Introducción al análisis cuantitativo de datos lingüísticos

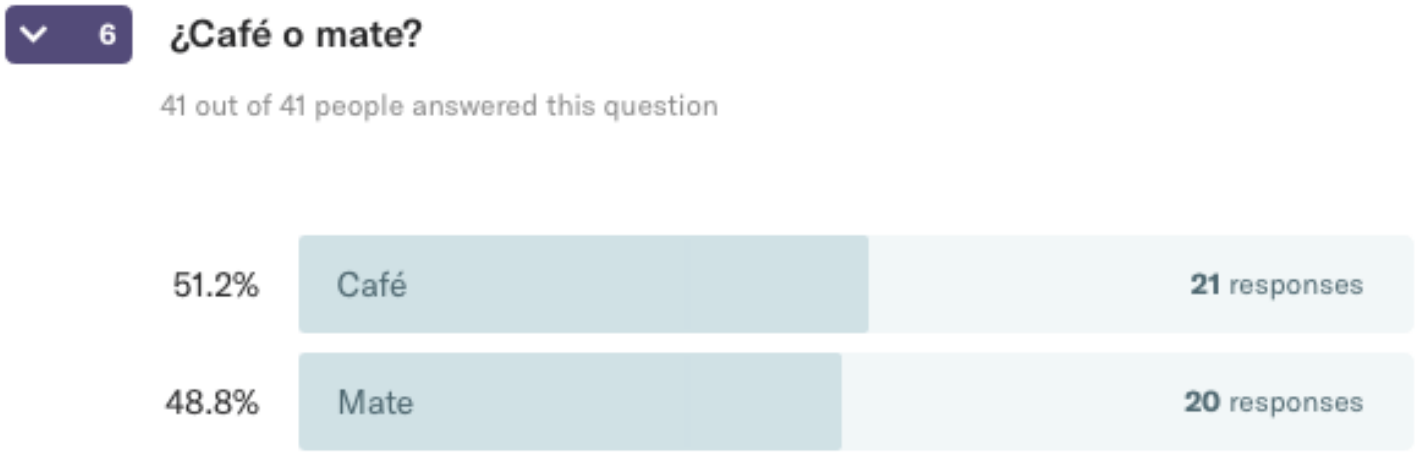
Bloque 1.3:
Tipos de variables y sus características

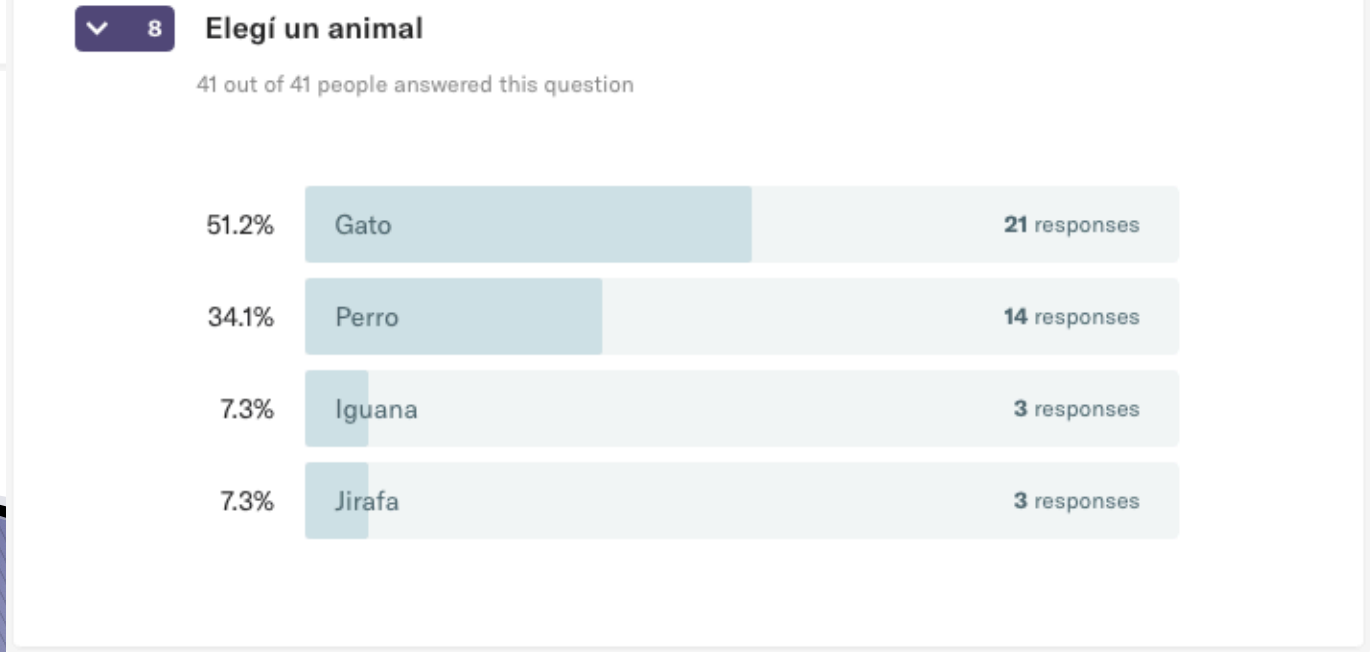
Ezequiel Koile (MPI-SSH)
Carolina Gattei (IFIBA - CONICET)

Tipos de variables

- Utilizamos *variables* para describir nuestros datos. Estas vienen en distintos tipos:

