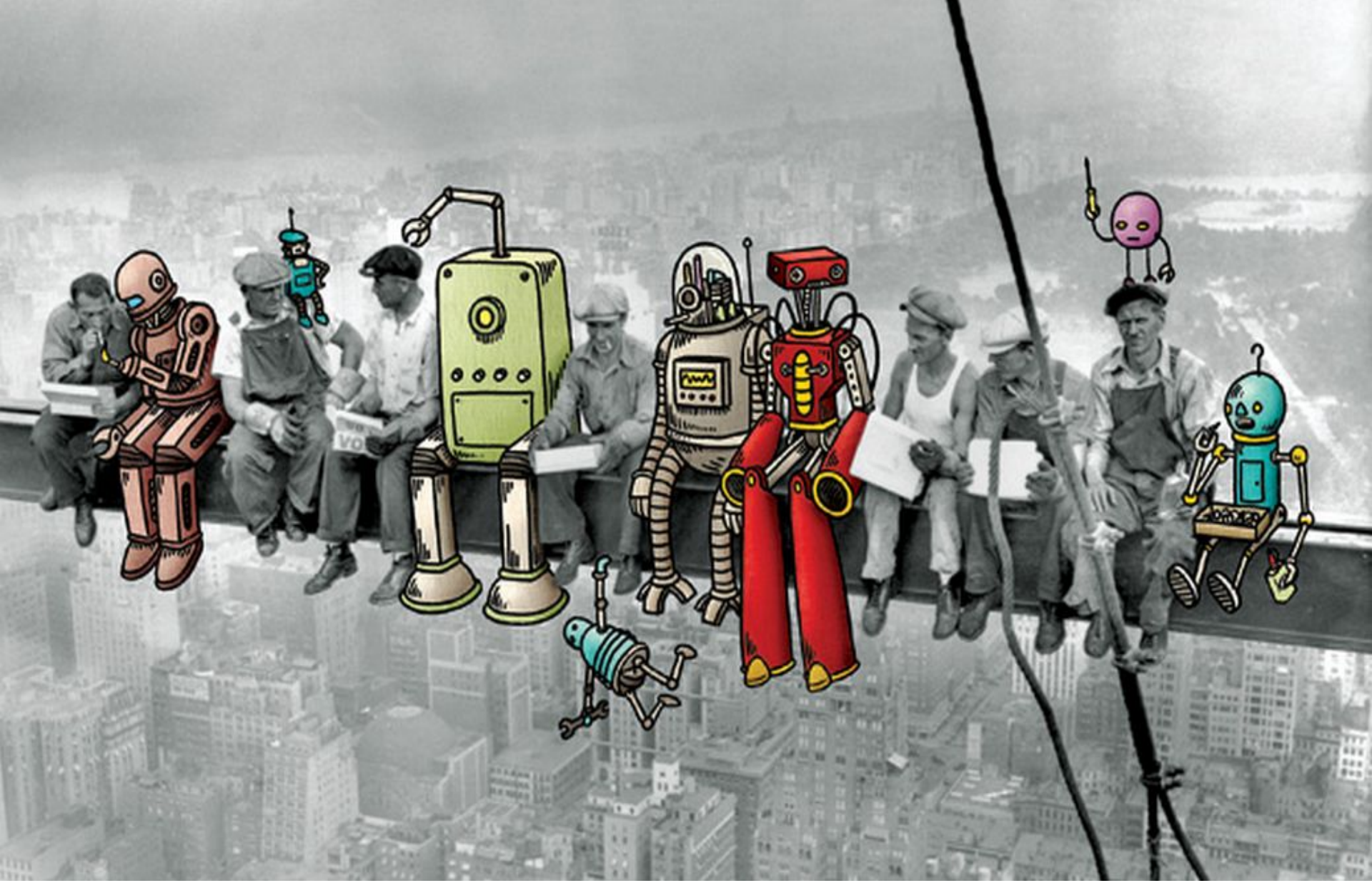


1. Introducción a la estadística

DF02 - Estadística descriptiva

DF - Data First - Data Analytics Journey

DF02 - Estadística descriptiva

1. Introducción a la estadística

- **Objetivos de aprendizaje**
  - Conocer la motivación principal que tiene la estadística y para qué nos puede servir en nuestro día a día
  - Entender conceptos básicos de la estadística

