Estadística aplicada en R

Conceptos básicos

-Marzo 2022-

Carlota Solano

carlota.solano.udina@upm.es

0. 1. ¿Qué es la estadística?

- Ciencia que estudia cómo obtener conclusiones empíricas mediante el uso de ***modelos matemáticos***.
- Es la recogida, agrupación, análisis e **interpretación** de datos.

- *Comparar y establecer relaciones entre cosas (e.g. grupos, modelos, valores reales y muestrales, …)*

0.2. ¿Para qué sirve?

- Sacar conclusiones a partir de observaciones
- Puente entre matemáticas y fenómenos reales

- *Comprobar si una hipótesis es cierta o no*

0.3. ¿Dónde se emplea?

- Ingeniería
- Biología
- Economía
- Psicología
- Lingüística
- …

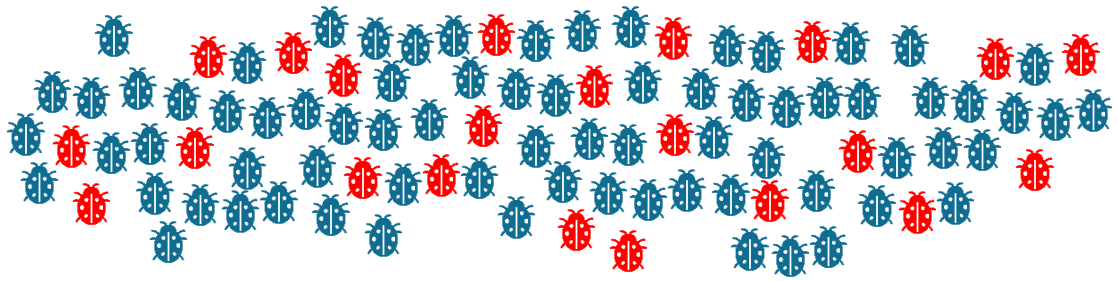
1. Conceptos básicos

1.1. Descripción de datos

- Tipos de datos → Tipo de test → f(datos var. respuesta)
  - Categóricos
    - Ordinal- e.g. excelente, bien, regular, mal, fatal
    - Nominal- e.g. azul, verde, naranja, amarillo
  - Numéricos (*)
    - **Continuos -** e.g. 1, 2.5, 89,0.006, pi
    - Discretos- e.g. 0 & 1 (binario), números enteros (conteos)

  - Paramétricos (*)
  - No paramétricos

- Población vs. Muestra → porque es imposible muestrear toda una población



1. Conceptos básicos

1.1. Descripción de datos

- Medidas de centralidad:

E.g.

1;15; 8; 46; 3; 4; 9; 6; 7; 91; 6; 28; 6

- Media → 17.69

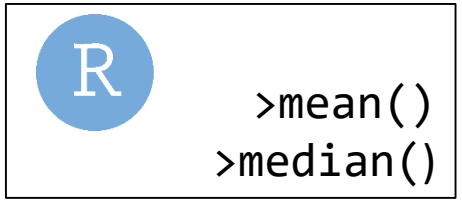
$$\bar{x} = \frac{\sum_{i=1}^{n} x_1 + x_2 + \cdots + x_i}{n}$$

- Mediana → 7

1; 3; 4; 6; 6; 6; 7; 8; 9; 15; 28; 46; 91

- Moda → 6

1; 3; 4; 6; 6; 6; 7; 8; 9; 15; 28; 46; 91



1. Conceptos básicos

1.1. Descripción de datos

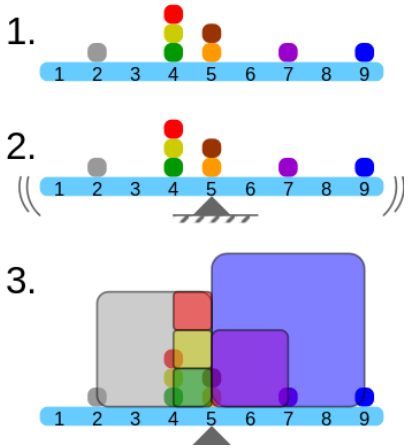
- Medidas de dispersión: variación de los datos de una muestra

  - Varianza

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

  - Desviación estándar (sd):

