

Estadística aplicada en R

Conceptos básicos



-Febrero 2021-

Carlota Solano
Álvaro Arredondo

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*

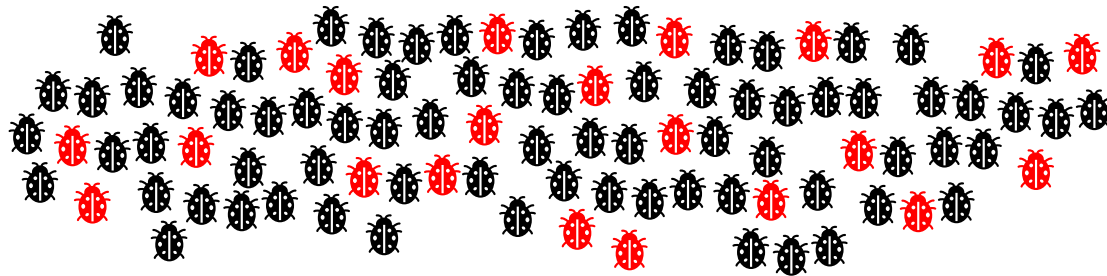
0.3. ¿Dónde se emplea?

- Ingeniería
- Biología
- Economía
- Psicología
- Lingüística
- ...

1. Conceptos básicos

1.1. Descripción de datos

- Tipos de datos \longrightarrow Tipo de test \rightarrow f(datos var. respuesta)
 - Categoricos
 - 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)
- Población vs. **Muestra** \rightarrow porque es imposible muestrear toda una población



1. Conceptos básicos

1.1. Descripción de datos

- Medidas de centralidad:

- Media → 17.69
- Mediana → 7
- Moda → 6

E.g.

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



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



```
>mean()  
>median()
```

1. Conceptos básicos

1.1. Descripción de datos

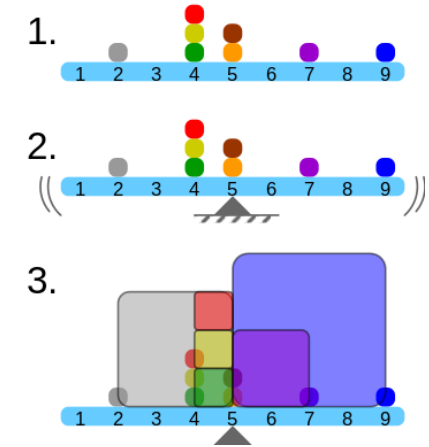
- Medidas de dispersión: variación de los datos de una muestra

- 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

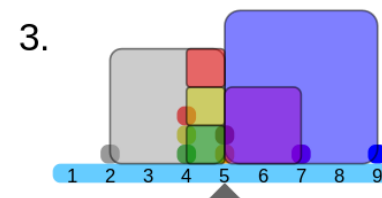
- Medidas de dispersión: variación de los datos de una muestra

- 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)