Tipos de variables

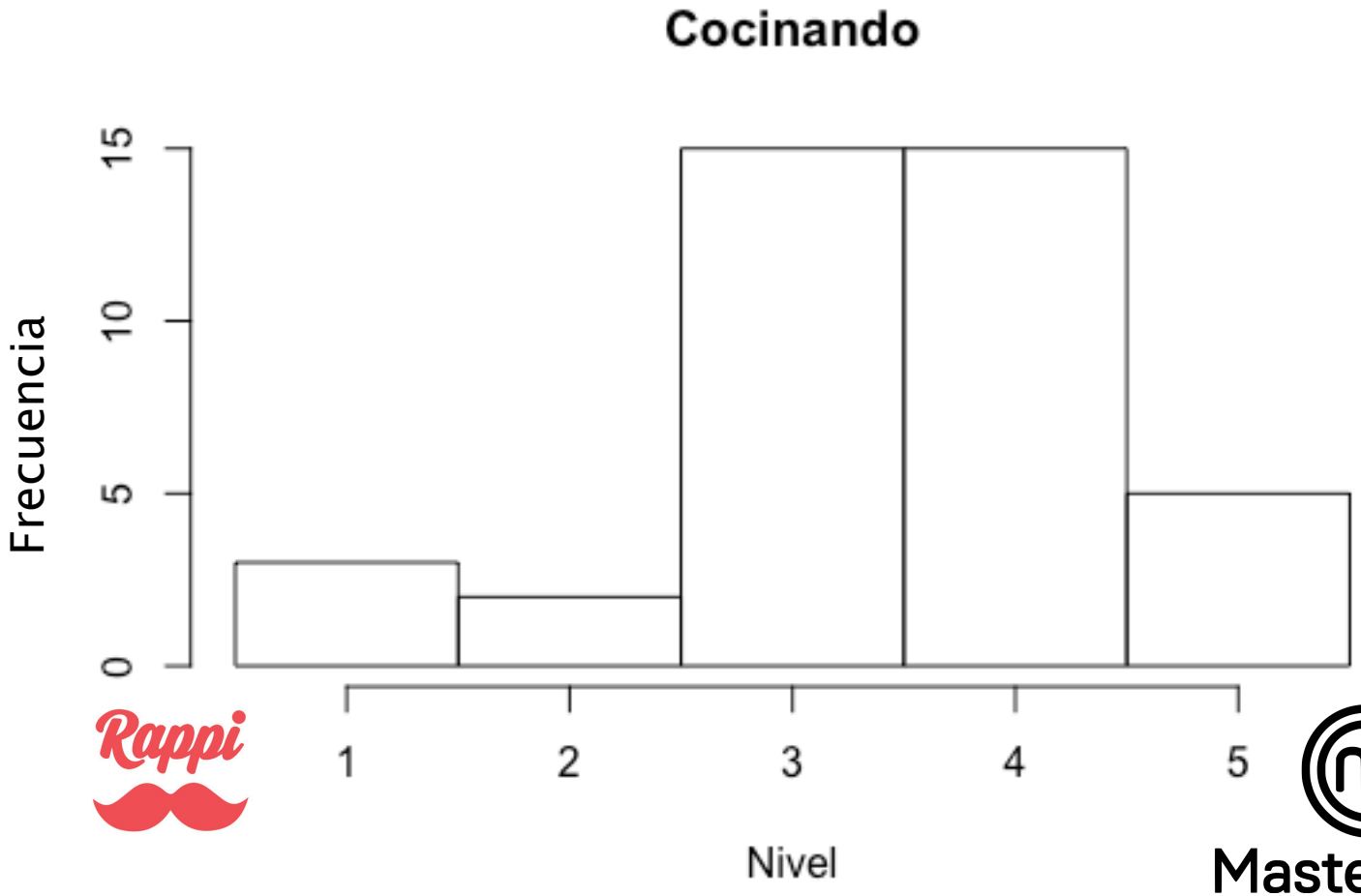




Tipos de variables

- Utilizamos *variables* para describir nuestros datos. Estas vienen en distintos tipos:
- **Variables nominales**: Dos o más *categorías,* mutuamente excluyentes. Si son solamente dos, estamos en presencia de una *variable binaria.*