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$



1. Conceptos básicos

1.1. Descripción de datos

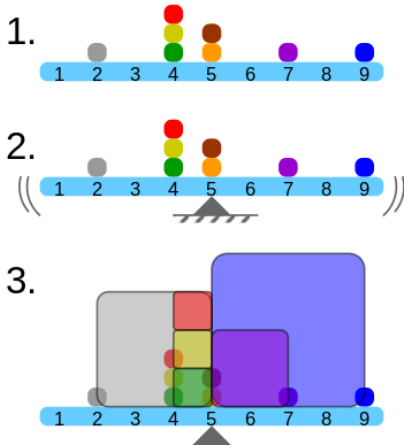
- Medidas de dispersión: variación de los datos de una muestra

  - Varianza

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

  - Desviación estándar (sd):

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$



- Medidas de precisión: cómo de lejos está la media calculada del valor real

  - Error estándar (se)

$$se = \sqrt{\frac{s^2}{n}}$$

  - Intervalo de confianza (95%CI) *95CI= ±1.96 ·se*

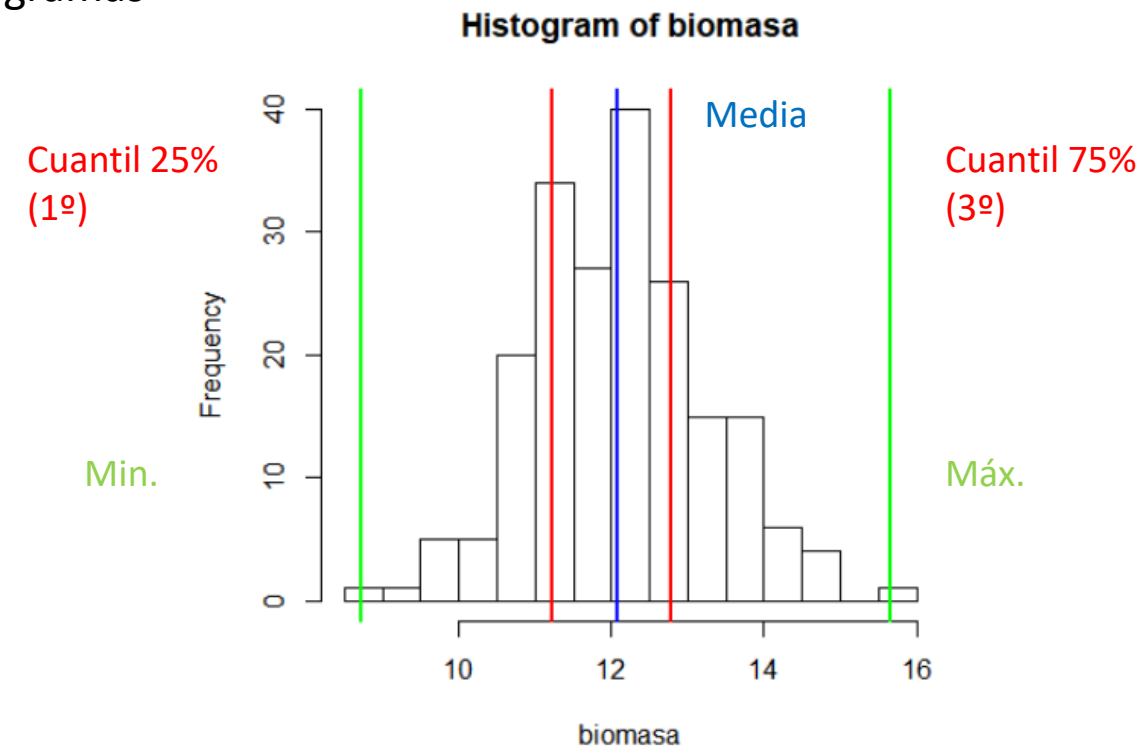
```
>variance()
>sd() == >sqrt(variance())
>se()
```

Varianza y sd: https://www.youtube.com/watch?v=pW8GZujRcFI&ab_channel=jpsauro01

1. Conceptos básicos

1.1. Descripción de datos

- Histogramas



*Cuantil: cierto % de los datos es menor o igual a el valor del cuantil

```
>hist()
>summary()
```

1. Conceptos básicos

1.2. Hipótesis

Hipótesis nula (H0): La diferencia/relación esperada no existe
Hipótesis alternativa (Ha): La diferencia/relación esperada existe

“Siempre” esperamos rechazar H0 y aceptar Ha

1. Conceptos básicos

1.2. Hipótesis

Hipótesis nula (H0): La diferencia/relación esperada no existe
Hipótesis alternativa (Ha): La diferencia/relación esperada existe