1. Introducción a la estadística

1. Introducción a la estadística

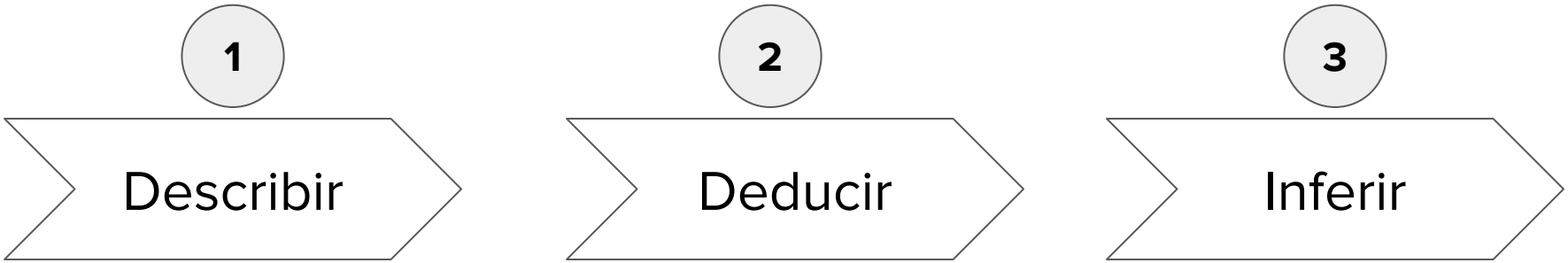
¿Para qué sirve la estadística?

- La estadística tiene por objetivo extraer conocimiento a partir de información (principalmente) numérica
- Se desarrolla **observando hechos**, formulando leyes que los explican y realizando experimentos para validar o rechazar dichas leyes
- Los modelos que crea la ciencia son de tipo bien **determinista** o **aleatorio** (estocástico)
- La **estadística** se utiliza como **tecnología** al **servicio** de las ciencias donde la variabilidad y la incertidumbre forman parte de su naturaleza
- La **estadística** ayuda a investigar en t**odas las áreas de las ciencias** donde la **variabilidad** no es la excepción sino la regla

1. Introducción a la estadística

La estadística es la ciencia de la....

- **Descripción**
  - Sistematización, recogida, ordenación y presentación de los datos referentes a un fenómeno que presenta variabilidad o incertidumbre para su estudio metódico, con objeto de ...
- **Probabilidad**
  - ... deducir las leyes que rigen esos fenómenos, ...
- **Inferencia**
  - ... y poder de esa forma hacer previsiones sobre los mismos, tomar decisiones u obtener conclusiones.



1. Introducción a la estadística

Pasos en un estudio estadístico

- **Plantear hipótesis sobre una población**
  - *Los fumadores "faltan más al puesto de trabajo" que los no fumadores en el Banco Santander*
- **Decidir qué datos recoger** (diseño de experimentos)
  - Qué individuos pertenecerán al estudio (muestras)
    - Fumadores y no fumadores matriculados.
    - ¿Cómo se eligen? ¿Descartamos los que practican deportes?
  - Qué datos recoger de los mismos (variables)
    - Número de ausencias
    - ¿Sexo? ¿repetidores? ¿Otros factores?

- **Recoger los datos (muestreo)**
  - ¿Aleatorio? ¿Estratificado? ¿Sistemáticamente?
- **Describir (resumir) los datos obtenidos**
  - Número medio de ausencias en fumadores y no (estadísticos)
  - % de alumnos ausentes por fumadores y sexo (frecuencias), gráficos,...
- **Realizar una inferencia sobre la población**
  - Los fumadores faltan a clase al menos 2 días/año más de media que los no fumadores.
- **Cuantificar la confianza en la inferencia**
  - Nivel de confianza del 95%

1. Introducción a la estadística

Pasos en un estudio estadístico (II)

1. Introducción a la estadística

Población y muestra

- **Población (*'population'*)**
  - Es el conjunto sobre el que estamos interesados en obtener conclusiones (hacer inferencia).
  - Normalmente es demasiado grande para poder abarcarla.
- **Muestra (*'sample'*)**
  - Es un subconjunto de la población al que tenemos acceso y sobre el que realmente hacemos las observaciones (mediciones)
  - Debería ser "representativo"
  - Está formado por miembros "seleccionados" de la población (individuos, unidades).