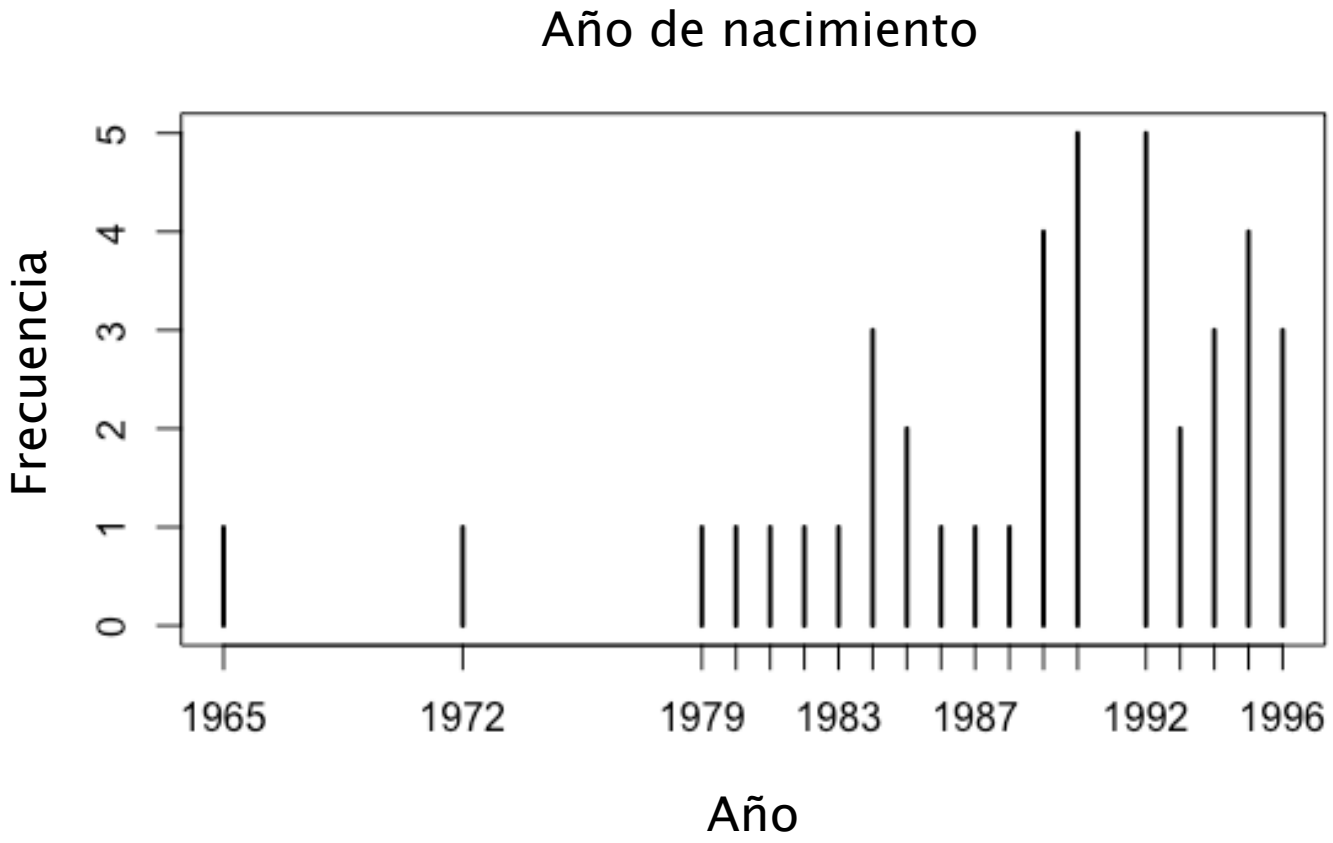
Tipos de variables



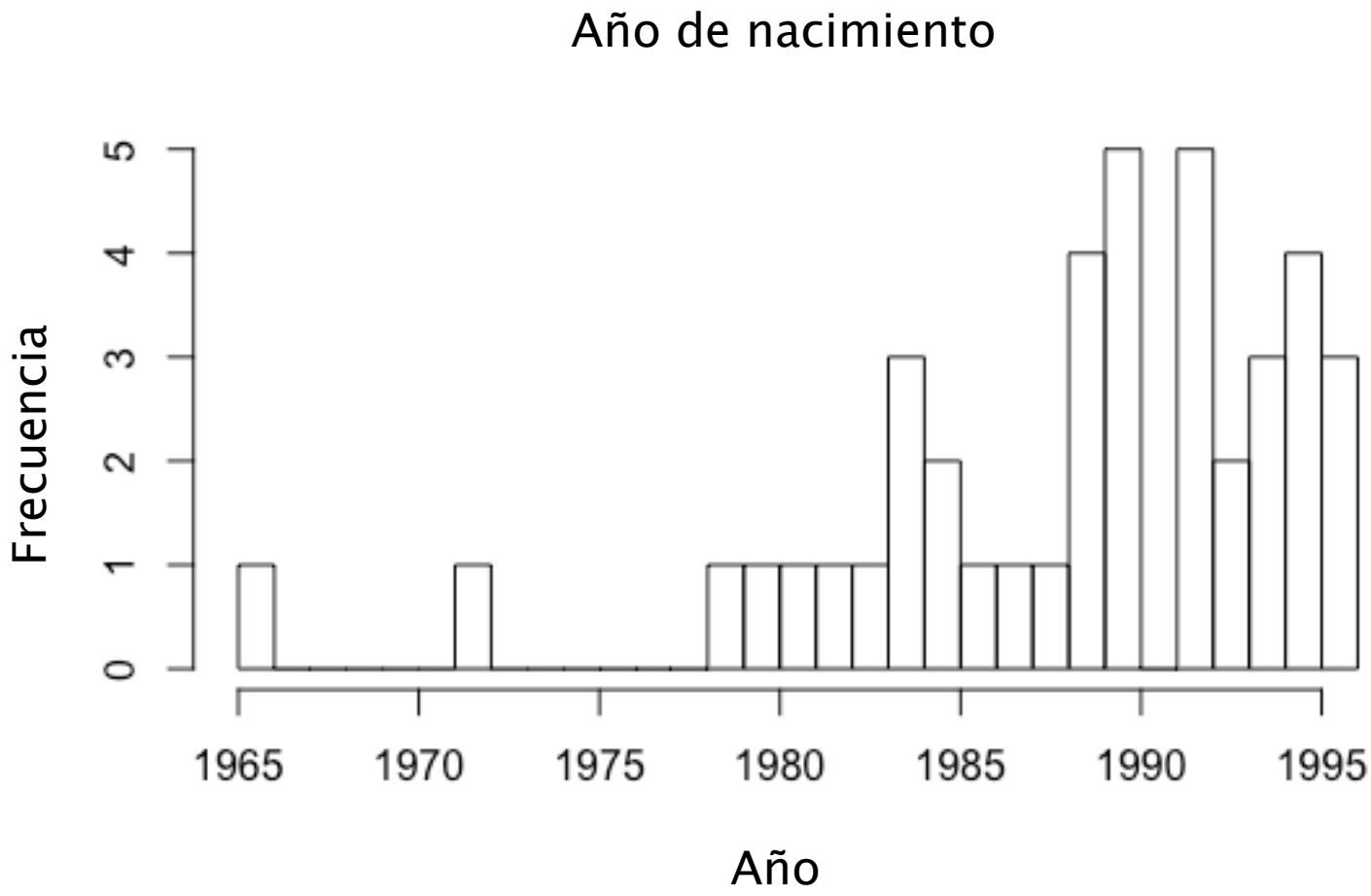
Tipos de variables

- Utilizamos *variables* para describir nuestros datos. Estas vienen en distintos tipos:
- **Variables nominales**: Dos o más *categorías,* mutuamente excluyentes. Si son solamente dos, estamos en presencia de una *variable binaria.*
- **Variables ordinales**: También son categorías, pero estas están ordenadas. No podemos asegurar que la diferencia entre dos categorías consecutivas sea la misma (¿la distancia de 3 a 4 es la misma que la de 4 a 5?