"Siempre" esperamos rechazar H0 y aceptar Ha

<u>¿Por qué H0?</u>

H0: Todos los cuervos son negros ←
Ha: Todos los cuervos no son negros

No implica que TODOS los cuervos sean negros, solo que falta evidencia para aceptar Ha → Nuestras observaciones no nos permiten rechazar la H0, pero encontrar un solo cuervo de otro color sí nos permite rechazar H0

1. Conceptos básicos

1.2. Hipótesis

Hipótesis nula (H0): La diferencia/relación esperada no existe
Hipótesis alternativa (Ha): La diferencia/relación esperada existe

"Siempre" esperamos rechazar H0 y aceptar Ha

¿Por qué H0?

H0: Todos los cuervos son negros ←
Ha: Todos los cuervos no son negros

No implica que TODOS los cuervos sean negros, solo que falta evidencia para aceptar Ha → Nuestras observaciones no nos permiten rechazar la H0, pero encontrar un solo cuervo de otro color sí nos permite rechazar H0

1.3. P-valor

Probabilidad de haber obtenido el resultado obtenido siendo H0 cierta

p-valor**: 0.05** → 5% probabilidad de obtener X resultados siendo H0 cierta → Significativo estadísticamente
→ rechazamos H0 y aceptamos Ha

p-valor: 0.8 → 80% probabilidad de obtener X resultados cuando la H0 sea cierta → No significativo estadísticamente
→ no podemos rechazar H0 pero no aceptamos Ha → **Solo encontramos evidencia contra la aceptación de H0**

Es importante pensar

0.05 → valor arbitrario para significancia

1. Conceptos básicos

1.4. Diseño experimental

1º Definir la pregunta / hipótesis a resolver
2º Cómo podemos contestar la pregunta == Qué modelo / test estadístico permite contestarla
3º Definir variables involucradas en la pregunta
4º Recoger los datos adecuadamente (muestreo), pensando en las variables.

Muestreos:

- Aleatorio simple → Elementos de la población son homogéneos respecto a la variable respuesta.
- Estratificado → Elementos de la población están divididos en clases.

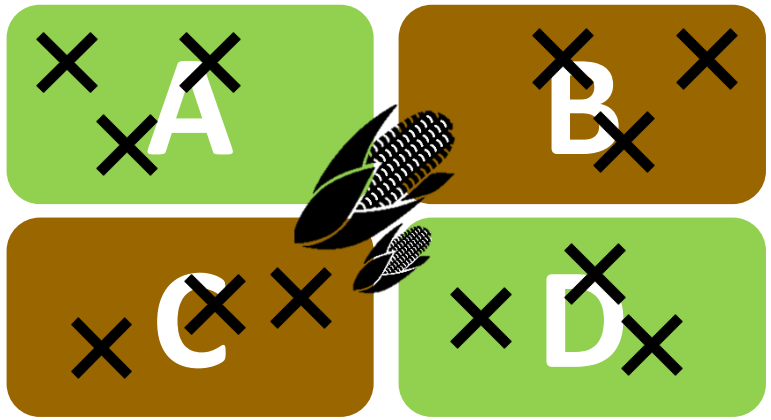
1. Conceptos básicos

1.4. Diseño experimental

1º Definir la pregunta / hipótesis a resolver
2º Cómo podemos contestar la pregunta == Qué modelo / test estadístico permite contestarla
3º Definir variables involucradas en la pregunta
4º Recoger los datos adecuadamente (muestreo), pensando en las variables.

E.g.:
1º ¿El uso de fertilizante provoca cambios en la producción de maíz?
H0: El uso de un fertilizante no provoca cambios en la producción de maíz.
2º Comparar la producción de campos de maíz con y sin fertilizante
3º Producción == gramos de maíz/ m2
Campo fertilizado vs. Sin fertilizar
4º (**Producción ~ Tratamiento**) + Muestreo

- Si los campos con y sin fertilizante son homogéneos, podemos llevar a cabo un muestreo aleatorio simple.
- Sin embargo, si existe heterogeneidad dentro de los tipos de campo, se debe tener en cuenta para que el muestreo sea lo más representativo posible, i.e. muestreo estratificado.