1. Introducción a la estadística

Variables

- Los conjuntos de datos que vamos a considerar proceden de medir una o más **variables** de un conjunto de **individuos**
- Una **variable** es una característica observable que varía entre los diferentes individuos de una población
  - La información que disponemos de cada individuo es resumida en variables.
- En los individuos de una población, de uno a otro es variable:
  - El grupo sanguíneo
    - {A, B, AB, O} Var. Cualitativa
  - Su nivel de felicidad “declarado”
    - {Deprimido, ..., Muy Feliz} Var. Ordinal
  - El número de hijos
    - {0,1,2,3,...} Var. Numérica discreta
  - La altura
    - {1,62 ; 1,74; ...} Var. Numérica continua

1. Introducción a la estadística

Variables (II)

- **Cualitativas**
  - Describen cualidades o atributos (*ejemplo*: color de pelo)
  - Si sus valores (modalidades) no se pueden asociar naturalmente a un número (no se pueden hacer operaciones algebraicas con ellos)
    - **Nominales**: Si sus valores no se pueden ordenar
      - Sexo, Grupo Sanguíneo, Religión, Nacionalidad, Fumar (Sí/No)
    - **Ordinales**: Si sus valores se pueden ordenar
      - Mejoría a un tratamiento, Grado de satisfacción, Intensidad del dolor
- **Cuantitativas o numéricas**
  - Si sus valores son numéricos (tiene sentido hacer operaciones algebraicas con ellos)
    - **Discretas**: toman un número pequeño de valores, normalmente enteros
      - Número de hijos, Número de cigarrillos, Num. de “cumpleaños”
    - **Continuas**: toman valores en un intervalo
      - Altura, Presión intraocular, Dosis de medicamento administrado, edad
- En general, la técnica estadística adecuada para analizar una variable depende de su tipo

1. Introducción a la estadística

Variables (III)

- Es buena idea **codificar** las variables como números para poder procesarlas con facilidad en un ordenador.
- Es conveniente asignar "**etiquetas**" a los valores de las variables para recordar qué significan los códigos numéricos.
  - **Sexo** (Cualit: Códigos arbitrarios)
    - 1 = Hombre
    - 2 = Mujer
  - **Raza** (Cualit: Códigos arbitrarios)
    - 1 = Blanca
    - 2 = Negra,...
  - **Felicidad** Ordinal: Respetar un orden al codificar.
    - 1 = Muy feliz
    - 2 = Bastante feliz
    - 3 = No demasiado feliz

1. Introducción a la estadística

Variables (IV)

- Se pueden asignar códigos a respuestas especiales como
  - 0 = No sabe
  - 99 = No contesta...
- Estas situaciones deberán ser tenidas en cuentas en el análisis
  - **Datos perdidos** ('missing data')
- Aunque se codifiquen como números, debemos recordar siempre el verdadero tipo de las variables y su significado cuando vayamos a usar programas de cálculo estadístico:
  - No todo está permitido con cualquier tipo de variable.

1. Introducción a la estadística

Variables (V)

- Los posibles valores de una variable suelen denominarse **modalidades**.
- Las modalidades pueden agruparse en **clases** (intervalos)
  - Edades:
    - Menos de 20 años, de 20 a 50 años, más de 50 años
  - Hijos:
    - Menos de 3 hijos, De 3 a 5, 6 o más hijos
- Las modalidades/clases deben forman un sistema **exhaustivo y excluyente**
  - **Exhaustivo**: No podemos olvidar ningún posible valor de la variable
    - Mal: ¿Cuál es su color del pelo: (Rubio, Moreno)?
    - Bien: ¿Cuál es su grupo sanguíneo?
  - **Excluyente**: Nadie puede presentar (; simultáneos de la variable
    - Estudio sobre el ocio
      - Mal: De los siguientes, qué le gusta: (deporte, cine)
      - Bien: Le gusta el deporte: (Sí, No)
      - Bien: Le gusta el cine: (Sí, No)
      - Mal: Cuántos hijos tiene: (Ninguno, Menos de 5, Más de 2)

