Bioestadística

Tema 1: Introducción a la estadística

¿Para qué sirve la estadística?

- La Ciencia se ocupa en general de fenómenos observables
- La Ciencia 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 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 **Bioestadística** [...] enseña y ayuda a investigar en todas las áreas de las **Ciencias de la Vida donde la variablidad** no es la excepción sino la **regla***"
  Carrasco de la Peña (1982)

Definición

La Estadística es la Ciencia de la

Descriptiva
- **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,

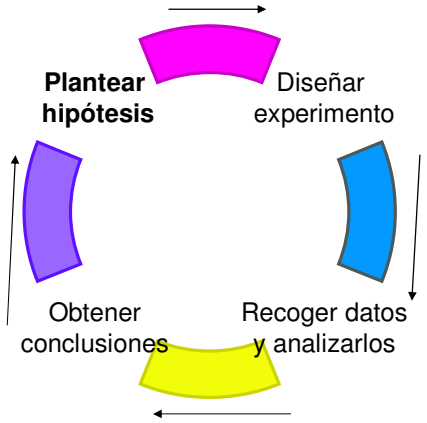
Inferencia
- y poder de esa forma hacer previsiones sobre los mismos, tomar **decisiones** u obtener **conclusiones**.

Pasos en un estudio estadístico

- **Plantear hipótesis sobre una *población***
  - **Los fumadores tienen *"más bajas"* laborales que los no fumadores**
  - ¿En qué sentido? ¿Mayor número? ¿Tiempo medio?
- **Decidir qué datos recoger (diseño de experimentos)**
  - Qué individuos pertenecerán al estudio (*muestras*)
    - Fumadores y no fumadores en edad laboral.
    - Criterios de exclusión ¿Cómo se eligen? ¿Descartamos los que padecen enfermedades crónicas?
  - Qué datos recoger de los mismos (*variables*)
    - Número de bajas
    - Tiempo de duración de cada baja
    - ¿Sexo? ¿Sector laboral? ¿Otros factores?
- **Recoger los datos (*muestreo*)**
  - ¿Estratificado? ¿Sistemáticamente?
- **Describir (resumir) los datos obtenidos**
  - tiempo medio de baja en fumadores y no (*estadísticos*)
  - % de bajas por fumadores y sexo (*frecuencias*), gráficos,...
- **Realizar una inferencia sobre la población**
  - Los fumadores están de baja al menos 10 días/año más *de media* que los no fumadores.
- **Cuantificar la confianza en la inferencia**
  - *Nivel de confianza del 95%*
  - *Significación del contraste: p=2%*

No tenéis que entenderlo (aún)

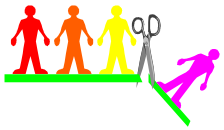
Método científico y estadística



Población y muestra

- **Población** (‘*population*’) es el conjunto sobre el que estamos interesados en obtener conclusiones (hacer inferencia).
  - Normalmente es demasiado grande para poder abarcarlo.

- **Muestra** (‘*sample*’) es un subconjunto suyo al que tenemos acceso y sobre el que realmente hacemos las observaciones (mediciones)
  - Debería ser “representativo”
  - Esta formado por miembros “seleccionados” de la población (individuos, unidades experimentales).



Variables

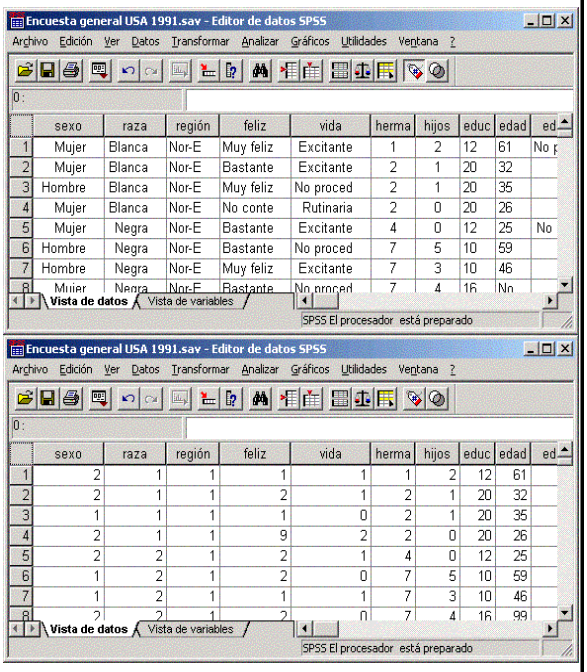
- Una **variable** es una característica observable *que varía entre los diferentes individuos* de una población. La información que disponemos de cada individuo es resumida en **variables**.

