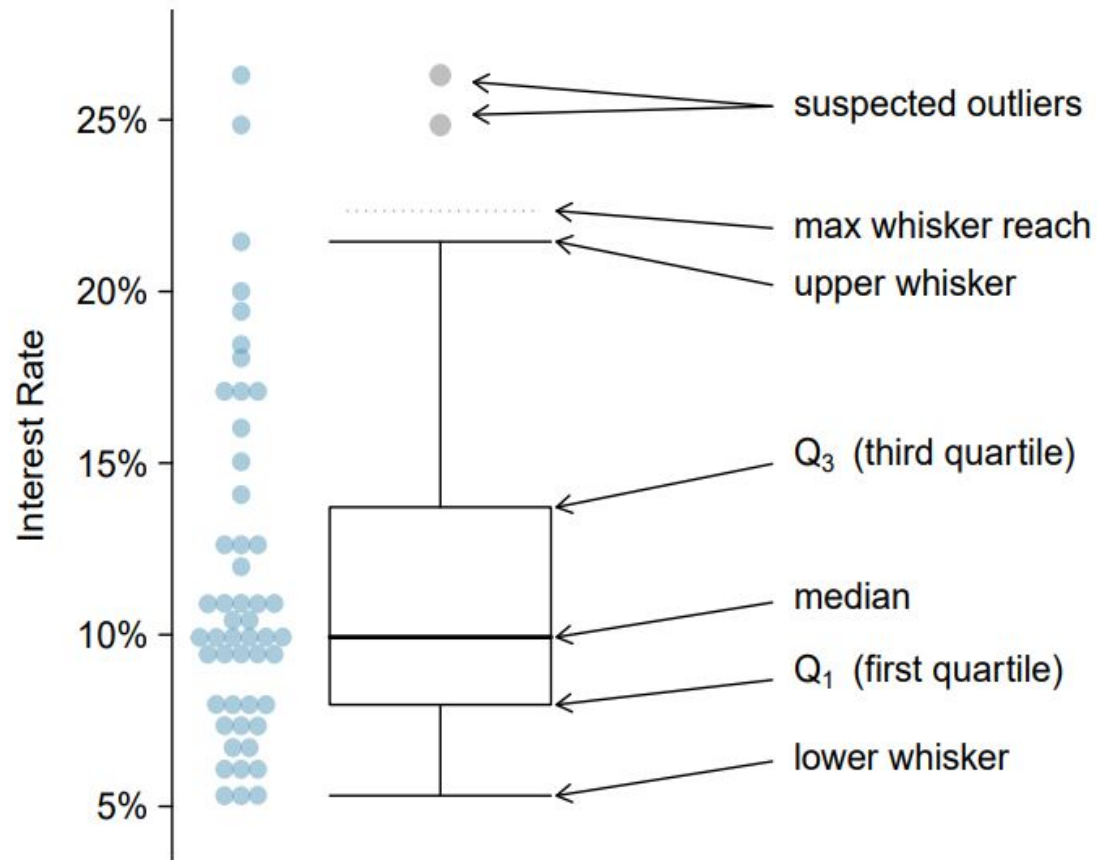
Boxplot

Ejemplo 2: tasas de interés



Boxplot

Ejemplo 2: tasas de interés



- En primer lugar se divide a los datos en dos mitades iguales, dadas por el valor mediano.
- La caja refleja donde se encuentra el 50% central de los datos (Q1-Q3)
- Los “whiskers” (líneas que salen de la caja) muestran la distribución fuera de los cuartiles, hasta ciertos límites (generalmente $1,5 * IQR$).
- Es un gráfico que permite visualizar valores típicos, hacia dónde se orientan los datos y outliers.
- En este caso se podría clasificar a 24,85% y 26,3% como outliers porque se alejan de la mayoría de los datos.



MEDIDAS DE VARIABILIDAD

VARIABILIDAD

Mientras el promedio provee un resumen de la muestra bajo estudio, la distribución es representada por la media siempre que la mayoría de las unidades asuman valores cercanos a ella.

Por VARIABILIDAD nos referimos a la tendencia que tienen los fenómenos naturales y sociales de manifestarse a sí mismos de diferentes maneras. Por lo tanto, la variabilidad de una distribución mide la tendencia de las unidades a asumir diferentes valores.

Con el fin de medir la variabilidad de una distribución, es posible emplear indicadores que resuman las desviaciones de cada valor y la tendencia central, o entre valores característicos de la distribución (ejemplo entre dos cuartiles)



INDICES DE VARIABILIDAD

Las herramientas que miden la variabilidad de la distribución se llaman ÍNDICES DE VARIABILIDAD y tienen dos propiedades:

- 1) Alcanza su menor valor cuando todas las unidades tienen el mismo valor
- 2) Aumenta su valor cuando hay “diversidad” de valores en distintas unidades

Algunos de estos índices basados en la desviación con respecto a la media son:

- VARIANZA
- DESVIACIÓN ESTÁNDAR
- COEFICIENTE DE VARIACIÓN



| Asignatura | Juan | María |
|---------------------|------|-------|
| Matemática | 7 | 4 |
| Lengua y literatura | 9 | 5 |
| Física | 4 | 6 |
| Biología | 2 | 7 |

Media:

5,5

5,5

¿Se puede decir que el rendimiento de ambos alumnos es igual?
¿Qué otro/s elemento/s podrían considerarse para analizarlo?