1. Introducción a la estadística

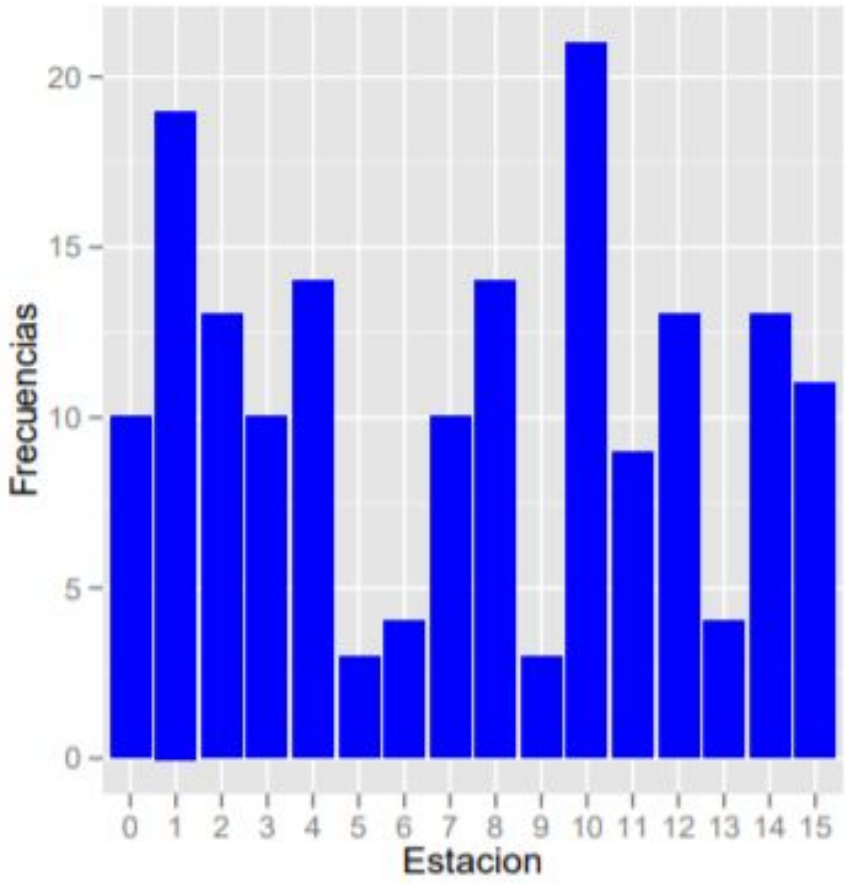
Distribución de una variable

- La **distribución de una variable** viene determinada por los **valores** que toma esa **variable** y la **frecuencia** con las que los toma
  - La *frecuencia absoluta* de un valor (o de un intervalo) es el número de individuos para los que la variable toma ese valor (o pertenece a ese intervalo)
  - La frecuencia relativa es igual a la frecuencia absoluta dividida por el número de datos n
    - Siempre es un número entre 0 y 1
- Aspectos interesantes de una distribución
  - Su **posición**: en torno a qué valor central toma valores la variable
  - Su **dispersión**: el grado de concentración de los valores que toma la variable alrededor de su posición central
  - Su **forma**: por ejemplo, la simetría, es decir, si los valores se reparten de la misma forma a uno y otro lado del centro

1. Introducción a la estadística

Distribución de una variable - Sectores o barras (sólo variables cualitativas o discretas)

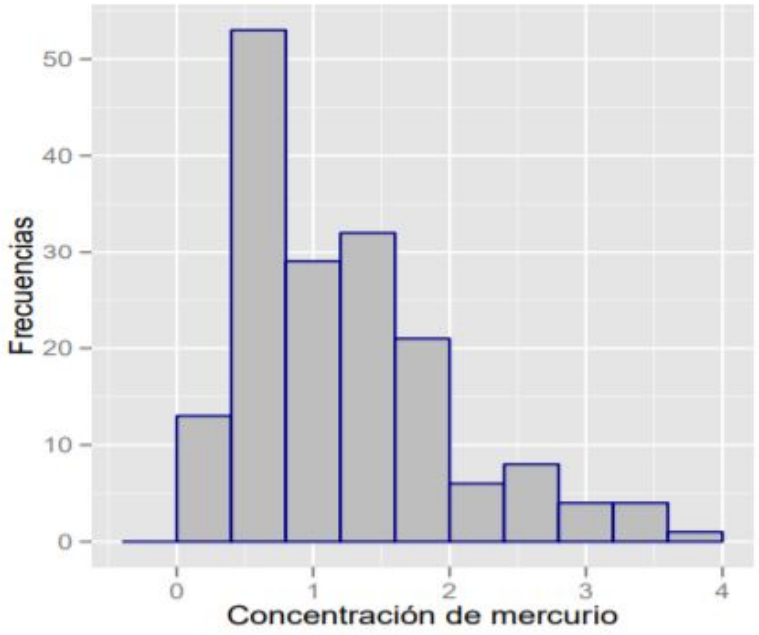
Número de observaciones en cada río



1. Introducción a la estadística

Distribución de una variable - Histogramas (sólo variables continuas)