1. Conceptos básicos

1.4. Diseño experimental

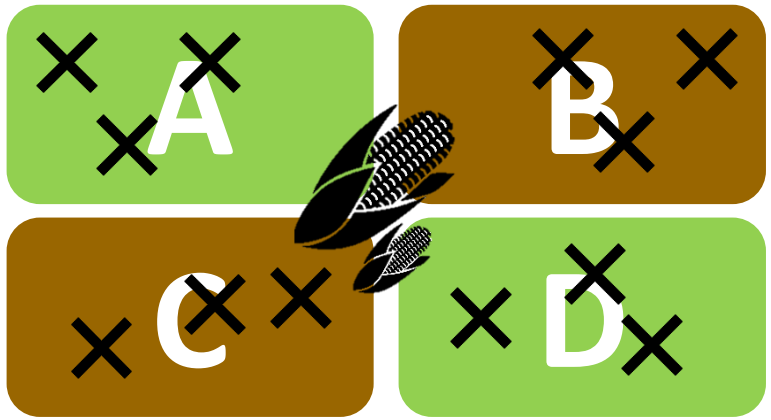
1º Definir la pregunta / hipótesis a resolver
2º Cómo podemos contestar la pregunta == Qué modelo / test estadístico permite contestarla
3º Definir variables involucradas en la pregunta
4º Recoger los datos adecuadamente (muestreo), pensando en las variables.

E.g.:
1º ¿El uso de fertilizante provoca cambios en la producción de maíz?
H0: El uso de un fertilizante no provoca cambios en la producción de maíz.
2º Comparar la producción de campos de maíz con y sin fertilizante
3º Producción == gramos de maíz/ m2
Campo fertilizado vs. Sin fertilizar
4º (**Producción ~ Tratamiento**) + Muestreo + **Recogida de datos** (¡pensando en facilitar su posterior análisis!):

Campo	Con fertilizante	Sin fertilizante
A	15.49	-
B	-	11.25
C	-	13.02
D	18.30	-

Factor
Niveles

Campo	Tratamiento	Producción
A	Fertilizante	15.49
B	Nada	11.25
C	Nada	13.02
D	Fertilizante	18.3



1. Conceptos básicos

1.5. Modelo estadístico

Abstracción simplificada de una realidad compleja → El modelo nunca será capaz de reflejar exactamente la realidad → *residuos (errores)*

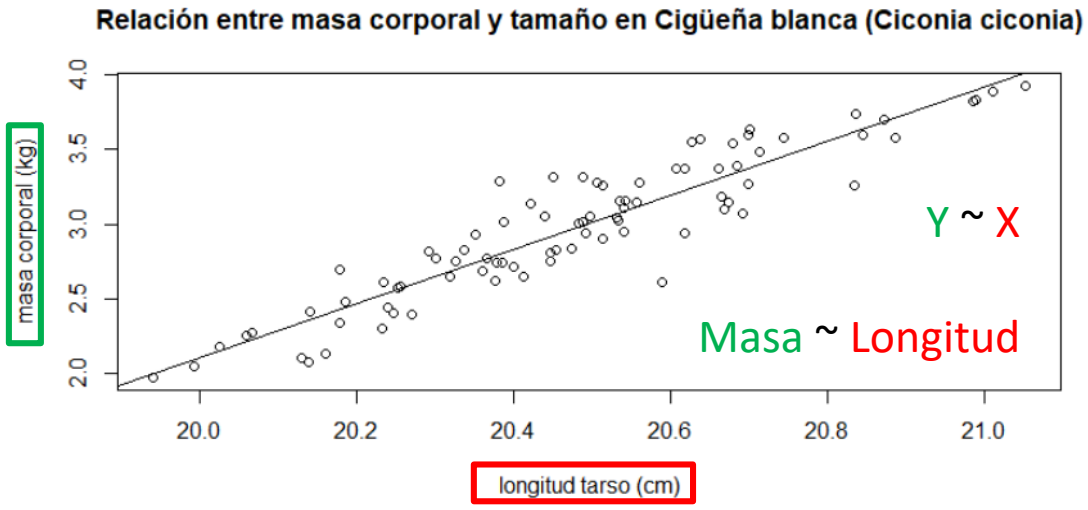
Ecuación matemática que relaciona variables.

1.6. Variable

Característica observable en los individuos de una población, con más de un valor entre individuos.

E.g. Sexo de los participantes de una encuesta, tamaño del tarso de un gorrión, precio de aguacates en la provincia de Soria, …

- Variable explicativa (variable independiente; x): aquella que influye en variable respuesta.
- Variable respuesta (variable dependiente; y): aquella que queremos describir



1. Conceptos básicos

Ejercicios: 1.Ejer_ConceptosBasicos