LA VARIANZA

La varianza de una serie de observaciones x_1, x_2, \dots, x_N es la diferencia de cada dato con respecto a la media, al cuadrado.

$$\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \dots + (x_N - \mu)^2}{N} = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

El numerador, es decir, $\sum_{i=1}^N (x_i - \mu)^2$, se llama DESVIACIÓN

PROPIEDADES DE LA VARIANZA

- 1) La varianza es 0 (cero) solamente si todos los valores son iguales (y, por lo tanto, coinciden con el promedio)
- 2) Las desviaciones de cada valor y la media son cuadradas. Esto es así porque al tomar el valor al cuadrado, cada diferencia es positiva y las desviaciones mayores “pesan” más, ya que aumentan considerablemente en comparación con las menores.

EL CASO DE LA DISTRIBUCIÓN DE FRECUENCIAS

Para el caso de una frecuencia distribuida la fórmula considera el peso que otorga las cantidad de unidades que caen en ese valor, es decir:

En los casos
de distribución
absoluta

$$\frac{1}{N} \sum_{i=1}^k (x_i - \mu)^2 n_i$$

En los casos
de distribución
relativa

$$\sum_{i=1}^k (x_i - \mu)^2 f_i$$



| NOTAS MATEMÁTICA (x_i) | CURSO A (n_i) | CURSO B (n_i) |
|----------------------------|-------------------|-------------------|
| 0 | 5 | 1 |
| 1 | 5 | 4 |
| 2 | 5 | 15 |
| 3 | 5 | 4 |
| 4 | 5 | 1 |
| TOTAL ALUMNOS | 25 | 25 |

Media: 2 2

¿Se puede decir que el rendimiento de ambas clases es igual?
¿Cómo se calcula la varianza del anterior ejemplo?



Muestra 1

$$\sigma_1^2 = \frac{(0-2)^2 \cdot 5 + (1-2)^2 \cdot 5 + (2-2)^2 \cdot 5 + (3-2)^2 \cdot 5 + (4-2)^2 \cdot 5}{25}$$

$$\sigma_1^2 = 50 / 25$$

$$\sigma_1^2 = 2$$

Muestra 2

$$\sigma_2^2 = \frac{(0-2)^2 \cdot 1 + (1-2)^2 \cdot 4 + (2-2)^2 \cdot 15 + (3-2)^2 \cdot 4 + (4-2)^2 \cdot 1}{25}$$

$$\sigma_2^2 = 16 / 25$$

$$\sigma_2^2 = 0,64$$

Tiene una varianza menor que la Muestra 1, lo que indica que sus valores están menos dispersos alrededor de la media

$$\frac{1}{N} \sum_{i=1}^k (x_i - \mu)^2 n_i$$

Referencias:

x_i , μ , n_i

DESVIACIÓN ESTÁNDAR

La varianza tiene la desventaja de no utilizar la misma unidad de medida que la variable, por ello, frecuentemente se recurre a la DESVIACIÓN ESTÁNDAR, la cual se obtiene tomando la raíz cuadrada de la varianza

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

La desviación estándar se expresa en la misma unidad de medida de la variable considerada.

PROPIEDADES DE LA DESVIACIÓN ESTÁNDAR

- 1) Es siempre NO-NEGATIVA y da CERO (0) sólo si todas las observaciones son iguales (por lo tanto iguales a la media)
- 2) No varía si le agregamos el mismo número a (sea positivo o negativo) a cada observación
- 3) Se multiplica por $|b|$ si multiplicamos todas las observaciones por un número b (sea positivo o negativo)

Retomando...

Varianza muestral

$$S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

← Corrección para muestras pequeñas
Relación con los grados de libertad

Recordar que \bar{x} es el promedio aritmético (media muestral)