- Se divide el rango de los datos en un número adecuado de intervalos
- Sobre cada intervalo se dibuja un rectángulo cuya área es proporcional a la frecuencia (relativa o absoluta) de datos en el intervalo



1. Introducción a la estadística

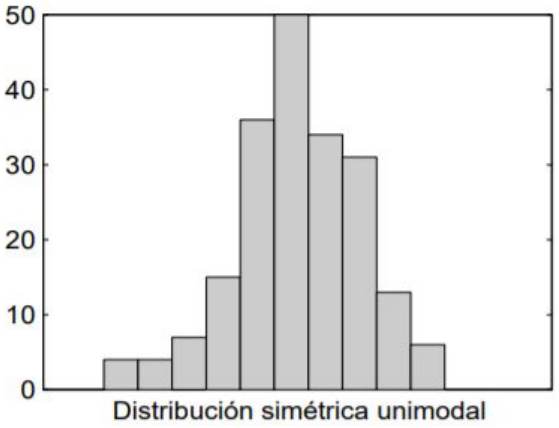
Distribución de una variable - Histogramas (sólo variables continuas) (II)

- Si la base de todos los rectángulos es la misma, la altura es proporcional a la frecuencia
- ¿Cuántas modas hay?
- ¿Hay algún dato atípico en relación al resto?
- ¿Es simétrica la distribución?
- En caso de asimetría, ¿es asimétrica a la izquierda o a la derecha?
- ¿En torno a qué valor aproximado están centrados los datos?
- ¿Están muy dispersos los datos en torno a ese centro?

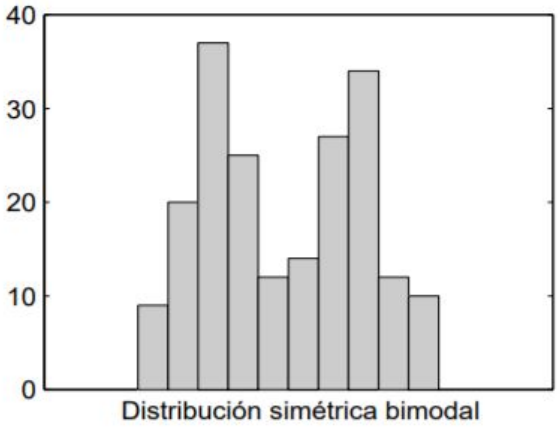
1. Introducción a la estadística

Distribución de una variable - Histogramas (sólo variables continuas) (III)