Tipos de variables



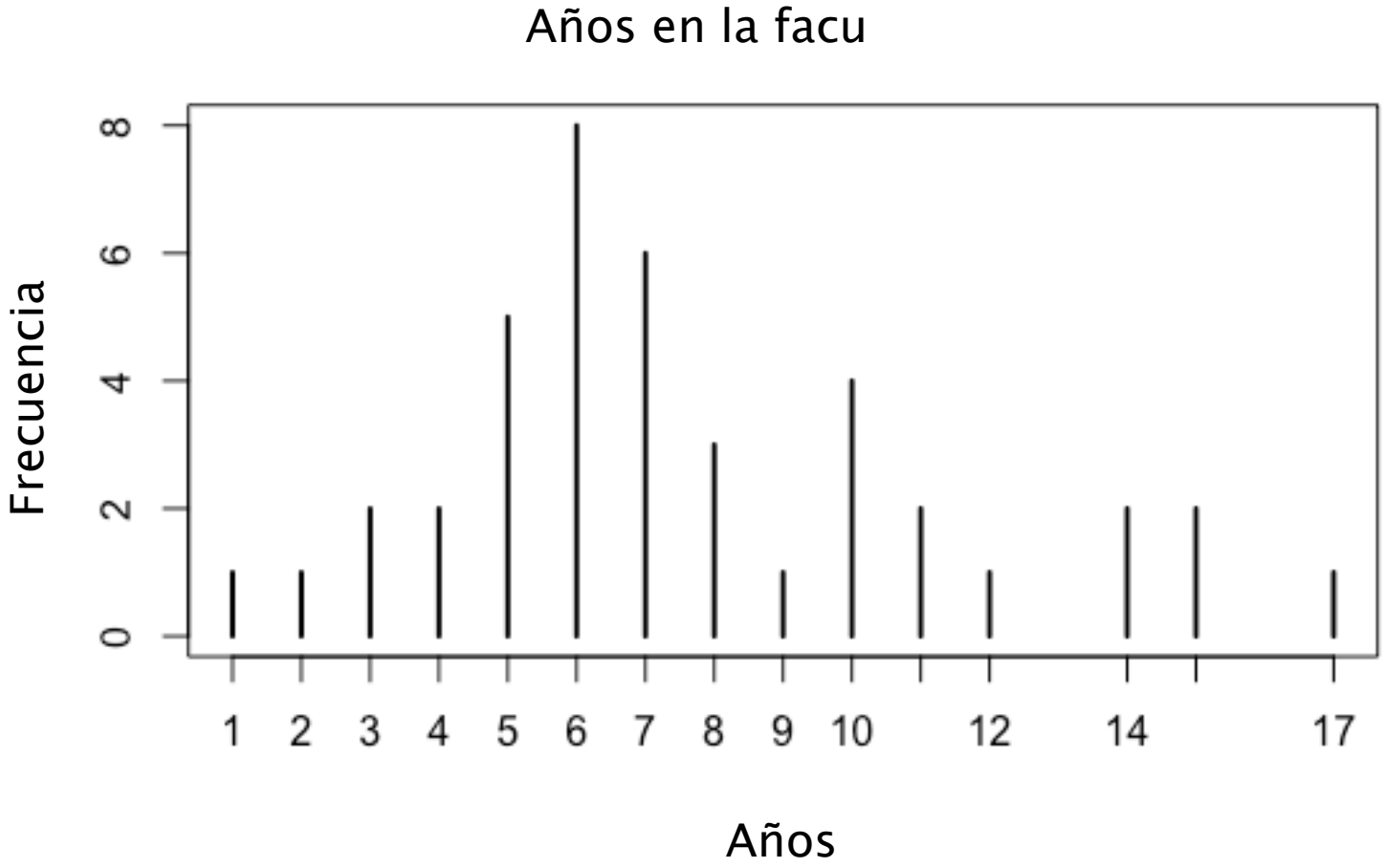
Tipos de variables



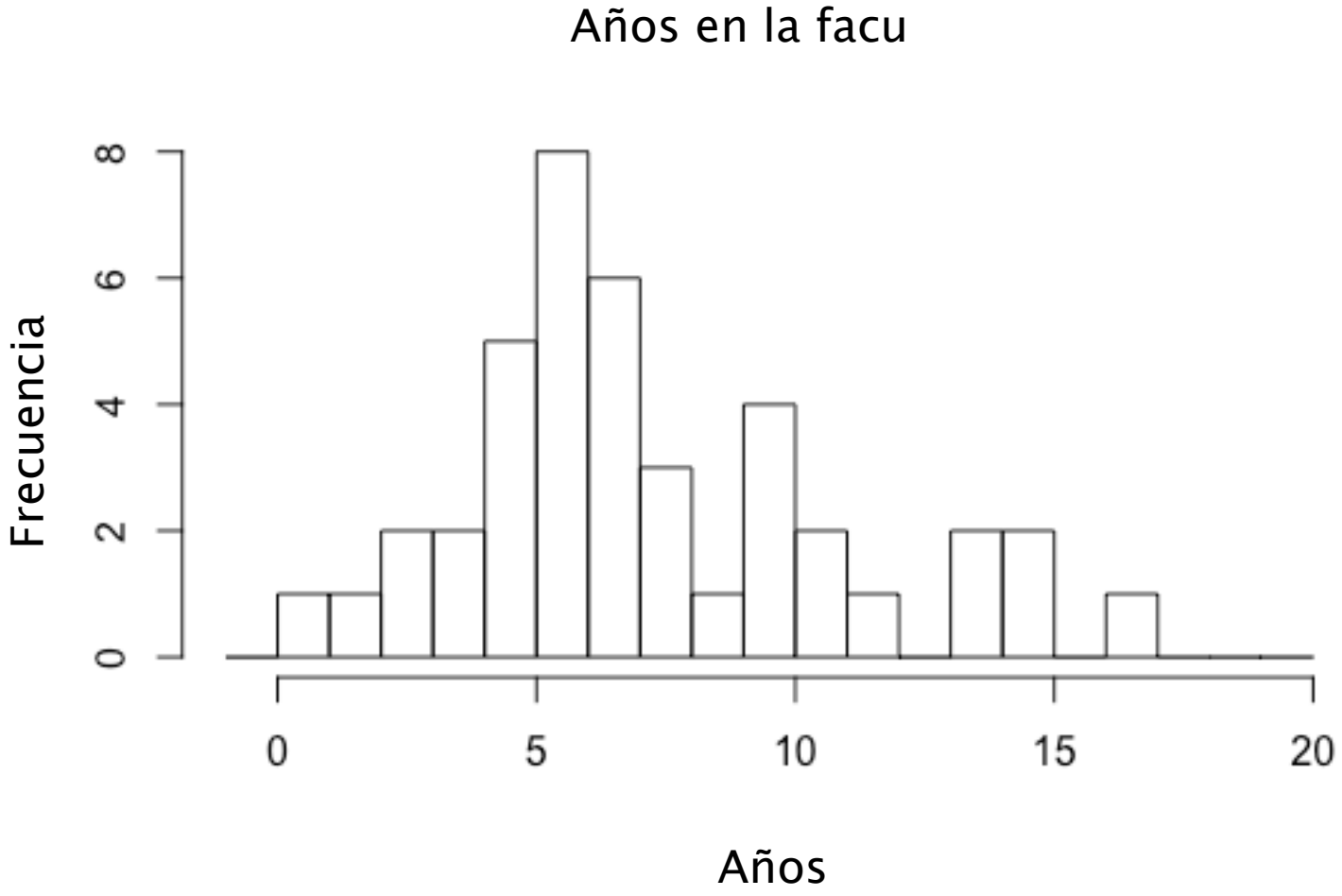
Tipos de variables

- Utilizamos *variables* para describir nuestros datos. Estas vienen en distintos tipos:
- **Variables nominales**: Dos o más *categorías,* mutuamente excluyentes. Si son solamente dos, estamos en presencia de una *variable binaria.*
- **Variables ordinales**: **También son categorías, pero estas están ordenadas. No podemos asegurar que la diferencia entre dos categorías consecutivas sea la misma (¿la distancia de 3 a 4 es la misma que la de 4 a 5?**
- **Intervalo**: Intervalos iguales en la escala representan diferencias iguales entre los puntos de la escala. *No incluyen el valor cero* (o, si lo incluyen, este es arbitrario).

Tipos de variables



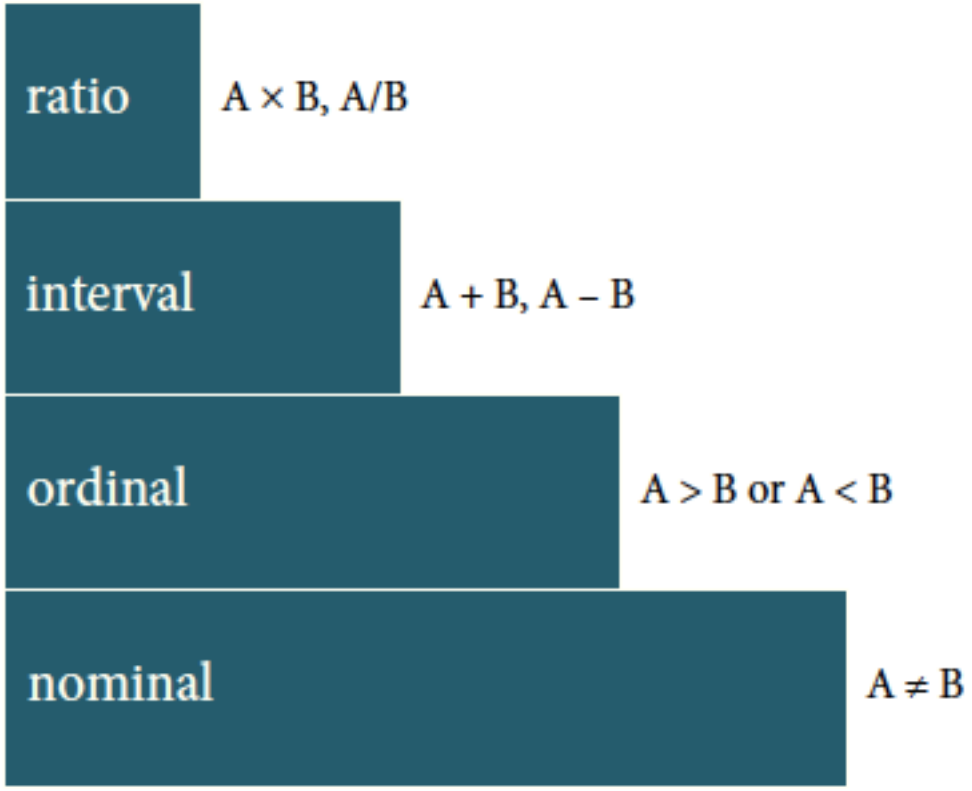
