# Análisis de Datos

## Estadística Computacional

Juan Zamora Osorio
juan.zamora@pucv.cl

Instituto de Estadística
Pontificia Universidad Católica de Valparaíso

28 de agosto de 2023

# Terminología

## Dato

- Del latín *datum*, es una representación simbólica, un atributo o una característica de una entidad o cosa.
- *Observación* de un fenómeno con una escala de medida.

## Población objetivo (universo de estudio) ($\Omega$)

- Colección de todas las unidades posibles consideradas.

## Muestra

- Lo que conocemos de una población.
- Es una imagen (sinopsis) del todo.
- $M = \{\omega_1, \omega_2, \ldots, \omega_n\} \subseteq \Omega$ muestra de $n$ datos.

# Terminología

## Dato

- Del latín *datum*, es una representación simbólica, un atributo o una característica de una entidad o cosa.
- *Observación* de un fenómeno con una escala de medida.

## Cuidado

- Recordar que el dato no corresponde a la realidad misma, sino a su representación

# Datos categóricos / discretos / nominales / cualitativos

## Ejemplos

- Género: masculino, femenino, no binari@.
- Nacionalidad: Chile, Argentina, Perú, etc.
- Color de ojos.
- Ha salido fuera del país: sí, no.

# Datos ordinales

## Ejemplos

- Edad: 0, 1 año, 2 años, etc.
- Cantidad de hijo/as: 0, 1, 2, 3, etc.
- Talla de zapatos.
- Niveles de educación (?).

# Datos continuos

## Ejemplos

- Tiempo que demora en caer un objeto.
- Distancia al lanzar una piedra.
- Precio de una acción en la bolsa.
- Temperatura.
- Ángulo que apunta una ruleta.
- Segundos desde 1970 (?).

# Escalas de medida

| Escala | Operaciones empíricas básicas | Estructura matemática de grupo | Estadísticos invariantes |
|---|---|---|---|
| **Nominal** | Igualdad | Grupo de permutaciones $x' = f(x)$, con $f(x)$ cualquier sustitución uno a uno | Número de casos, moda, correlación de contingencia |
| **Ordinal** | Mayor o menor | Grupo isotónico $x' = f(x)$, con $f(x)$ cualquier función monótona | Mediana, percentiles (cuantiles) |
| **Intervalar** | Igualdad de intervalos o diferencias | Grupo lineal general $x' = ax + b$ | Media, desviación estándar, correlaciones |
| **De razón** | Igualdad de razones | Grupo de similaridad $x' = ax'$ | Coeficiente de variación |

# Terminología – Tipos de datos

## Según cantidad de variables

- Univariados.
- Bivariados.
- Multivariados.

## Cuidado

Muchos fenómenos se representan por medio de datos mezclados (estructurados o no estructurados).

# Conjunto de datos

## Qué es

- Recordar: una representación de la cosa.
- Descripciones de distintas instancias de un mismo fenómeno.
- Similar a una “muestra”.
- Representados como tuplas, matrices, tensores, grafos, imágenes, etc.

$$
\begin{array}{c}
\begin{array}{ccccc} \text{Dato} & \text{Atributo } 1 & \text{Atributo } 2 & \cdots & \text{Atributo } p \end{array} \\
\begin{pmatrix}
1 & x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\
2 & x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
n & x_{n,1} & x_{n,2} & \cdots & x_{n,p}
\end{pmatrix}_{n \times p}
\end{array}
$$

# Conjunto de datos como matriz

## Ejemplo: despachos de una cadena de pizzerías

$$\begin{array}{c}
\begin{array}{ccccc}
\text{Entrega} & \text{Tiempo entrega} & \text{Temperatura} & \cdots & \text{Sucursal}
\end{array} \\
\begin{pmatrix}
1 & 35{,}1 & 68{,}3 & \cdots & \text{Este (1)} \\
2 & 25{,}2 & 71{,}0 & \cdots & \text{Este (1)} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1266 & 35{,}7 & 60{,}8 & \cdots & \text{Oeste (2)}
\end{pmatrix}
\end{array}$$

## Decisiones / transformaciones

- ¿Unidades?
- ¿Cómo representar datos categóricos?
- 1, 2, 3, 4.
- 0001, 0010, 0100, 1000.

# Frecuencias

## Ejemplo: alumnos presentes

- S, N, S, N, S, S, S, N, S, S.

## Frecuencias

- Absoluta:
  - $n_S = 7$.
  - $n_N = 3$.
- Relativa:
  - $f_S = \frac{n_S}{n} = \frac{7}{10} = 0{,}7$.
  - $f_N = \frac{n_N}{n} = \frac{3}{10} = 0{,}3$.