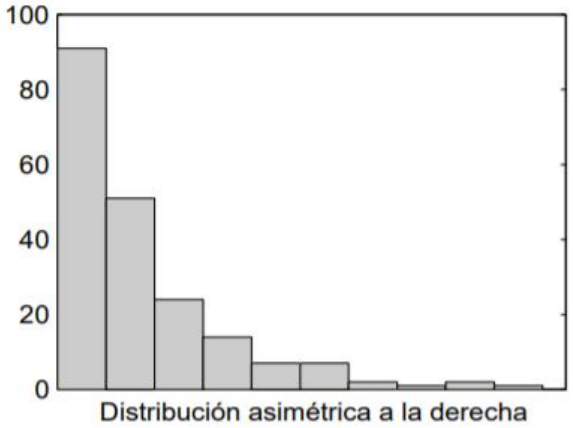
- **Tipos de simetría**



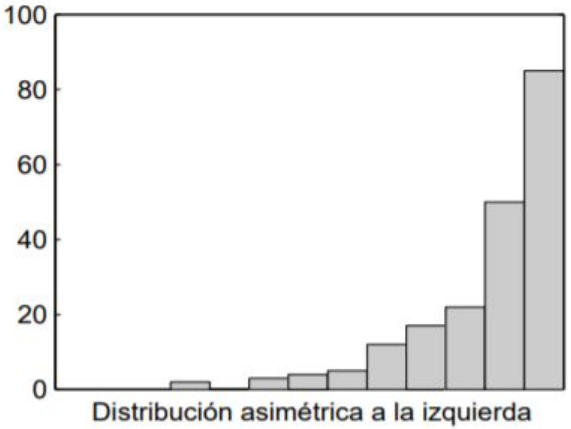
Distribución simétrica unimodal



Distribución simétrica bimodal



Distribución asimétrica a la derecha



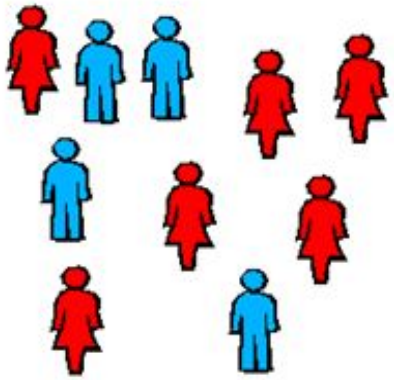
Distribución asimétrica a la izquierda

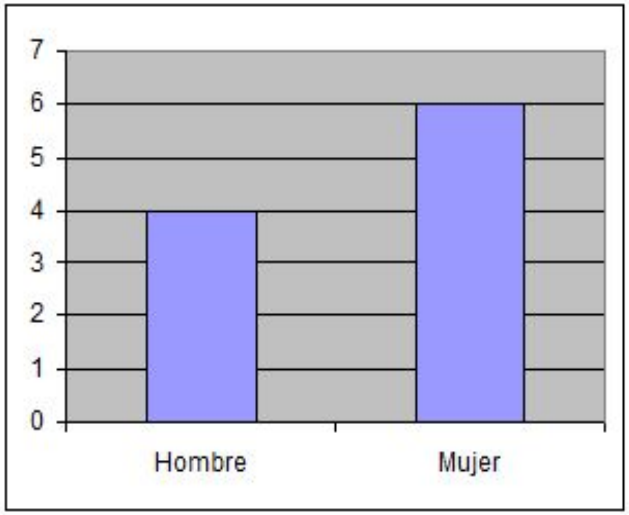
1. Introducción a la estadística

Presentación ordenada de datos

- Las tablas de frecuencias y las representaciones gráficas son dos maneras equivalentes de presentar la información
- Las dos exponen ordenadamente la información recogida en una muestra.

Género	Frec.
Hombre	4
Mujer	6





1. Introducción a la estadística

Tablas de frecuencia

- Exponen la información recogida en la muestra, de forma que no se pierda nada de información (o poca).
  - **Frecuencias absolutas**: Contabilizan el número de individuos de cada modalidad
  - **Frecuencias relativas (porcentajes)**: Idem, pero dividido por el total
  - **Frecuencias acumuladas**: Sólo tienen sentido para variables ordinales y numéricas
    - Muy útiles para calcular cuantiles (ver más adelante)
      - ¿Qué porcentaje de individuos tiene menos de 3 hijos? Sol: 83,8
      - ¿Entre 4 y 6 hijos? Soluc 1ª: 8,4%+3,6%+1,6%= 13,6%. Soluc 2ª: 97,3% - 83,8% = 13,5%

Sexo del encuestado

		Frecuencia	Porcentaje	Porcentaje válido
Válidos	Hombre	636	41,9	41,9
	Mujer	881	58,1	58,1
	Total	1517	100,0	100,0

Nivel de felicidad

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Muy feliz	467	30,8	31,1	31,1
	Bastante feliz	872	57,5	58,0	89,0
	No demasiado feliz	165	10,9	11,0	100,0
	Total	1504	99,1	100,0	
Perdidos	No contesta	13	,9		
Total		1517	100,0		

Número de hijos

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	0	419	27,6	27,8	27,8
	1	255	16,8	16,9	44,7
	2	375	24,7	24,9	69,5
	3	215	14,2	14,2	83,8
	4	127	8,4	8,4	92,2
	5	54	3,6	3,6	95,8
	6	24	1,6	1,6	97,3
	7	23	1,5	1,5	98,9
	Ocho o más	17	1,1	1,1	100,0
	Total	1509	99,5	100,0	
Perdidos	No contesta	8	,5		
Total		1517	100,0		

1. Introducción a la estadística

Tablas de frecuencia (II)