Tipos de variables



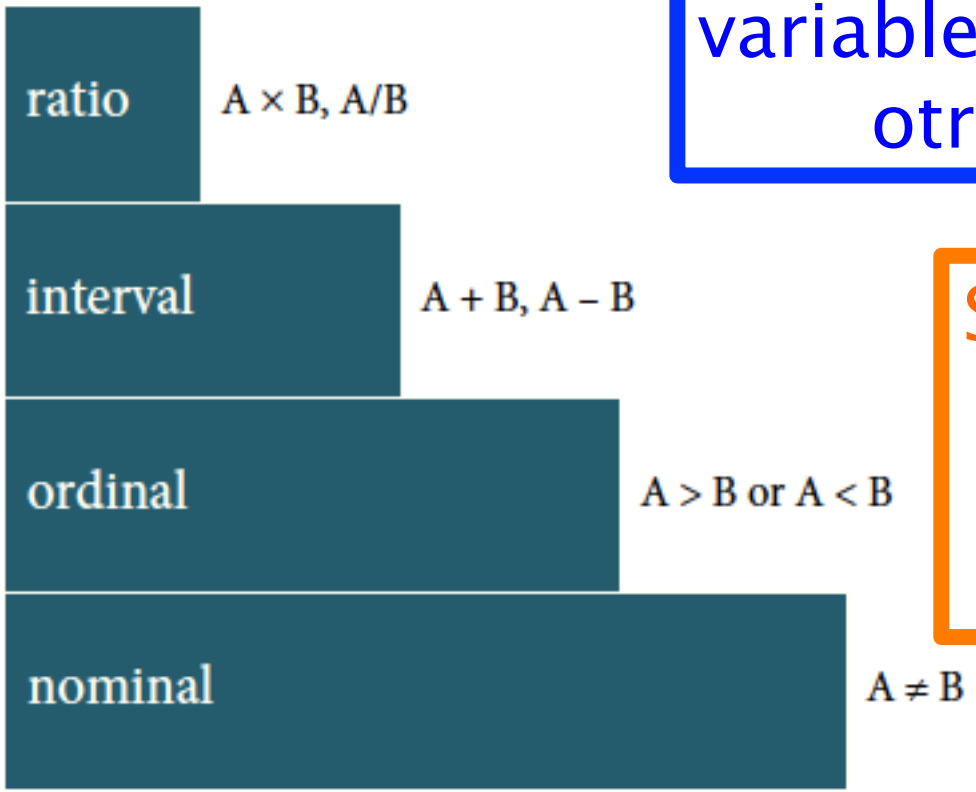
Tipos de variables

- Utilizamos *variables* para describir nuestros datos. Estas vienen en distintos tipos:
- **Variables nominales**: Dos o más *categorías,* mutuamente excluyentes. Si son solamente dos, estamos en presencia de una *variable binaria.*
- **Variables ordinales**: **También son categorías, pero estas están ordenadas. No podemos asegurar que la diferencia entre dos categorías consecutivas sea la misma (¿la distancia de 3 a 4 es la misma que la de 4 a 5?**
- **Intervalo**: Intervalos iguales en la escala representan diferencias iguales entre los puntos de la escala. *No incluyen el valor cero* (o, si lo incluyen, este es arbitrario).
- **Ratio**: Como las variables intervalo, pero *aquí sí hay un valor cero significativo:* los cocientes son relevantes (4 es el doble de 2).