# ¿Y los datos continuos?

## Agrupando

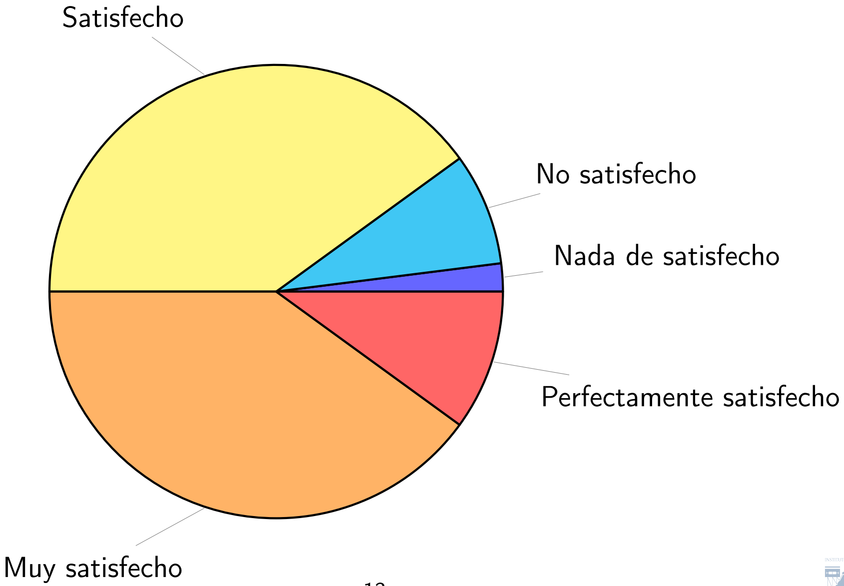
- Podemos agrupar por intervalos.

## Ejemplo: notas de control

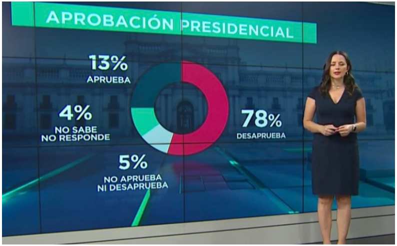
- 28, 35, 42, 90, 70, 56, 75, 66, 30, 89, 75, 64, 81, 69, 55, 83, 72, 68, 73, 16.

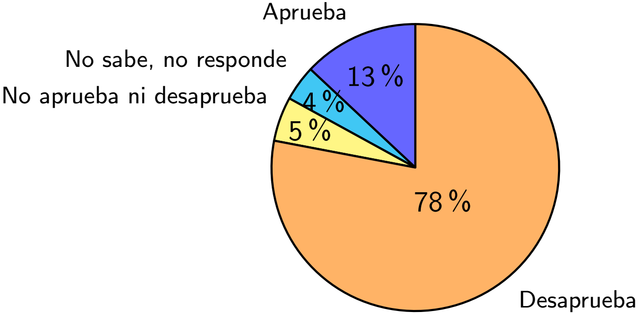
| 0 – 20 | 21 – 40 | 41 – 60 | 61 – 80 | 81 – 100 |
|---|---|---|---|---|
| $n_1 = 1$ | $n_2 = 3$ | $n_3 = 3$ | $n_4 = 9$ | $n_5 = 4$ |
| $f_1 = \frac{1}{20}$ | $f_2 = \frac{3}{20}$ | $f_3 = \frac{3}{20}$ | $f_4 = \frac{9}{20}$ | $f_5 = \frac{4}{20}$ |

# Visualización: gráfico de torta



# Visualización: gráfico de torta

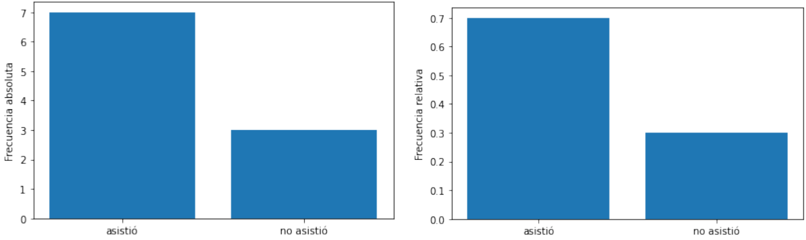


# Visualización: gráfico de barras

## Ejemplo: alumnos presentes

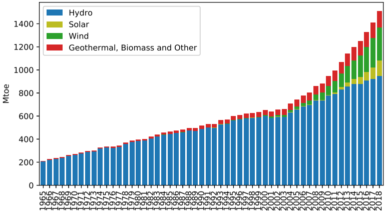
- S, N, S, N, S, S, S, N, S, S.

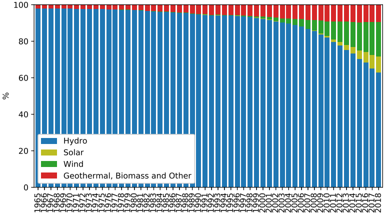


# Visualización: gráfico de barras apilado

## Frecuencia absoluta versus relativa

- ¿Cuál es la diferencia entre utilizar la frecuencia absoluta y la relativa?

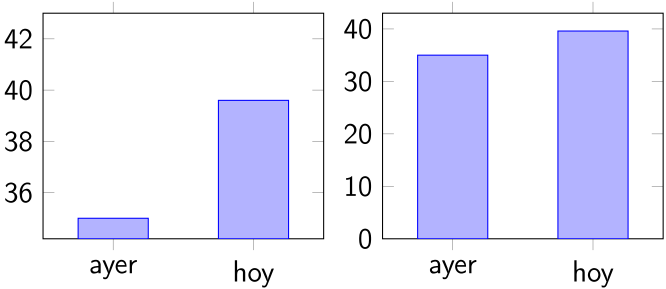


# Visualización: gráfico de barras


# Visualización: gráfico de barras

Porcentaje que se paga de impuestos:



# Gráfico de tallo-hojas

## Notas de control

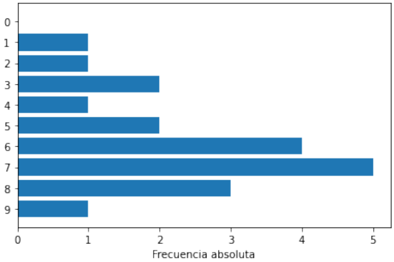
- 28, 35, 42, 90, 70, 56, 75, 66, 30, 89, 75, 64, 81, 69, 55, 83, 72, 68, 73, 16.