- Exponen la información recogida en la muestra, de forma que no se pierda nada de información (o poca).
  - **Frecuencias absolutas**: Contabilizan el número de individuos de cada modalidad
  - **Frecuencias relativas (porcentajes)**: Idem, pero dividido por el total
  - **Frecuencias acumuladas**: Sólo tienen sentido para variables ordinales y numéricas
    - Muy útiles para calcular cuantiles (ver más adelante)
      - ¿Qué porcentaje de individuos tiene menos de 3 hijos? Sol: 83,8
      - ¿Entre 4 y 6 hijos? Soluc 1ª: 8,4%+3,6%+1,6%= 13,6%. Soluc 2ª: 97,3% - 83,8% = 13,5%

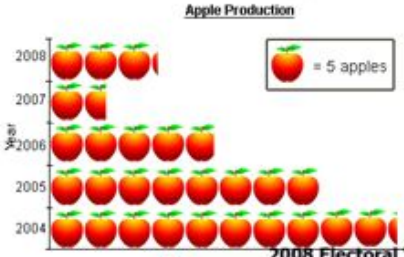
Número de hijos

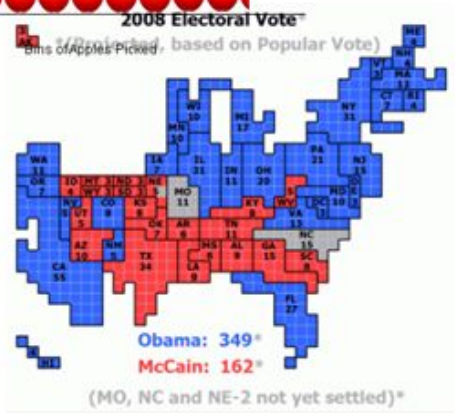
	Frec.	Porcent. (válido)	Porcent. acum.	
0	419	27,8	27,8	
1	255	16,9	44,7	
2	375	24,9	69,5	≥50%
3	215	14,2	83,8	
4	127	8,4	92,2	
5	54	3,6	95,8	
6	24	1,6	97,3	
7	23	1,5	98,9	
Ocho+	17	1,1	100,0	
Total	1509	100,0		

1. Introducción a la estadística

Gráficos para variables cualitativas

- **Diagramas de barras**
  - Alturas proporcionales a las frecuencias (abs. o rel.)
  - Se pueden aplicar también a variables discretas
- **Diagramas de sectores (tortas, polares)**
  - No usarlo con variables ordinales.
  - El área de cada sector es proporcional a su frecuencia (abs. o rel.)
- **Pictogramas**
  - Fáciles de entender.
  - El área de cada modalidad debe ser proporcional a la frecuencia. ¿De los dos, cuál es incorrecto?.