Tipos de variables



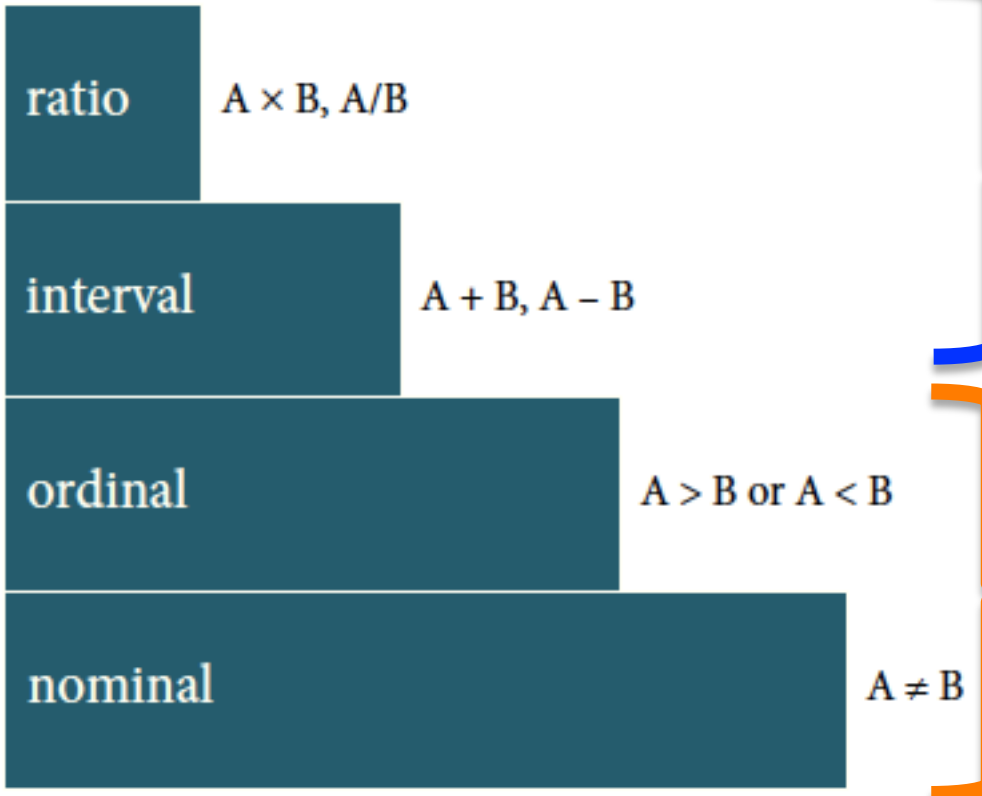
Tipos de variables



¿Podemos transformar variables de un tipo en otro? ¿Cómo?

Solo hacia abajo (¡y siempre perderemos información!)

Tipos de variables



Cuantitativas
o
Numéricas

Cualitativas
o
Categóricas

Medidas de tendencia central

- Queremos conocer “el valor más típico” para las alturas de un grupo de *N* personas. ¿Qué valor tomamos?
  1. Sumamos todas las alturas y las dividimos por *N.*
  2. Las multiplicamos y calculamos la raíz *N*–ésima.
  3. Promediamos los cuadrados de cada altura y le tomamos la raíz cuadrada al resultado.
  4. Tomamos el valor que más se repite
  5. Sumamos el valor más alto y el más bajo y los dividimos por 2.
  6. Promediamos las inversas de las alturas e invertimos el resultado.
  7. Tomamos el valor que deja a la mitad de los demás valores por encima y a la otra mitad por debajo.
  8. Promediamos todos los valores excepto el10% más alto y el 10% más bajo.
  9. Promediamos todos los valores en la mitad central de la distribución de alturas.
  10. Promediamos dos valores: el que deja al 75% de los demás por debajo, y el que deja al 25% de los demás por debajo.
  11. Igual que el caso anterior, pero agregamos el valor medio al promedio (dos veces).

Medidas de tendencia central

- Queremos conocer “el valor más típico” para las alturas de un grupo de *N* personas. ¿Qué valor tomamos?
  1. Sumamos todas las alturas y las dividimos por *N*.
  2. Las multiplicamos y calculamos la raíz *N*–ésima.
  3. Promediamos los cuadrados de cada altura y le tomamos la raíz cuadrada al resultado.
  4. Tomamos el valor que más se repite
  5. Sumamos el valor más alto y el más bajo y los dividimos por 2.
  6. Promediamos las inversas de las alturas e invertimos el resultado.
  7. Tomamos el valor que deja a la mitad de los demás valores por encima y a la otra mitad por debajo.
  8. Promediamos todos los valores excepto el10% más alto y el 10% más bajo.
  9. Promediamos todos los valores en la mitad central de la distribución de alturas.
  10. Promediamos dos valores: el que deja al 75% de los demás por debajo, y el que deja al 25% de los demás por debajo.
  11. Igual que el caso anterior, pero agregamos el valor medio al promedio (dos veces).

tipo Media

tipo Moda

tipo Mediana

1. Media (promedio)

- Media o promedio aritmético
- Se suman los valores y se divide por la cantidad de estos
- Pros:
  - Usa todas las observaciones
  - Las desviaciones suman cero (igual dispersión a izquierda y a derecha)
  - La suma de las desviaciones cuadráticas es mínima
  - Muestras tomadas de la misma población suelen tener promedios similares
  - Yapa: Todo el mundo sabe de qué se trata, es fácil de calcular, aparece naturalmente en cálculos estadísticos más avanzados, etc.
- Contras:
  - Demasiado sensible a *outliers*
- *Definida solamente para variables numéricas* (ratio o intervalo)