| | |
|---|---|
| 0 | |
| 1 | 6 |
| 2 | 8 |
| 3 | 05 |
| 4 | 2 |
| 5 | 56 |
| 6 | 4689 |
| 7 | 02355 |
| 8 | 139 |
| 9 | 0 |

# Gráfico de tallo-hojas

## Notas de control

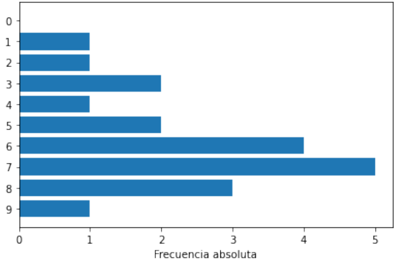
- 28, 35, 42, 90, 70, 56, 75, 66, 30, 89, 75, 64, 81, 69, 55, 83, 72, 68, 73, 16.



# Gráfico de tallo-hojas

## Notas de control

- 28, 35, 42, 90, 70, 56, 75, 66, 30, 89, 75, 64, 81, 69, 55, 83, 72, 68, 73, 16.



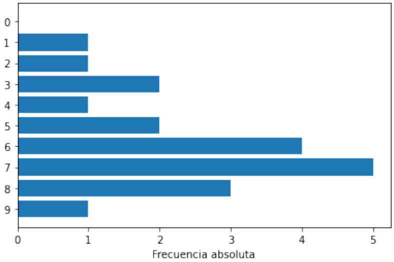

## Heurística: ley de Benford

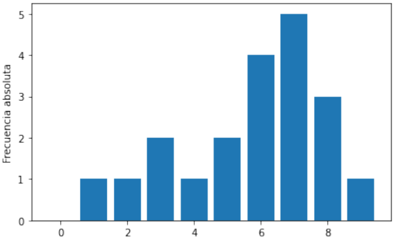
Dependiendo del dominio, tienden a agruparse en los números más pequeños, especialmente dominios abiertos de razón.

# Histograma

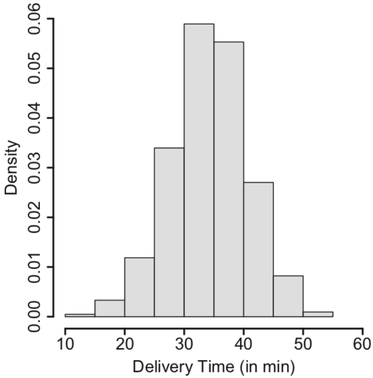
## Notas de control

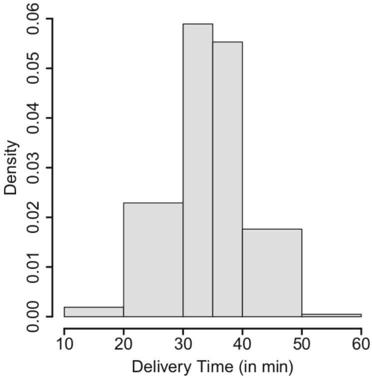
- 28, 35, 42, 90, 70, 56, 75, 66, 30, 89, 75, 64, 81, 69, 55, 83, 72, 68, 73, 16.