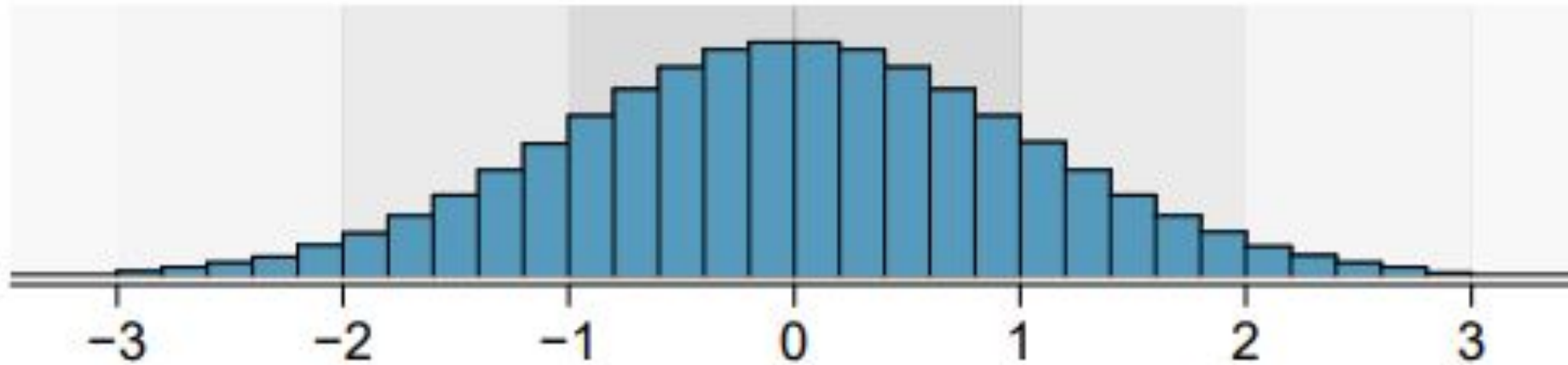
Desviación estándar

$$S = \sqrt{S^2}$$

La diferencia de cada dato con respecto a la media (en la fórmula de la varianza) se eleva al cuadrado para que los datos no se cancelen. Pero esta elevación al cuadrado dificulta la interpretación de la varianza.

Por este motivo se suele utilizar la desviación estándar, cuya unidad se corresponde con la de los datos originales

Desviación estándar y la distribución normal



En general, cuando los datos siguen una distribución normal alrededor del 70% de las obs se encontrarán a una desviación estándar de la media y un 95% a dos desviaciones estándar.

COMPARAR LA VARIABILIDAD DE DOS MUESTRAS

La varianza y la desviación estándar no se ajustan lo suficientemente bien al momento de querer comparar la variabilidad entre dos muestras de considerables magnitudes o que son medidas en diferentes unidades.

Se necesita una medida relativa de variabilidad: el ratio (multiplicado por 100) de una medida de absoluta variabilidad respecto de la media, este es EL
COEFICIENTE DE VARIACIÓN

EL COEFICIENTE DE VARIACION

Con el fin de comparar dos o más muestras en términos de su variabilidad, el coeficiente de variación es el más adecuado:

$$CV = \frac{\sigma}{\mu} \cdot 100$$

Al calcular la medición relativa de variabilidad, se elimina -en algún sentido- el nivel de influencia que tiene el promedio de la variable sobre la medida de variabilidad considerada.

EJEMPLO DE COEFICIENTE DE VARIACIÓN

La media y la desviación estándar de dos stocks de empresas de electrodomésticos en el año actual han sido:

| EMPRESA | PRECIO MEDIO | VARIANZA |
|-----------|--------------|----------|
| INDESIT | 9,87 | 1,34 |
| DE LONGHI | 3,03 | 0,32 |

¿Qué stock debería comprar el/la inversora si está interesada en comprar el stock con menos volatilidad?


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

$$CV = \frac{\sigma}{\mu} \cdot 100$$



EJEMPLO DE COEFICIENTE DE VARIACIÓN

La media y la desviación estándar de dos stocks de empresas de electrodomésticos en el año actual han sido:


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

$$CV = \frac{\sigma}{\mu} \cdot 100$$

| EMPRESA | PRECIO MEDIO | VARIANZA |
|-----------|--------------|----------|
| INDESIT | 9,87 | 1,34 |
| DE LONGHI | 3,03 | 0,32 |

¿Qué stock debería comprar el/la inversora si está interesada en comprar el stock con menos volatilidad?

+

| EMPRESA | PRECIO MEDIO | VARIANZA | DESVIACIÓN ESTÁNDAR | COEFICIENTE DE VARIACIÓN |
|-----------|--------------|----------|---------------------|--------------------------|
| INDESIT | 9,87 | 1,34 | 1,15 | 0,12 o 12% |
| DE LONGHI | 3,03 | 0,32 | 0,56 | 0,19 o 19% |

Indesit tiene la menor volatilidad en sus precios

Rango o recorrido $R = \max - \min$

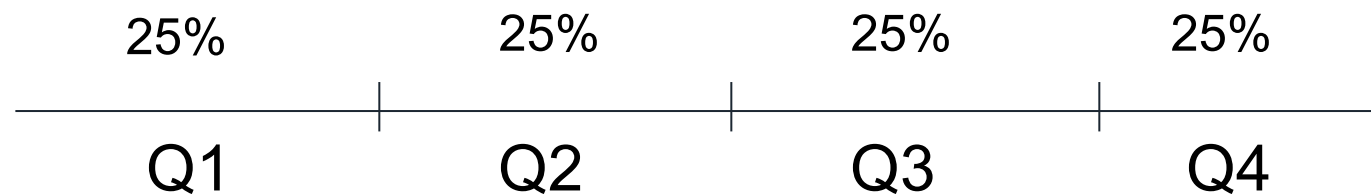


Máximo valor de los datos - mínimo valor de los datos

Problema: sensible a valores extremos

Ejemplo de uso: tiempo de reacción mínima y máxima de una persona a un medicamento.