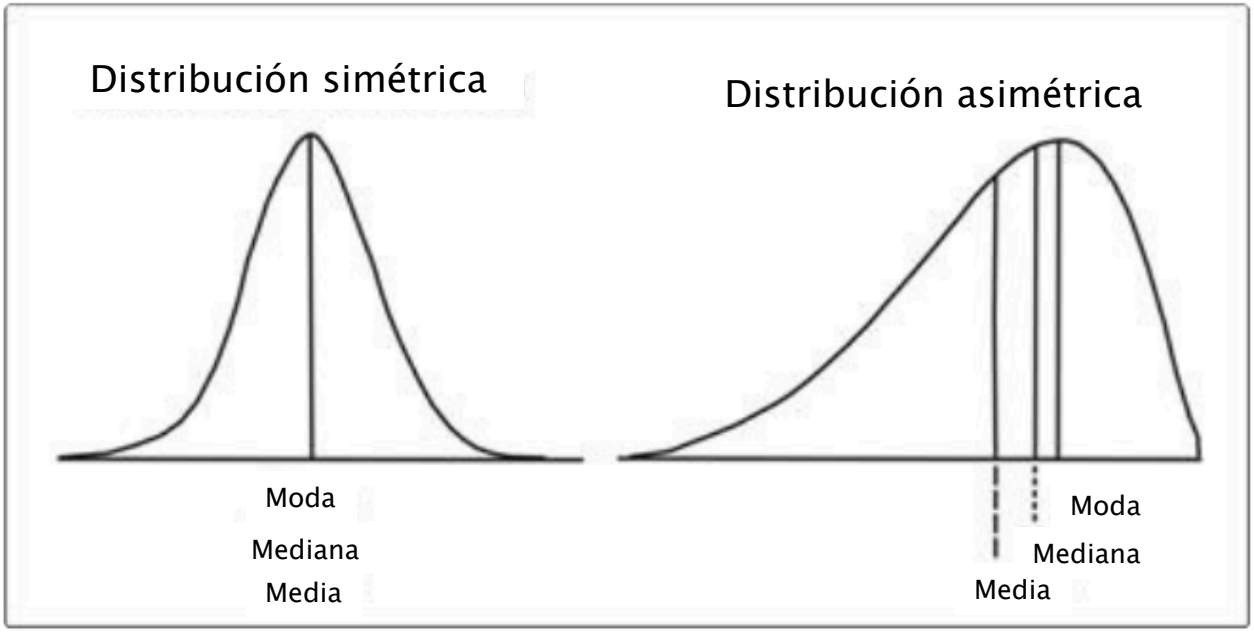
2. Mediana

- Es el valor “en el medio” si ordenamos todas las observaciones en orden ascendente: un 50% de las demás observaciones será mayor, y el otro 50% será menor.
- Pros:
  - Robustez (no es sensible a *outliers*)
- Contras:
  - No toma en cuenta todas las observaciones
  - La mediana de dos grupos combinados no puede expresarse en términos de las medias individuales de cada grupo.
- *Definida para variables numéricas* (ratio o intervalo) *y ordinales*

3. Moda

- Es el valor más frecuente de una variable
- Pros:
  - Puede usarse para variables nominales
- Contras:
  - No siempre representa “el valor más típico”
  - Puede no ser única (distribuciones bimodal y multimodal)
  - No está algebraicamente definida
- *Definida para todos los tipos de variables: numéricas* (ratio o intervalo) *y categóricas* (ordinales o nominales)

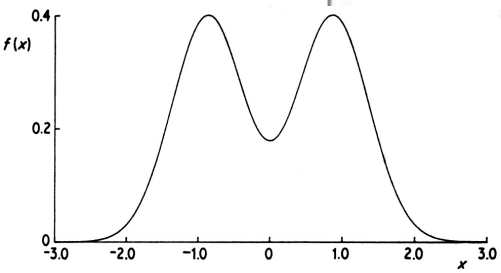
Comparación de las tres medidas



Source - CDC - Course SS1978 - Lesson 2 Overview

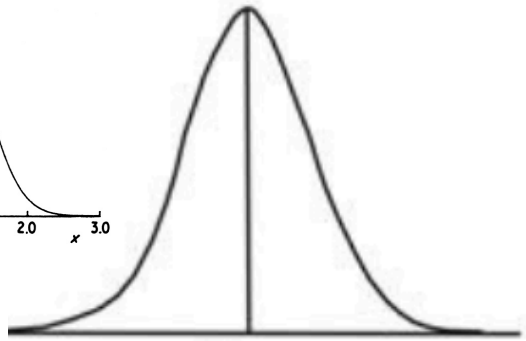
Comparación de las tres medidas

¡No siempre!