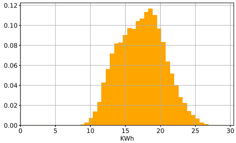
# Histograma – continuo



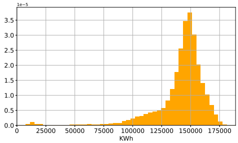



# Histograma – ejemplos demanda eléctrica

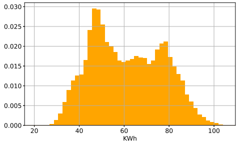
Pan de Azúcar



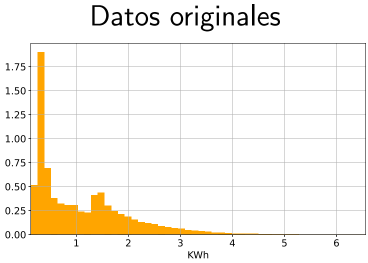

Quillota

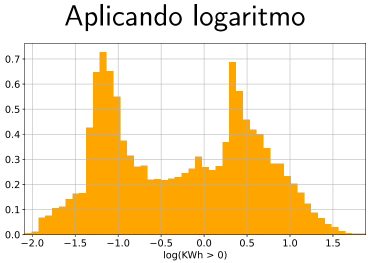



San Joaquín

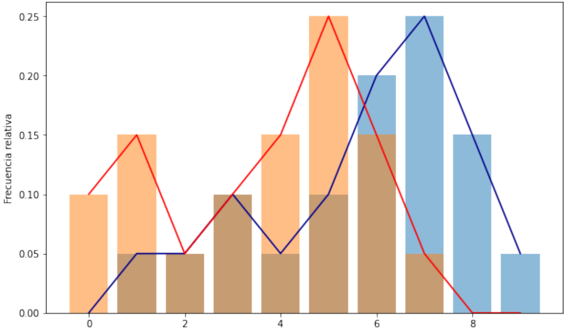


# Histograma – ejemplo demanda eléctrica

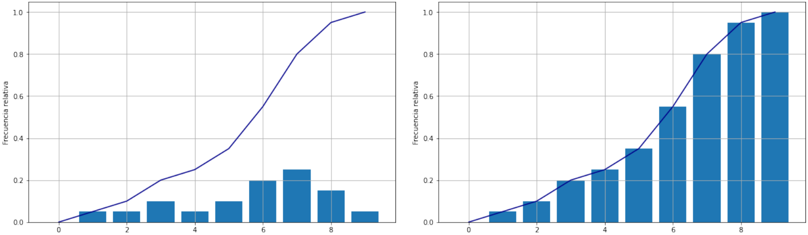


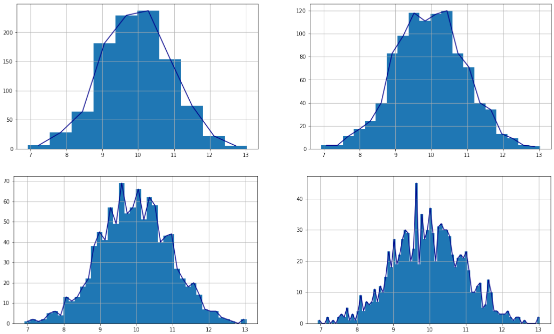
# Histograma – polígono de frecuencias



# Histograma – frecuencia acumulada



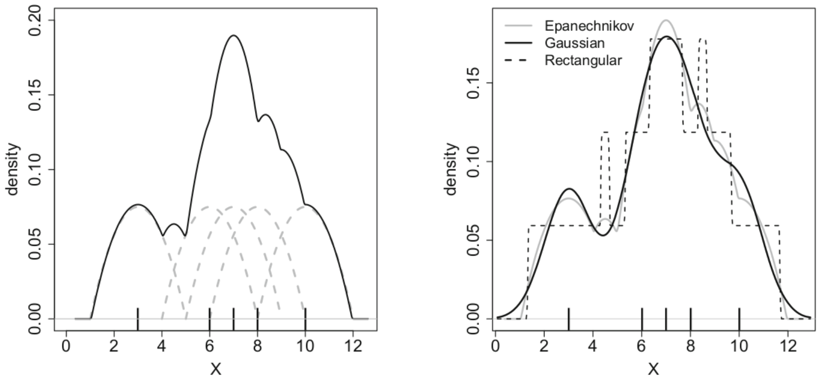
# Histograma – ¿cuántos intervalos?




## Heurística: regla de Sturges

- Se puede elegir cantidad $k = 1 + \log n$, donde $k$ es la cantidad de clases (intervalos) y $n$ la cantidad de datos.
- Recordar que **es solo una heurística**.

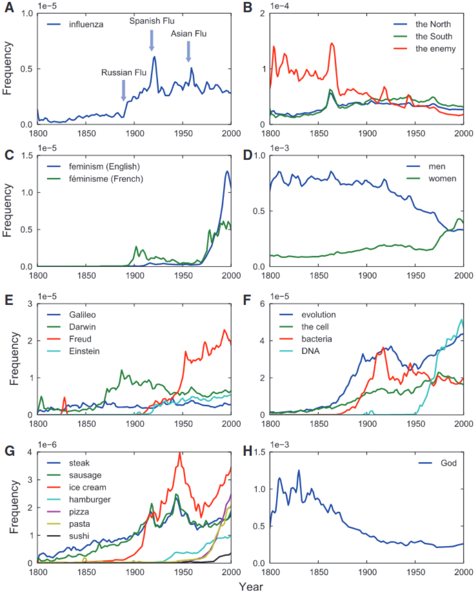
# Nota al margen – Kernel Density Plot




## Otros asuntos

- Mezcla de gaussianas (normales).
- Ventanas de Parzen.

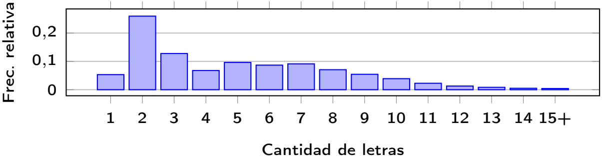
# Un ejemplo: palabras en libros en inglés



# Preguntas

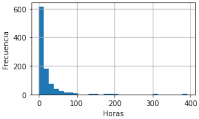
## Largo de palabras en castellano

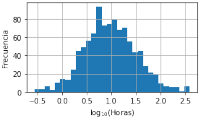
¿Con qué tipo de dato se está trabajando? ¿Qué significa que la frecuencia relativa de las palabras de largo 15 o más sea 0,004?




## Horas de juego de juegos video

¿Qué histograma considera más apropiado para explorar las propiedades de la cantidad de horas?



# Medidas de localización

## Media muestral (media aritmética)

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{\sum_{i=1}^{n} x_i}{n}.$$

- Busca estimar la *media* de la población, denotada como $\mu$.

## Ejemplo: temperaturas máximas durante enero

- 22, 24, 21, 22, 25, 26, 25, 24, 23, 25, 25, 26, 27, 25, 26, 25, 26, 27, 27, 28, 29, 29, 29, 28, 30, 29, 30, 31, 30, 28, 29.
- $\bar{x} = \frac{22+24+21+\cdots+28+29}{31} = 26{,}48$ ºC.