Rango intercuartílico

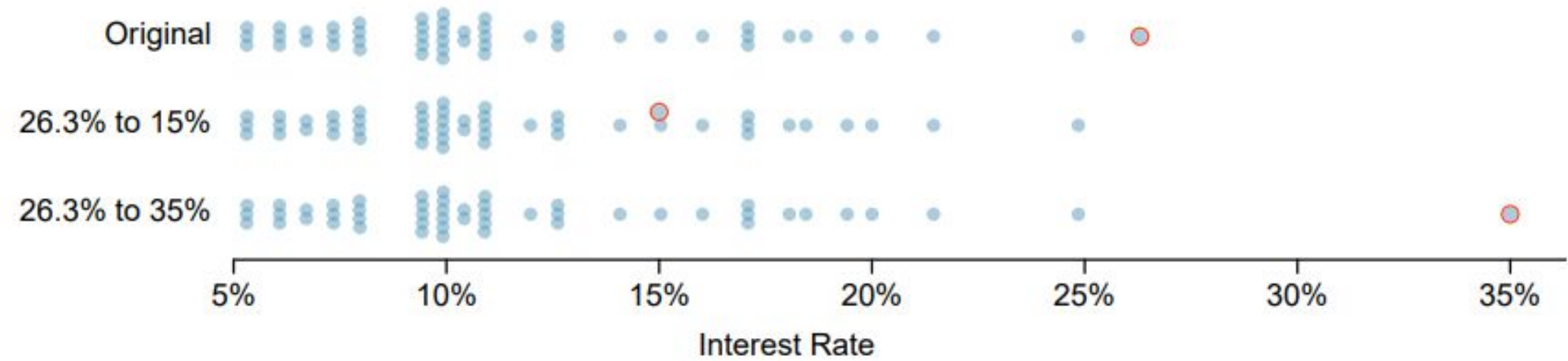


$\text{IQR} = Q3 - Q1$



Por ejemplo: al analizar la velocidad de distintos corredores de una maratón, encontramos que las personas más “promedio” se encuentran entre el 2do y 3er cuartil. Queremos medir la diferencia de velocidad en este grupo, deshaciéndonos de los casos que corren o muy lento o muy rápido

Robustez frente a outliers: el IQR y la mediana



| scenario | robust | | not robust | |
|-----------------------------|--------|-------|------------|-------|
| | median | IQR | \bar{x} | s |
| original interest_rate data | 9.93% | 5.76% | 11.57% | 5.05% |
| move 26.3% → 15% | 9.93% | 5.76% | 11.34% | 4.61% |
| move 26.3% → 35% | 9.93% | 5.76% | 11.74% | 5.68% |

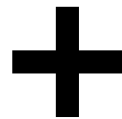


Denominamos estadísticos robustos al IQR y la mediana porque se ven poco afectadas por obs. extremas. Por el contrario, la desvest y la media son más sensibles a ellas.



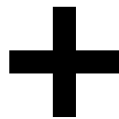
1. Estadística

- A) Terminologías y definiciones
- B) Distribuciones de Frecuencia
- C) Medidas de Tendencia Central
- D) Variabilidad**
- E) Distribución Bivariada
- F) Dependencia de la media
- G) Regresión Lineal Simple
- H) Correlación



1. Estadística

- A) Terminologías y definiciones
- B) Distribuciones de Frecuencia
- C) Medidas de Tendencia Central
- D) Variabilidad
- E) Análisis bivariado**
- F) Dependencia de la media
- G) Regresión Lineal Simple
- H) Correlación





ANÁLISIS BIVARIADO

DATOS BIVARIADOS

Los datos bivariados son una secuencia de **pares de valores** de **dos variables** observadas a lo largo del conjunto de unidades estadísticas.

Cuando el número de observaciones es grande, los datos bivariados suelen arreglarse en la forma de una **tabla de contingencia**.