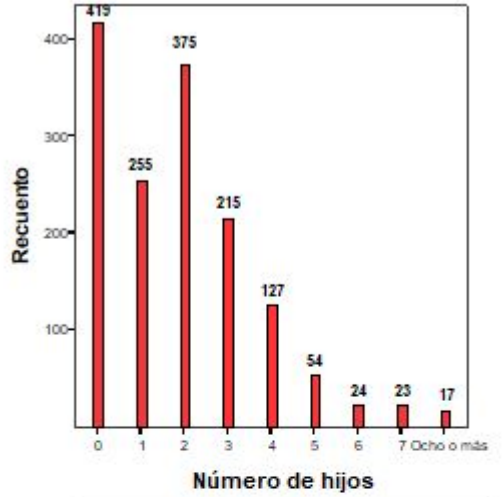


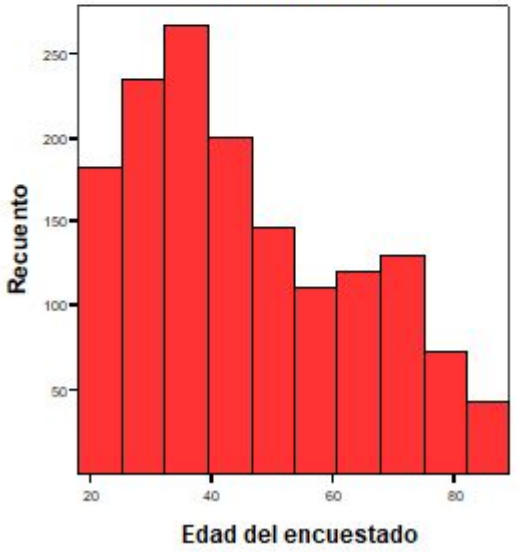


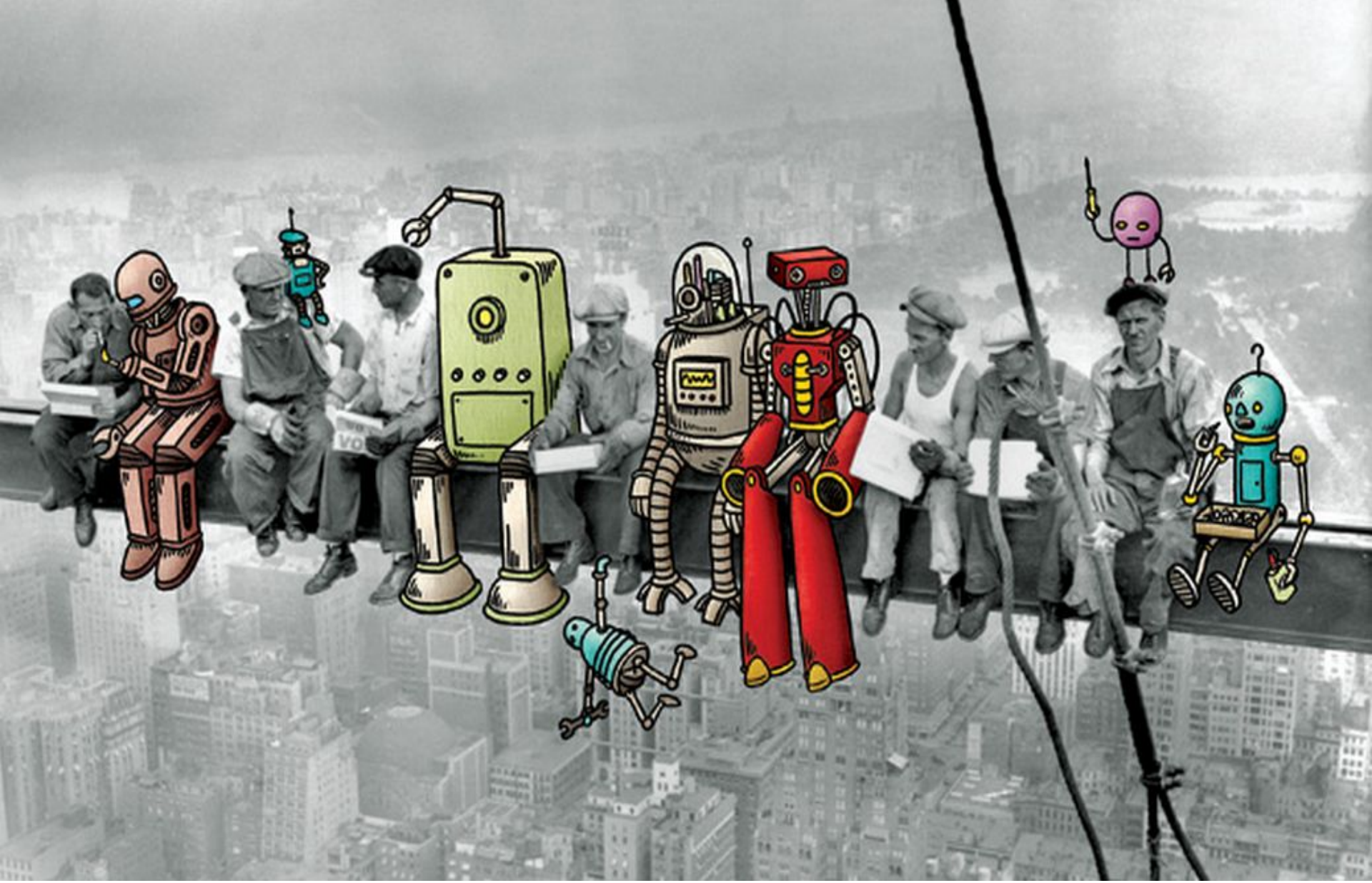
1. Introducción a la estadística

Gráficos diferenciales para variables numéricas

- Son diferentes en función de que las variables sean discretas o continuas
- Valen con frecuencias absolutas o relativas
  - Diagramas barras para variables discretas
    - Se deja un hueco entre barras para indicar los valores que no son posibles
  - Histogramas para variables continuas
    - El área que hay bajo el histograma entre dos puntos cualesquiera indica la cantidad (porcentaje o frecuencia) de individuos en el intervalo.







1. Introducción a la estadística

DF02 - Estadística descriptiva

DF - Data First - Data Analytics Journey