# Media muestral – promedio

## Ejemplo: temperaturas máximas durante enero

- 22, 24, 21, 22, 25, 26, 25, 24, 23, 25, 25, 26, 27, 25, 26, 25, 26, 27, 27, 28, 29, 29, 29, 28, 30, 29, 30, 31, 30, 28, 29.

| $< 20$ | $(20, 25]$ | $(25, 30]$ | $(30, 35]$ | $> 35$ |
|---|---|---|---|---|
| $n_1 = 0$ | $n_2 = 12$ | $n_3 = 18$ | $n_4 = 1$ | $n_5 = 0$ |
| $f_1 = 0$ | $f_2 = \frac{12}{31}$ | $f_3 = \frac{18}{31}$ | $f_4 = \frac{1}{31}$ | $f_5 = 0$ |

## Datos agrupados

- Se debe elegir **marca de clase** (representante de la clase).
- Ejemplo: valor medio de cada intervalo: $m_j = \frac{e_{j-1}+e_j}{2}$.
- Media aritmética con pesos $\bar{x} = \frac{1}{n}\sum_{j=1}^{k} n_j m_j = \sum_{j=1}^{k} f_j m_j$.
- Ejemplo: $\bar{x} = 0 + \frac{12}{31}22{,}5 + \frac{18}{31}27{,}5 + \frac{1}{31}32{,}5 + 0 \approx 25{,}7$ ºC.

# Media muestral – promedio

## Ejemplo: temperaturas máximas durante enero

- 22, 24, 21, 22, 25, 26, 25, 24, 23, 25, 25, 26, 27, 25, 26, 25, 26, 27, 27, 28, 29, 29, 29, 28, 30, 29, 30, 31, 30, 28, 29.

| $< 20$ | $(20, 25]$ | $(25, 30]$ | $(30, 35]$ | $> 35$ |
|---|---|---|---|---|
| $n_1 = 0$ | $n_2 = 12$ | $n_3 = 18$ | $n_4 = 1$ | $n_5 = 0$ |
| $f_1 = 0$ | $f_2 = \frac{12}{31}$ | $f_3 = \frac{18}{31}$ | $f_4 = \frac{1}{31}$ | $f_5 = 0$ |

## Datos agrupados

- Para evitar el problema: media de medias.

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{k} n_j \bar{x}_j = \sum_{j=1}^{k} f_j \bar{x}_j.$$

- Ejemplo: $\bar{x} = 0 + \frac{12}{31} 23{,}83 + \frac{18}{31} 28 + \frac{1}{31} 31 + 0 = 26{,}48$ ºC.

# Medidas de localización

## Mediana muestral

$$\tilde{x}_{0,5} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{si } n \text{ es impar,} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) & \text{si } n \text{ es par.} \end{cases}$$

- $x_i$ observación $i$-ésima.
- $x_{(i)}$ estadística de orden $i$-ésima.
- $x_{(1)} = \min\{x_i\}_{i=1}^{n}$ y $x_{(n)} = \max\{x_i\}_{i=1}^{n}$.

### Ejemplo: temperaturas máximas durante enero

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 21 | 22 | 22 | 23 | 24 | 24 | 25 | 25 | 25 | 25 | 25 | 25 | 26 | 26 | 26 | 26 |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| 27 | 27 | 27 | 28 | 28 | 28 | 29 | 29 | 29 | 29 | 29 | 30 | 30 | 30 | 31 | |
| 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | |

- $\tilde{x}_{0,5} = 26$ ºC.

# Mediana muestral – datos agrupados

## Ejemplo: temperaturas máximas durante enero

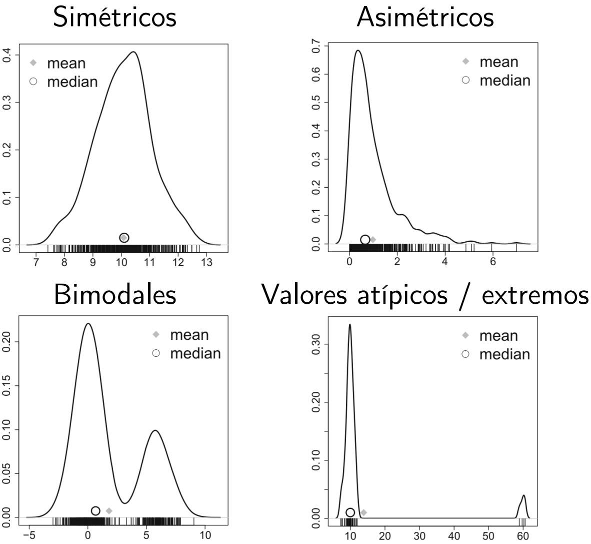
- 22, 24, 21, 22, 25, 26, 25, 24, 23, 25, 25, 26, 27, 25, 26, 25, 26, 27, 27, 28, 29, 29, 29, 28, 30, 29, 30, 31, 30, 28, 29.

| $< 20$ | $(20, 25]$ | $(25, 30]$ | $(30, 35]$ | $> 35$ |
|---|---|---|---|---|
| $n_1 = 0$ | $n_2 = 12$ | $n_3 = 18$ | $n_4 = 1$ | $n_5 = 0$ |
| $f_1 = 0$ | $f_2 = \frac{12}{31}$ | $f_3 = \frac{18}{31}$ | $f_4 = \frac{1}{31}$ | $f_5 = 0$ |

## Datos agrupados

- Clase $m$ es la clase que cumple que $\sum_{j=1}^{m-1} f_j < 0{,}5$ y $\sum_{j=1}^{m} f_j \geq 0{,}5$.
- Mediana interpolada $\tilde{x}_{0,5} = e_{m-1} + \frac{0{,}5 - \sum_{j=1}^{m-1} f_j}{f_m}(e_m - e_{m-1})$.
- Ejemplo: $\tilde{x}_{0,5} = 25 + \frac{0{,}5 - \frac{12}{31}}{\frac{18}{31}} 5 \approx 25{,}97$ ºC.