Bimodal simétrica

Distribución simétrica

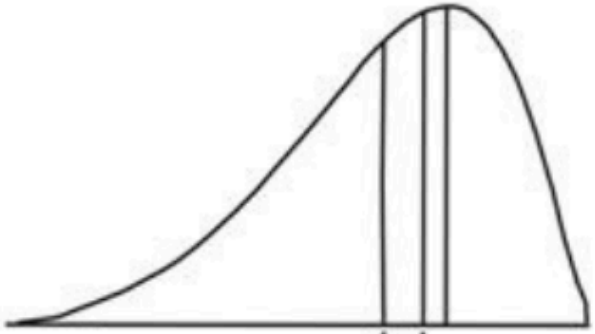


Moda

Mediana

Media

Distribución asimétrica



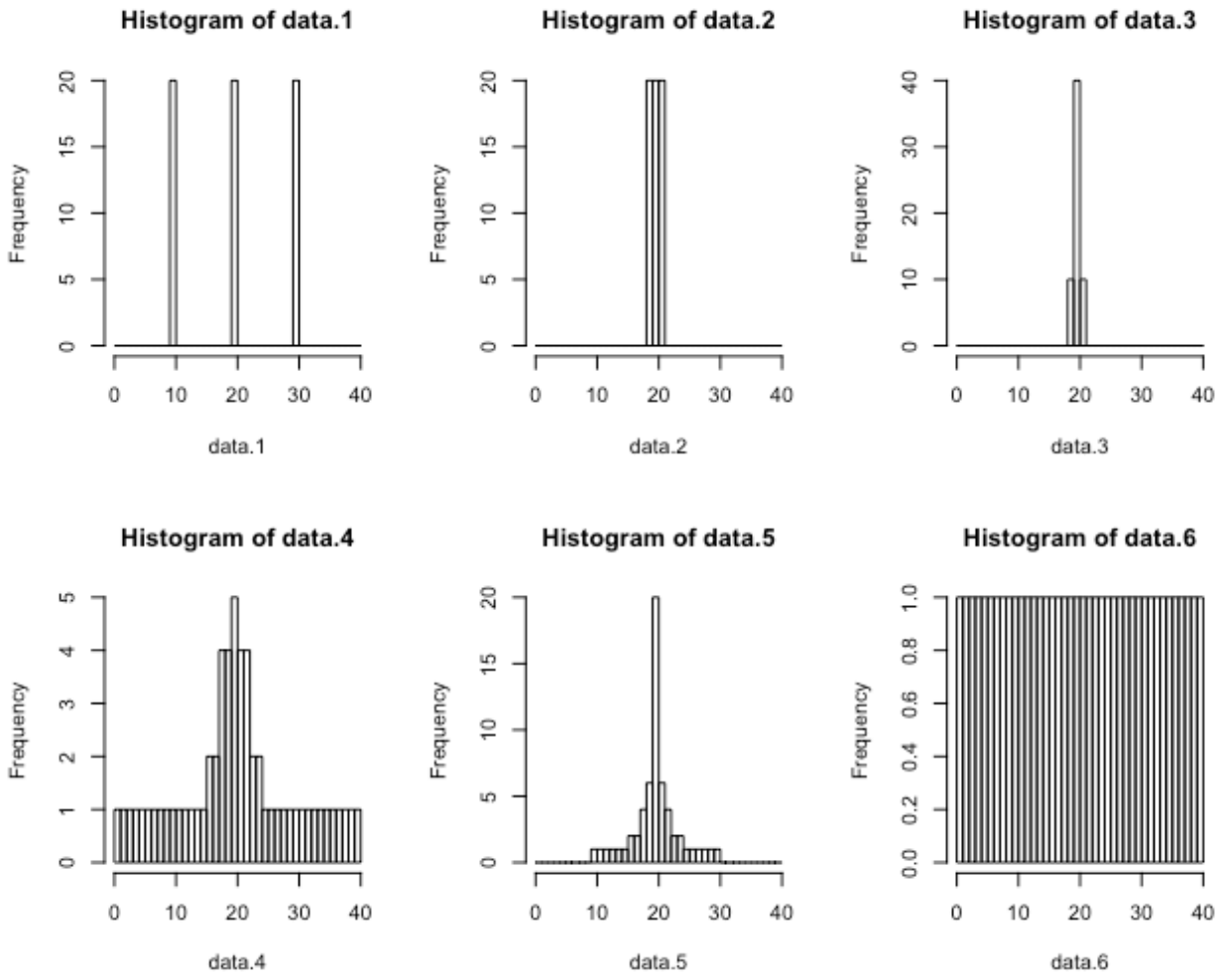
Moda

Mediana

Media

Source - CDC - Course SS1978 - Lesson 2 Overview

Medidas de dispersión



Medidas de dispersión

- Queremos saber “cuán dispersa” es una distribución
  1. Calculamos la distancia entre el valor mínimo y el máximo de nuestras observaciones.
  2. Promediamos la desviación (absoluta) de cada observación respecto de la media.
  3. Promediamos el cuadrado de la desviación de cada observación respecto de la media (y tomamos la raíz cuadrada)
  4. Tomamos el intervalo “central”, que deja un 25% de las observaciones por debajo y un 25% de las observaciones por encima
  5. Tomamos la mediana de las desviaciones de la mediana

Medidas de dispersión

- Queremos saber "cuán dispersa" es una distribución
  1. Calculamos la distancia entre el valor mínimo y el máximo de nuestras observaciones.
  2. Promediamos la desviación (absoluta) de cada observación respecto de la media.
  3. Promediamos el cuadrado de la desviación de cada observación respecto de la media (y tomamos la raíz cuadrada)
  4. Tomamos el intervalo "central", que deja un 25% de las observaciones por debajo y un 25% de las observaciones por encima
  5. Tomamos la mediana de las desviaciones de la mediana

1. Rango

- Es simplemente la extensión del intervalo en el que entran todas las observaciones
- rango = valor máximo – valor mínimo

2. Varianza y desviación estándar

- Es el promedio de las desviaciones cuadráticas de la media.
- El valor cuadrático se llama **varianza.**
- La raíz cuadrada de la varianza es la **desviación estándar**, y tiene las mismas unidades que la variable original

3. Cuartiles y cuantiles

- La mediana es el valor que deja al 50% de los valores por debajo
- Le llamaremos el **percentil 50%** o **50mo percentil**
- Si dividimos nuestra muestra en cuatro partes iguales (con la misma cantidad de observaciones), definimos los **cuaRtiles**:
  - Al percentil 0% le llamamos el cuartil cero (mínimo)
  - Al percentil 25% le llamamos primer cuartil
  - Al percentil 50% le llamamos segundo cuartil (mediana)
  - Al percentil 75% le llamamos tercer cuartil
  - Al percentil 100% le llamamos cuarto cuartil (máximo)
- También podemos llamarles cuaNtiles:
  - Al percentil 30% lo llamamos cuantil **0.3**

3. Cuartiles y cuantiles

Cuartil	Percentil	Cuantil	En criollo
0mo	0mo o 0%	0	Mínimo
1ro	25to o 25%	0.25	
2do	50mo o 50%	0.5	Meidana
3ro	75to o 75%	0.75	
4to	100mo o100%	1	Máximo

Rango intercuartil: 3° cuartil – 1° cuartil

4. Desviación absoluta de la Mediana (MAD)

- Es la mediana de las desviaciones absolutas (de las observaciones con respecto a la mediana)