- En los individuos de la *población* española, de uno a otro ***es variable***:
  - El grupo sanguíneo
    - {A, B, AB, O} ← Var. Cualitativa
  - Su nivel de felicidad "declarado"
    - {Deprimido, Ni fu ni fa, Muy Feliz} ← Var. Ordinal
  - El número de hijos
    - {0,1,2,3,...} ← Var. Numérica discreta
  - La altura
    - {1'62 ; 1'74; ...} ← Var. Numérica continua

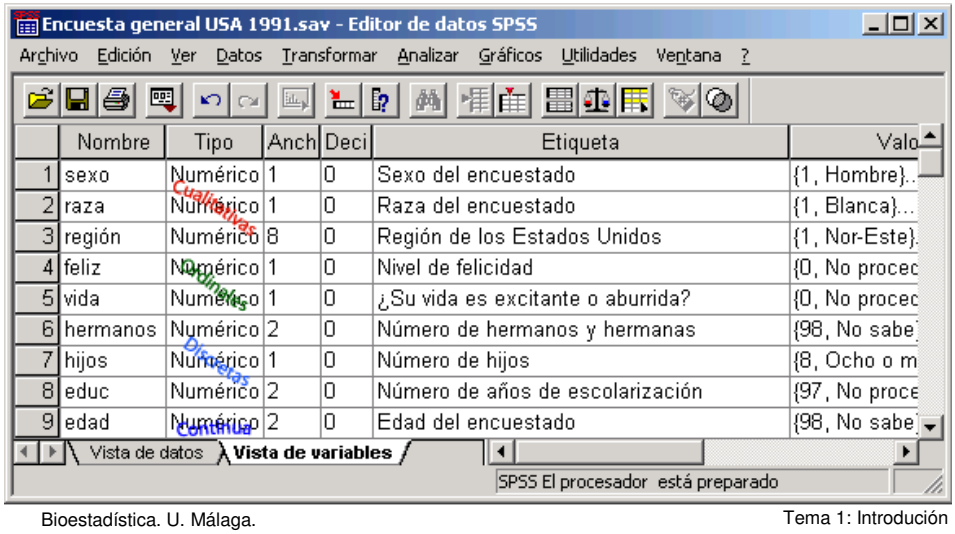
Tipos de variables

- Cualitativas
  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: Si toma valores enteros
    - Número de hijos, Número de cigarrillos, Num. de "cumpleaños"
  - Continuas: Si entre dos valores, son posibles infinitos valores intermedios.
    - Altura, Presión intraocular, Dosis de medicamento administrado, edad

- 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
- 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')



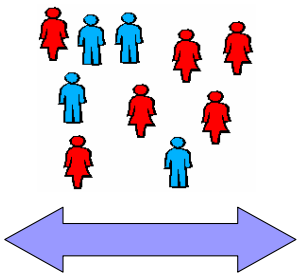
- Aunque se codifiquen como números, debemos recordar siempre el verdadero tipo de las variables y su significado cuando vayamos a usar programas de cálculo estadístico.
- No todo está permitido con cualquier tipo de variable.

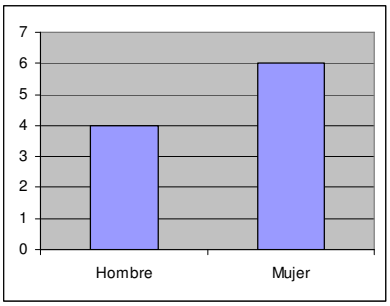


- Los posibles valores de una variable suelen denominarse modalidades.
- Las modalidades pueden agruparse en clases (intervalos)
  - Edades:
    - Menos de 20 años, de 20 a 50 años, más de 50 años
  - Hijos:
    - Menos de 3 hijos, De 3 a 5, 6 o más hijos
- Las modalidades/clases deben forman un sistema 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 dos valores 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)

Presentación ordenada de datos

Género	Frec.
Hombre	4
Mujer	6





- 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.

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		

Datos desordenados y ordenados en tablas

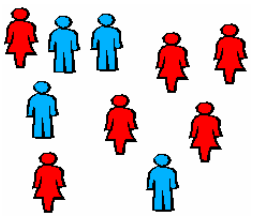
- Variable: Género
  - Modalidades:
    - H = Hombre
    - M = Mujer

Género	Frec.	Frec. relat. porcentaje
Hombre	4	4/10=0,4=40%
Mujer	6	6/10=0,6=60%
	10=tamaño muestral	

- Muestra:

  M H H M M H M M M H

  - equivale a HHHH MMMMMM



Ejemplo

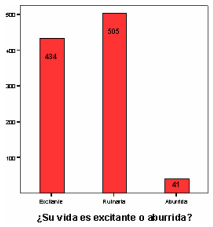
- ¿Cuántos individuos tienen menos de 2 hijos?
  - frec. indiv. sin hijos + frec. indiv. con 1 hijo = 419 + 255 = 674 individuos
- ¿Qué porcentaje de individuos tiene 6 hijos o menos?
  - 97,3%
- ¿Qué cantidad de hijos es tal que al menos el 50% de la población tiene una cantidad inferior o igual?
  - 2 hijos