# Media vs mediana



# Generalización de la mediana

## Cuantiles

$$\tilde{x}_\alpha = \begin{cases} x_{(\lceil n\alpha \rceil)} & \text{si } n\alpha \text{ no es entero,} \\ \frac{1}{2}\left(x_{(n\alpha)} + x_{(n\alpha+1)}\right) & \text{si } n\alpha \text{ es entero.} \end{cases}$$

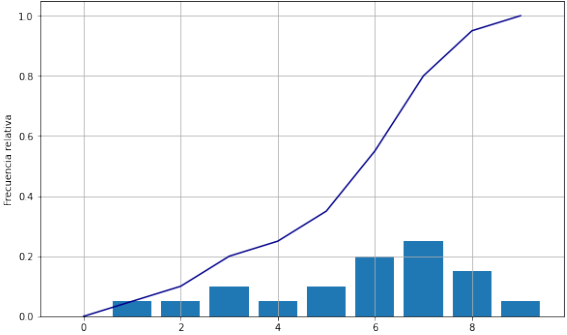
- ¿Es la única manera de interpolar?

### Ejemplo: temperaturas máximas durante enero

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 21 | 22 | 22 | 23 | 24 | 24 | 25 | 25 | 25 | 25 | 25 | 25 | 26 | 26 | 26 | 26 |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| 27 | 27 | 27 | 28 | 28 | 28 | 29 | 29 | 29 | 29 | 29 | 30 | 30 | 30 | 31 | |
| 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | |

- $\tilde{x}_{0,25} = x_{(\lceil 7\frac{3}{4} \rceil)} = x_{(8)} = 25$ ºC.
- $\tilde{x}_{0,5} = x_{(\lceil 15\frac{1}{2} \rceil)} = x_{(16)} = 26$ ºC.
- $\tilde{x}_{0,75} = x_{(\lceil 23\frac{1}{4} \rceil)} = x_{(24)} = 29$ ºC.

# Cuantiles y frecuencia acumulada




- Deciles, quintiles, cuartiles, etc.

# Medidas de dispersión / variabilidad

## Rango

- $R = x_{(n)} - x_{(1)} = \text{máx}\{x_i\}_{i=1}^{n} - \text{mín}\{x_i\}_{i=1}^{n}$.

## Rango intercuantil

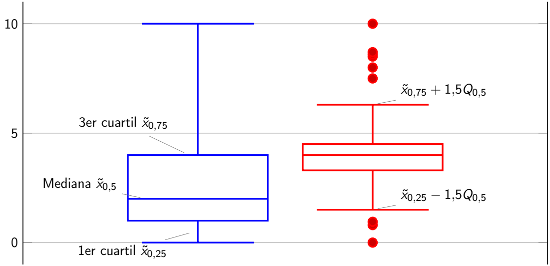
- $Q_\alpha = \tilde{x}_{\frac{1+\alpha}{2}} - \tilde{x}_{\frac{1-\alpha}{2}}$.
- Ejemplo: intercuartil $\alpha = 0{,}5$: $Q_{0,5} = \tilde{x}_{0,75} - \tilde{x}_{0,25}$.

## Ejemplo: temperaturas máximas durante enero

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 21 | 22 | 22 | 23 | 24 | 24 | 25 | 25 | 25 | 25 | 25 | 25 | 26 | 26 | 26 | 26 |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| 27 | 27 | 27 | 28 | 28 | 28 | 29 | 29 | 29 | 29 | 29 | 30 | 30 | 30 | 31 | |
| 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | |

- $R = 31 - 21 = 10$ ºC.
- $Q_{0,5} = 29 - 25 = 4$ ºC.

# Gráfico de caja (*boxplot*)



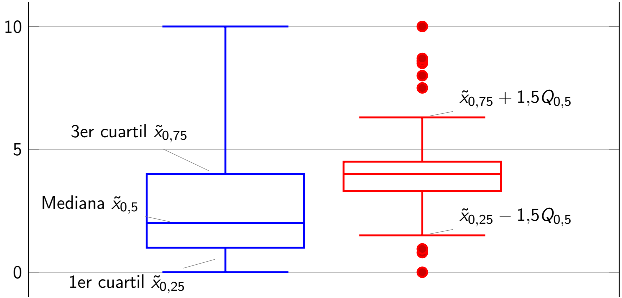

## Valores atípicos / extremos (*outliers*)

- $Q_{0,5} = \tilde{x}_{0,75} - \tilde{x}_{0,25}$.
- $x_i$ valor atípico si $x_i - \tilde{x}_{0,75} > 1{,}5Q_{0,5}$, o si $\tilde{x}_{0,25} - x_i > 1{,}5Q_{0,5}$.
- $x_i$ valor extremo si $x_i - \tilde{x}_{0,75} > 3Q_{0,5}$, o si $\tilde{x}_{0,25} - x_i > 3Q_{0,5}$.
- Estas definiciones varían según libro, biblioteca, etc. Es configurable.

# Asimetría (*skewness*)

## Coeficiente/índice de Bowley / Yule

$$\frac{(\tilde{x}_{0,75} - \tilde{x}_{0,5}) - (\tilde{x}_{0,5} - \tilde{x}_{0,25})}{\tilde{x}_{0,75} - \tilde{x}_{0,25}} = \frac{\tilde{x}_{0,25} + \tilde{x}_{0,75} - 2\tilde{x}_{0,5}}{\tilde{x}_{0,75} - \tilde{x}_{0,25}}.$$



# Medidas de localización

## Moda

- El valor más frecuente.
- Puede haber más de un valor modal.