Se trata de una tabla de **dos entradas** donde los **s** valores de la variable X se enlistan encolumnadamente, mientras los valores **t** de la variable Y se enlistan a lo largo de las filas. El resultado es una tabla con celdas de valores **$s.t$**

En los márgenes se encuentra la suma de casos que caen en el valor X_{s0} (última columna) y los que caen en el valor Y_{0t} (última fila). La intersección entre la suma de todos los valores de X_{s0} y Y_{0t} es el total de las unidades de análisis (N)

DATOS BIVARIADOS

| Variable <i>X</i> | Variable <i>Y</i> | | | | | | Total |
|------------------------------|-------------------------------|-------------------------------|-----|--------------------------------|-----|--------------------------------|-------------------------------|
| | <i>y</i> ₁ | <i>y</i> ₂ | ... | <i>y</i> _{<i>j</i>} | ... | <i>y</i> _{<i>t</i>} | |
| <i>X</i> ₁ | <i>n</i> ₁₁ | <i>n</i> ₁₂ | ... | <i>n</i> _{1<i>j</i>} | ... | <i>n</i> _{1<i>t</i>} | <i>n</i> ₁₀ |
| <i>X</i> ₂ | <i>n</i> ₂₁ | <i>n</i> ₂₂ | ... | <i>n</i> _{2<i>j</i>} | ... | <i>n</i> _{2<i>t</i>} | <i>n</i> ₂₀ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| <i>X</i> _{<i>l</i>} | <i>n</i> _{<i>l</i>1} | <i>n</i> _{<i>l</i>2} | ... | <i>n</i> _{<i>l j</i>} | ... | <i>n</i> _{<i>l t</i>} | <i>n</i> _{<i>l</i>0} |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| <i>X</i> _{<i>s</i>} | <i>n</i> _{<i>s</i>1} | <i>n</i> _{<i>s</i>2} | ... | <i>n</i> _{<i>s j</i>} | ... | <i>n</i> _{<i>s t</i>} | <i>n</i> _{<i>s</i>0} |
| Total | <i>n</i> ₀₁ | <i>n</i> ₀₂ | ... | <i>n</i> _{0<i>j</i>} | ... | <i>n</i> _{0<i>t</i>} | <i>N</i> |

$$\sum_{i=1}^s n_{i0} = \sum_{j=1}^t n_{0j} = N$$



EJEMPLO

| CANTIDAD DE INTEGRANTES POR HOGAR (X) | CANTIDAD DE HABITACIONES POR HOGAR (Y) | | | | | TOTAL |
|---------------------------------------|--|----|---|---|---|-------|
| | 2 | 3 | 4 | 5 | 6 | |
| 1 | 2 | 3 | 0 | 0 | 0 | 5 |
| 2 | 0 | 4 | 3 | 0 | 1 | 8 |
| 3 | 0 | 2 | 2 | 4 | 0 | 8 |
| 4 | 0 | 1 | 0 | 3 | 1 | 5 |
| 5 | 0 | 0 | 0 | 1 | 1 | 2 |
| TOTAL | 2 | 10 | 5 | 8 | 3 | 28 |

→ 5 casos tienen 1 integrante

→ 8 casos tienen 2 integrantes

→ 8 casos tienen 3 integrantes

→ 5 casos tienen 4 integrantes

→ 2 casos tienen 5 integrantes

→ 28 casos en total

2 casos
2 hab.

10 casos
3 hab.

5 casos
4 hab.

8 casos
5 hab.

3 casos
6 hab.

28 casos
en total

$$\sum_{i=1}^s n_{i0} = \sum_{j=1}^t n_{0j} = N$$

DISTRIBUCIONES MARGINALES

Como se mencionó, a los márgenes de la tabla de contingencia, se ubican los casos que caen en cada valor X_{s0} (última columna) y los que caen en cada valor Y_{0t} (última fila).

Como pueden suponer, se trata de la distribución de frecuencia de cada variable. En la tabla se las reconoce como LA DISTRIBUCIÓN MARGINAL de X o Y en cada caso. En la última columna se encuentra la distribución marginal de X, mientras que en la última fila se encuentra la distribución marginal de Y

| CANTIDAD DE INTEGRANTES POR HOGAR (X) | Distribución de Frecuencias de X |
|---------------------------------------|----------------------------------|
| | TOTAL |
| 1 | 5 |
| 2 | 8 |
| 3 | 8 |
| 4 | 5 |
| 5 | 2 |
| | 28 |

| CANTIDAD DE HABITACIONES POR HOGAR (Y) | Distribución de Frecuencias de Y |
|--|----------------------------------|
| | TOTAL |
| 2 | 2 |
| 3 | 10 |
| 4 | 5 |
| 5 | 8 |
| 6 | 3 |
| | 28 |

DISTRIBUCIÓN DE FRECUENCIAS CONDICIONALES PARA X e Y

Cada valor que asume X o Y dentro de la tabla de contingencia, se lo llama ***distribución de frecuencia condicional***.

Ahora bien, la misma varía si es en función de X o Y, por lo tanto, las frecuencias consignadas **en las filas** con un valor x_i estático son las **distribuciones de frecuencia condicional de Y**, en tanto que para un valor de X específico (condicional) se mira la distribución de cada valor de Y.