Número de hijos

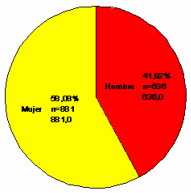
	Frec.	Porcent. (válido)	Porcent. acum.	
0	419	27,8	27,8	
1	255	16,9	44,7	
2	375	24,9	69,5	≥50%
3	215	14,2	83,8	
4	127	8,4	92,2	
5	54	3,6	95,8	
6	24	1,6	97,3	
7	23	1,5	98,9	
Ocho+	17	1,1	100,0	
Total	1509	100,0		

Gráficos para v. cualitativas

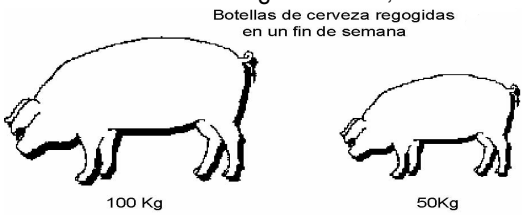
- Diagramas de barras
  - Alturas proporcionales a las frecuencias (abs. o rel.)
  - Se pueden aplicar también a variables discretas

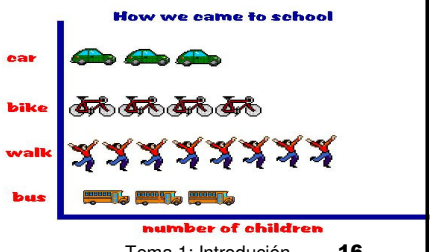


- Diagramas de sectores (tartas, polares)
  - No usarlo con variables ordinales.
  - El área de cada sector es proporcional a su frecuencia (abs. o rel.)



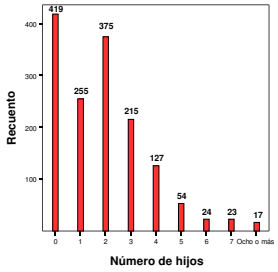
- Pictogramas
  - Fáciles de entender.
  - El área de cada modalidad debe ser proporcional a la frecuencia. ¿De los dos, cuál es incorrecto?.

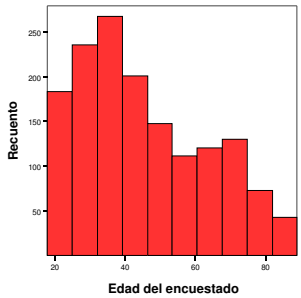




Gráficos diferenciales para variables numéricas

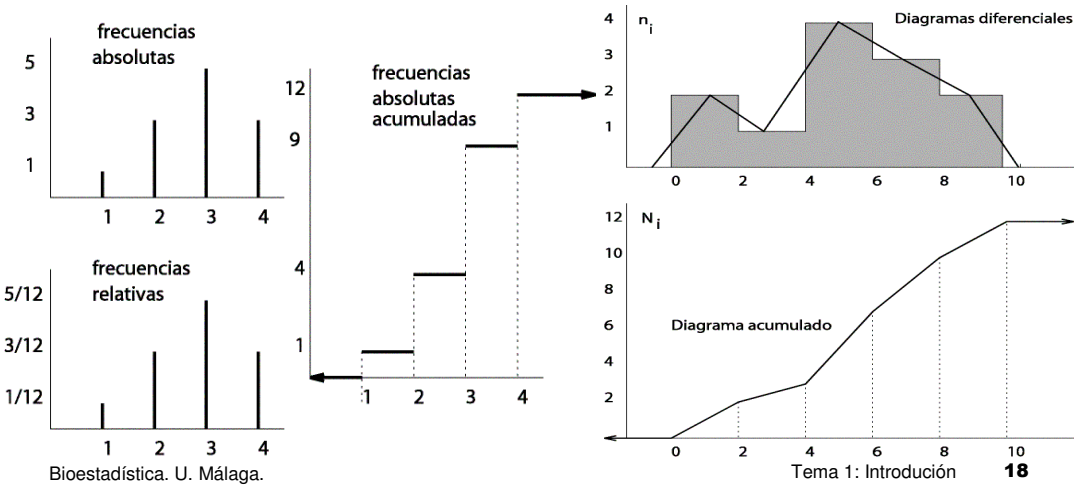
- Son diferentes en función de que las variables sean discretas o continuas. Valen con frec. absolutas o relativas.
  - **Diagramas barras para v. discretas**
    - Se deja un hueco entre barras para indicar los valores que no son posibles
  - **Histogramas para v. continuas**
    - El área que hay bajo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





Diagramas integrales

- Cada uno de los anteriores diagramas tiene su correspondiente diagrama integral. Se realizan a partir de las frecuencias acumuladas. Indican, para cada valor de la variable, la cantidad (frecuencia) de individuos que poseen un valor inferior o igual al mismo. No los construiremos en clase. Se pasan de los diferenciales a los integrales por integración y a la inversa por derivación (en un sentido más general del que visteis en bachillerato.)



¿Qué hemos visto?

- Definición de estadística
- Población
- Muestra
- Variables
  - Cualitativas
  - Numéricas
- Presentación ordenada de datos
  - Tablas de frecuencias
    - absolutas
    - relativas
    - acumuladas
  - Representaciones gráficas
    - Cualitativas
    - Numéricas
      - Diferenciales
      - Integrales