## Ejemplo: temperaturas máximas durante enero

- 22, 24, 21, 22, 25, 26, 25, 24, 23, 25, 25, 26, 27, 25, 26, 25, 26, 27, 27, 28, 29, 29, 29, 28, 30, 29, 30, 31, 30, 28, 29.

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 21 | 22 | 22 | 23 | 24 | 24 | 25 | 25 | 25 | 25 | 25 | 25 | 26 | 26 | 26 | 26 |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| 27 | 27 | 27 | 28 | 28 | 28 | 29 | 29 | 29 | 29 | 29 | 30 | 30 | 30 | 31 | |
| 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | |

- Moda = 25 ºC.

# Moda – datos agrupados

## Moda

- El intervalo / clase más frecuente.
- ¿Cuál es su representante?

## Ejemplo: temperaturas máximas durante enero

- 22, 24, 21, 22, 25, 26, 25, 24, 23, 25, 25, 26, 27, 25, 26, 25, 26, 27, 27, 28, 29, 29, 29, 28, 30, 29, 30, 31, 30, 28, 29.

| $< 20$ | $(20, 25]$ | $(25, 30]$ | $(30, 35]$ | $> 35$ |
|---|---|---|---|---|
| $n_1 = 0$ | $n_2 = 12$ | $n_3 = 18$ | $n_4 = 1$ | $n_5 = 0$ |
| $f_1 = 0$ | $f_2 = \frac{12}{31}$ | $f_3 = \frac{18}{31}$ | $f_4 = \frac{1}{31}$ | $f_5 = 0$ |

- Clase modal es intervalo $(25, 30]$ ºC.

# Medidas de dispersión / variabilidad / volatilidad

## Desviación absoluta

$$D(A) = \frac{1}{n} \sum_{i=1}^{n} |x_i - A|.$$

Ejemplo: desviación absoluta de la media

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

Mínimo: se obtiene con la mediana $\tilde{x}_{0,5}$

$$D(\tilde{x}_{0,5}) = \frac{1}{n} \sum_{i=1}^{n} |x_i - \tilde{x}_{0,5}|.$$

# Medidas de dispersión / variabilidad / volatilidad

## Error cuadrático medio

$$s^2(A) = \frac{1}{n}\sum_{i=1}^{n}(x_i - A)^2.$$

Mínimo: se obtiene con la media $\bar{x}$, y a su valor se le llama varianza

$$s^2(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - \bar{x}^2) = \left(\frac{1}{n}\sum_{i=1}^{n}x_i^2\right) - \bar{x}^2.$$

## Varianza muestral

- $\bar{x}$ es una estimación de $\mu$, la media de la población.

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

# Varianza – desviación estándar

## Datos agrupados / Datos muestrales

$$s^2 \approx \frac{1}{n-1}\sum_{j=1}^{k} n_j(m_j-\bar{x})^2 \approx \frac{n}{n-1}\sum_{j=1}^{k} f_j(m_j-\bar{x})^2.$$

$$s^2 \approx \frac{1}{n-1}\sum_{j=1}^{k} n_j(\bar{x}_j-\bar{x})^2 \approx \frac{n}{n-1}\sum_{j=1}^{k} f_j(\bar{x}_j-\bar{x})^2.$$

## Ejemplo: temperaturas máximas durante enero

| $< 20$ | $(20, 25]$ | $(25, 30]$ | $(30, 35]$ | $> 35$ |
|---|---|---|---|---|
| $n_1 = 0$ | $n_2 = 12$ | $n_3 = 18$ | $n_4 = 1$ | $n_5 = 0$ |
| $f_1 = 0$ | $f_2 = \frac{12}{31}$ | $f_3 = \frac{18}{31}$ | $f_4 = \frac{1}{31}$ | $f_5 = 0$ |

# Nota al margen – varianza entre - intra clases

## Clase $j$

- Media $\bar{x}_j = \frac{1}{n_j}\sum_{i\in\text{Clase } j} x_i$, varianza $s_j^2 = \frac{1}{n}\sum_{i\in\text{Clase } j}(x_i - \bar{x}_j)^2$.

| Intervalo | $< 20$ | $(20, 25]$ | $(25, 30]$ | $(30, 35]$ | $> 35$ |
|---|---|---|---|---|---|
| $n_j$ | $n_1 = 0$ | $n_2 = 12$ | $n_3 = 18$ | $n_4 = 1$ | $n_5 = 0$ |
| $\bar{x}_j$ | – | 23,83 | 28 | 31 | – |
| $s_j^2$ | – | 1,972 | 2 | 0 | – |

## Varianza entre clases e intra clases

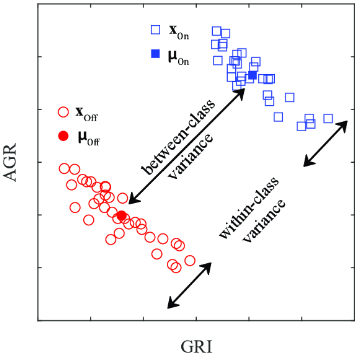
$$s^2 = \left[\frac{1}{n}\sum_{j=1}^{k} n_j(\bar{x}_j - \bar{x})^2\right] + \left[\frac{1}{n}\sum_{j=1}^{k} n_j s_j^2\right].$$

- $\frac{1}{31}\left(12[23{,}83 - 26{,}48]^2 + 18[28 - 26{,}48]^2 + 1[31 - 26{,}48]^2\right) + \frac{1}{31}(12 \cdot 1{,}972 + 18 \cdot 2 + 1 \cdot 0) \approx 4{,}71 + 1{,}925$ .