A la inversa,, las frecuencias consignadas en las columnas con un valor y_j estático son las **distribuciones de frecuencia condicional de X**, en tanto que para un valor de Y específico (condicional) se mira la distribución de cada valor de X

EJEMPLO - DISTRIBUCIÓN DE FRECUENCIAS CONDICIONALES PARA X e Y

| CANTIDAD DE INTEGRANTES POR HOGAR (X) | CANTIDAD DE HABITACIONES POR HOGAR (Y) | | | | | TOTAL |
|---------------------------------------|--|----|---|---|---|-------|
| | 2 | 3 | 4 | 5 | 6 | |
| 1 | 2 | 3 | 0 | 0 | 0 | 5 |
| 2 | 0 | 4 | 3 | 0 | 1 | 8 |
| 3 | 0 | 2 | 2 | 4 | 0 | 8 |
| 4 | 0 | 1 | 0 | 3 | 1 | 5 |
| 5 | 0 | 0 | 0 | 1 | 1 | 2 |
| TOTAL | 2 | 10 | 5 | 8 | 3 | 28 |

Distribución de frecuencia condicional del número de habitaciones para un hogar con 3 integrantes → **DISTRIBUCIÓN DE FRECUENCIA CONDICIONAL DE Y**

Distribución de frecuencia condicional de la cantidad de integrantes por hogar para un número de habitaciones igual a 6 → **DISTRIBUCIÓN DE FRECUENCIA CONDICIONAL DE X**

| CANTIDAD DE INTEGRANTES (X) | CANTIDAD DE HABITACIONES (Y=6) |
|-----------------------------|--------------------------------|
| 1 | 0 |
| 2 | 1 |
| 3 | 0 |
| 4 | 1 |
| 5 | 1 |
| TOTAL | 3 |

| CANTIDAD DE INTEGRANTES | CANTIDAD DE HABITACIONES (Y) | | | | | TOTAL |
|-------------------------|------------------------------|---|---|---|---|-------|
| | | 3 | 4 | 5 | 6 | |
| 3 | 0 | 2 | 2 | 4 | 0 | 8 |

DISTRIBUCIÓN DE FRECUENCIAS RELATIVAS
CONDICIONALES PARA X E Y

De la misma manera que la distribución de frecuencias relativas para una única variable, es posible obtener la frecuencia relativa condicional de una tabla de contingencias a partir de la división de la frecuencia observada por el total de los casos observados (N)

| CANTIDAD DE INTEGRANTES | CANTIDAD DE HABITACIONES | | | | | TOTAL |
|-------------------------|--------------------------|------|------|------|------|-------|
| | 2 | 3 | 4 | 5 | 6 | |
| 1 | 0,07 | 0,11 | 0 | 0 | 0 | 0,18 |
| 2 | 0 | 0,14 | 0,11 | 0 | 0,04 | 0,28 |
| 3 | 0 | 0,07 | 0,07 | 0,14 | 0 | 0,28 |
| 4 | 0 | 0,04 | 0 | 0,11 | 0,04 | 0,18 |
| 5 | 0 | 0 | 0 | 0,04 | 0,04 | 0,08 |
| TOTAL | 0,07 | 0,36 | 0,18 | 0,28 | 0,11 | 1 |

DISTRIBUCIÓN DE FRECUENCIAS RELATIVAS
CONDICIONALES PARA X (EN FUNCIÓN DE FILAS)

En el siguiente caso, obtenemos la distribución condicional de frecuencias relativas para cada fila (X) a partir de la división entre el valor de la frecuencia y el total de la columna marginal que corresponde a esa fila

| CANTIDAD DE INTEGRANTES | CANTIDAD DE HABITACIONES | | | | | TOTAL |
|-------------------------|--------------------------|-------|-------|-------|-------|-------|
| | 2 | 3 | 4 | 5 | 6 | |
| 1 | 0,400 | 0,600 | 0,000 | 0,000 | 0,000 | 1,00 |
| 2 | 0,000 | 0,500 | 0,375 | 0,000 | 0,125 | 1,00 |
| 3 | 0,000 | 0,250 | 0,250 | 0,500 | 0,000 | 1,00 |
| 4 | 0,000 | 0,200 | 0,000 | 0,600 | 0,200 | 1,00 |
| 5 | 0,000 | 0,000 | 0,000 | 0,500 | 0,500 | 1,00 |
| TOTAL | 0,07 | 0,36 | 0,18 | 0,28 | 0,11 | 1 |

DISTRIBUCIÓN DE FRECUENCIAS RELATIVAS
CONDICIONALES PARA Y (EN FUNCIÓN DE COLUMNAS)

En el siguiente caso, obtenemos la distribución condicional de frecuencias relativas para cada columna (Y) a partir de la división entre el valor de la frecuencia y el total de la columna marginal que corresponde a esa fila

| CANTIDAD DE INTEGRANTES | CANTIDAD DE HABITACIONES | | | | | TOTAL |
|-------------------------|--------------------------|-------|-------|-------|-------|-------|
| | 2 | 3 | 4 | 5 | 6 | |
| 1 | 1,000 | 0,300 | 0,000 | 0,000 | 0,000 | 0,18 |
| 2 | 0,000 | 0,400 | 0,600 | 0,000 | 0,333 | 0,28 |
| 3 | 0,000 | 0,200 | 0,400 | 0,500 | 0,000 | 0,28 |
| 4 | 0,000 | 0,100 | 0,000 | 0,375 | 0,333 | 0,18 |
| 5 | 0,000 | 0,000 | 0,000 | 0,125 | 0,333 | 0,08 |
| TOTAL | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1 |