# Nota al margen – varianza entre - intra clases

## Varianza entre clases e intra clases

$$s^2 = \left[\frac{1}{n}\sum_{j=1}^{k} n_j(\bar{x}_j - \bar{x})^2\right] + \left[\frac{1}{n}\sum_{j=1}^{k} n_j s_j^2\right].$$



# Algunas propiedades

## Transformaciones lineales $y_i = a + bx_i$

- $\bar{y} = a + b\bar{x}$.
- $s_y^2 = b^2 s_x^2$.
- $s_y = |b| s_x$.

## Ejemplo: estandarización

$$y_i = \frac{x_i - \bar{x}}{s_x}.$$

- $\bar{y} = 0$, $s_y = 1$.

## Otras transformaciones

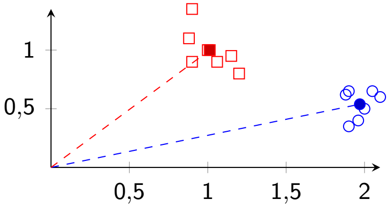
- $y_i = ax_i^2$, $y_i = a\sqrt{x_i}$.
- $y_i = \ln x_i$.

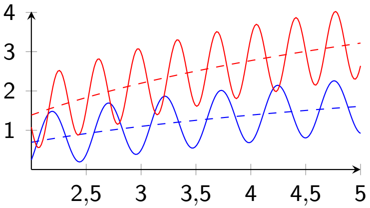
# Medidas de dispersión / variabilidad

## Coeficiente de variación

$$v = \frac{s}{\bar{x}}$$

- Solo para valores positivos de razón.
- Permite comparar poblaciones con unidades diferentes, ya que si $y_i = bx_i$, se cumple que $v_y = \frac{s_y}{\bar{y}} = \frac{bs_x}{b\bar{x}} = \frac{s_x}{\bar{x}} = v_x$.
- A $\frac{\bar{x}}{s}$ se le llama *razón señal ruido*.



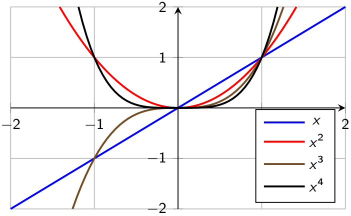
# Otras medidas: momentos

## No centrados

$$m_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

## Centrados

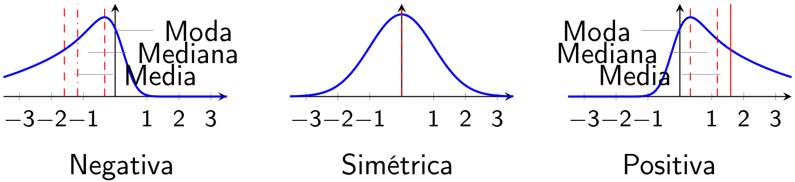
$$\bar{m}_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_1)^k$$



# Tercer momento: asimetría (*skewness*)

## Coeficiente de Fisher

$$\frac{\bar{m}_3}{s^3}.$$



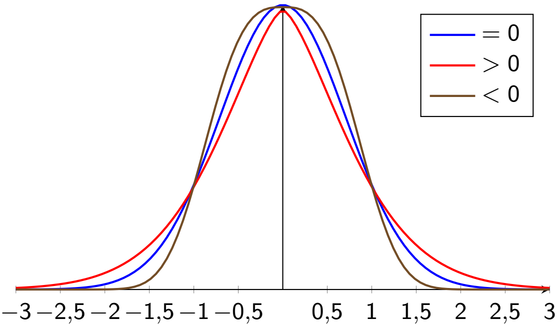

## Recordar

- Coeficiente de variación (inversa de razón señal ruido): $\frac{s}{\bar{x}} = \sqrt{\frac{\bar{m}_2}{\bar{x}^2}}$.

# Cuarto momento: curtosis / apuntamiento (*kurtosis*)

## Coeficiente

$$\frac{\bar{m}_4}{s^4} - 3.$$



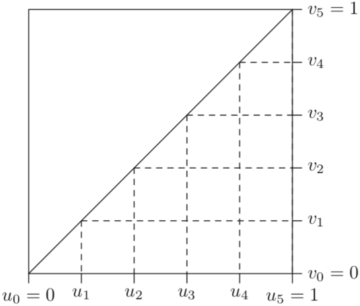

- ¿3?

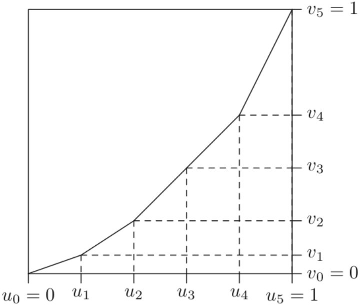
# Medidas de concentración

## Curva de Lorenz

$$u_i = \frac{i}{n},$$

$$v_i = \frac{\sum_{p=1}^{i} x_{(p)}}{\sum_{p=1}^{n} x_{(p)}}.$$



# Medidas de concentración

## Coeficiente de Gini

- $G = 2F$.
- $F = \sum_{i=1}^{n} F_i - \frac{1}{2}$, con $F_i = \frac{(u_{i-1}+u_i)(v_i - v_{i-1})}{2}$.
- $G = 1 - \frac{1}{n}\sum_{i=1}^{n}(v_{i-1} + v_i)$.