INDEPENDENCIA ENTRE VARIABLES

Este concepto es la clave para definir la asociación entre variables. Dada una tabla de contingencia, decimos que las variables X e Y son **estadísticamente independientes** si las distribuciones de frecuencias relativas condicionales son todas iguales (e iguales a la marginal)

Si esto no ocurre, entonces las variables están **asociadas**

| CANTIDAD DE INTEGRANTES | CANTIDAD DE HABITACIONES | | | | | TOTAL |
|-------------------------|--------------------------|-------|-------|-------|-------|-------|
| | 2 | 3 | 4 | 5 | 6 | |
| 1 | 1,000 | 0,300 | 0,000 | 0,000 | 0,000 | 0,18 |
| 2 | 0,000 | 0,400 | 0,600 | 0,000 | 0,333 | 0,28 |
| 3 | 0,000 | 0,200 | 0,400 | 0,500 | 0,000 | 0,28 |
| 4 | 0,000 | 0,100 | 0,000 | 0,375 | 0,333 | 0,18 |
| 5 | 0,000 | 0,000 | 0,000 | 0,125 | 0,333 | 0,08 |
| TOTAL | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1 |

(NO) INDEPENDENCIA ENTRE VARIABLES

Ejemplo para plantear una hipótesis

| | <i>Percentage Who Drop Out of High School</i> | <i>Percentage Who Complete High School</i> |
|---|---|--|
| Teen mothers | 70% | 30% |
| Women who delay parenthood until 21 or older | 24% | 76% |

SOURCE: Adapted from Berglas et al. (2003, p. 24).

¿Qué hipótesis está estudiando esta tabla de contingencia?

+ ¿Qué cuadrantes se comparan para evaluarla?

Cuando queremos ver la asociación entre variables es fundamental considerar el tipo de variables involucradas.

Para variables **cualitativas** son medidas de asociación comunes:


- Chi cuadrado. No estandarizado
- Phi (sólo para tablas 2x2). Estandarizado
- V de cramer. Estandarizado

(En cada caso indican magnitud y significatividad).



Cuando queremos ver la asociación entre variables es fundamental considerar el tipo de variables involucradas.

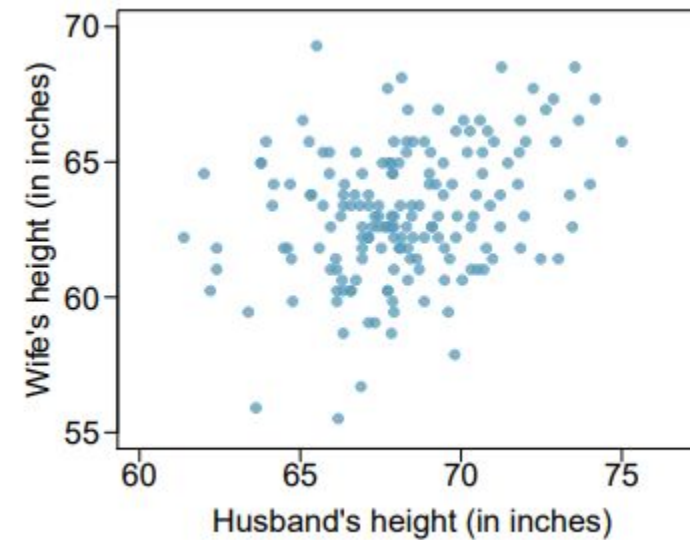
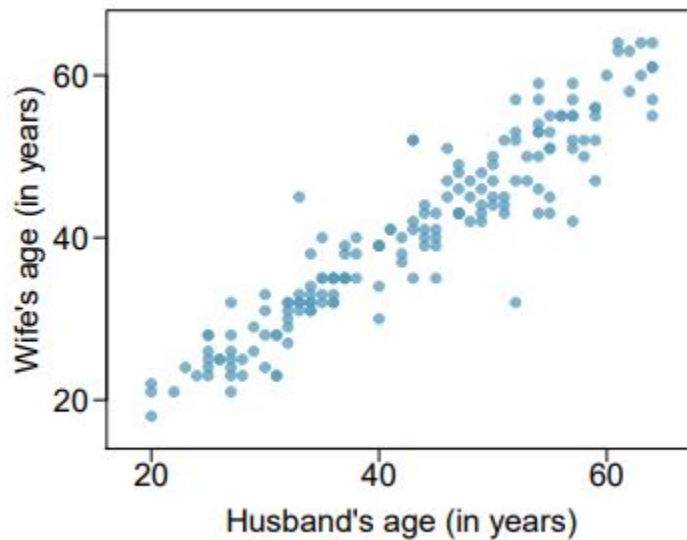
Para variables **cuantitativas** en general se usa el coeficiente de correlación de Pearson:

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$


En el numerador tenemos la covarianza entre x e y. El denominador tiene ambas desviaciones estándar para normalizar el coeficiente y que siempre esté entre -1 y 1

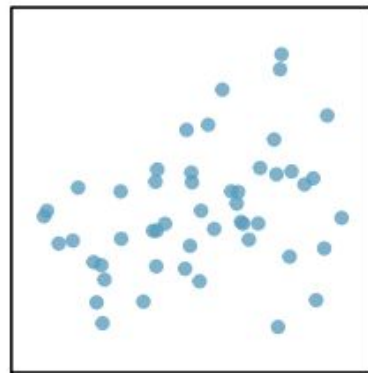
+ Permite identificar magnitud, dirección y significatividad.

Visualización de la relación entre 2 variables numéricas: scatterplots

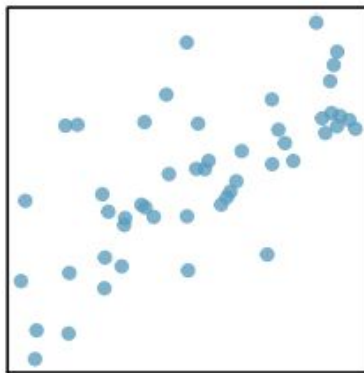


+ ¿Qué par de variables están más relacionadas?
¿Cómo es esa relación?

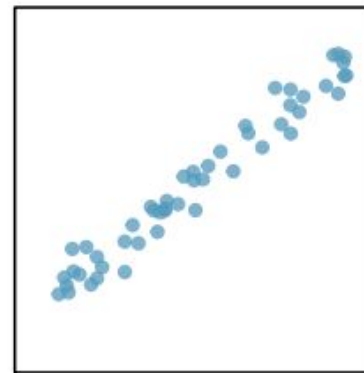
Visualización de la relación entre 2 variables numéricas: scatterplots para distintos r



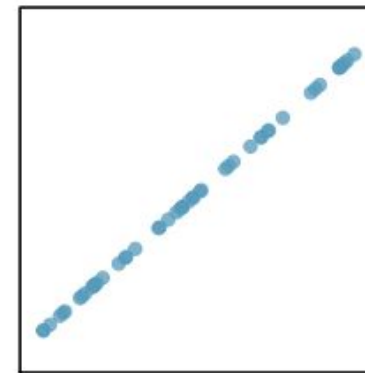
$R = 0.33$



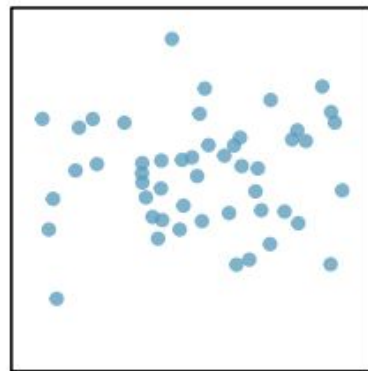
$R = 0.69$



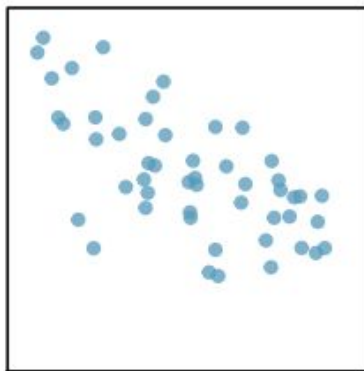
$R = 0.98$



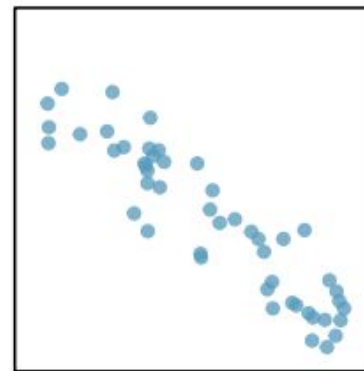
$R = 1.00$



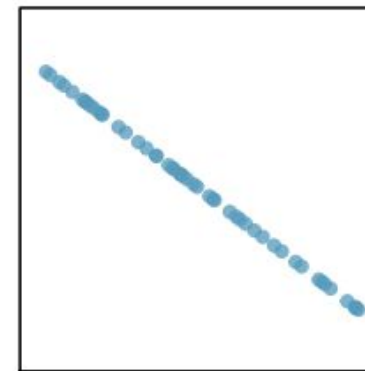
$R = 0.08$



$R = -0.64$



$R = -0.92$



$R = -1.00$





- A) Terminologías y definiciones
- B) Variables Aleatorias
- C) Distribuciones de Probabilidad
- D) Uniones de Distribuciones de Probabilidad