- Intervalo de confianza (95%CI) $95CI = \pm 1.96 \cdot se$

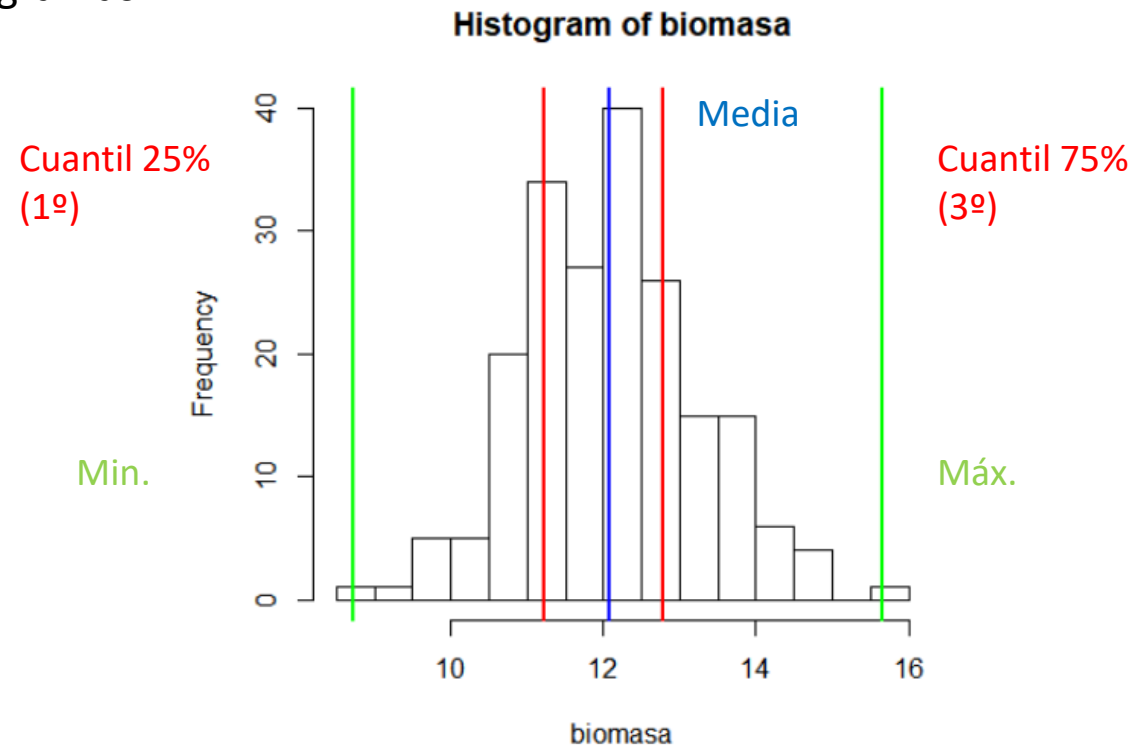
$$se = \sqrt{\frac{s^2}{n}}$$

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

1. Conceptos básicos

1.1. Descripción de datos

- Histogramas



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

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

1. Conceptos básicos

1.2. Hipótesis

Hipótesis nula (H_0): La diferencia/relación esperada no existe

Hipótesis alternativa (H_a): La diferencia/relación esperada existe

“Siempre” esperamos rechazar H_0 y aceptar H_a

¿Por qué H_0 ?

H_0 : Todos los cuervos son negros

H_a : Todos los cuervos no son negros

Falta evidencia para aceptar H_a , pero eso no implica que TODOS los cuervos sean negros → Nuestras observaciones no nos permiten rechazar la H_0 , pero encontrar un solo cuervo de otro color nos permite rechazar H_0



1.3. P-valor

Probabilidad de haber obtenido el resultado obtenido siendo H_0 cierta (i.e. .

p-valor: **0.05** → 5% probabilidad de obtener X resultados siendo H_0 cierta → Significativo estadísticamente
→ rechazamos H_0 y aceptamos H_a

p-valor: 0.8 → 80% probabilidad de obtener X resultados cuando la H_0 sea cierta → No significativo estadísticamente
→ no podemos rechazar H_0 y no aceptamos H_a → **Solo encontramos evidencia contra el rechazo de H_0**

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.

Muestreos:

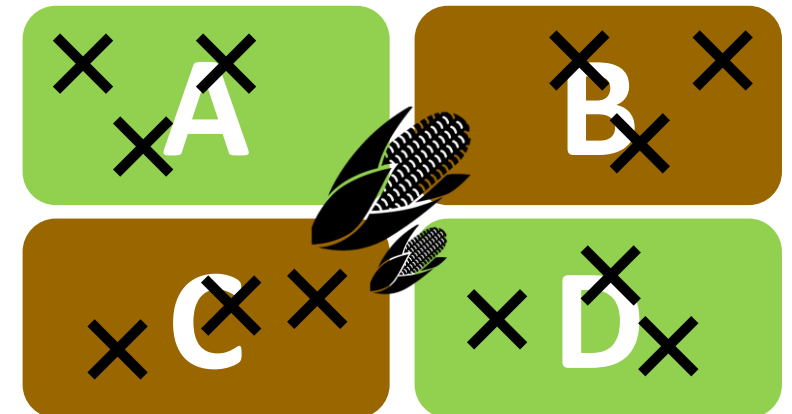
- Aleatorio simple → Elementos de la población son homogéneos respecto a la variable respuesta.
- Estratificado → Elementos de la población están divididos en clases.

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
Tratamiento = 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.

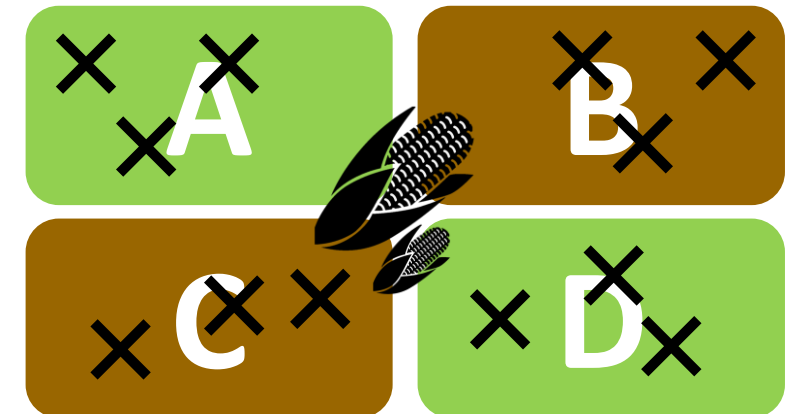
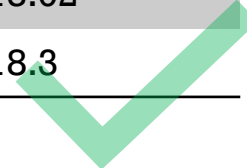
E.g.:

- 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 | - |



| 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)*

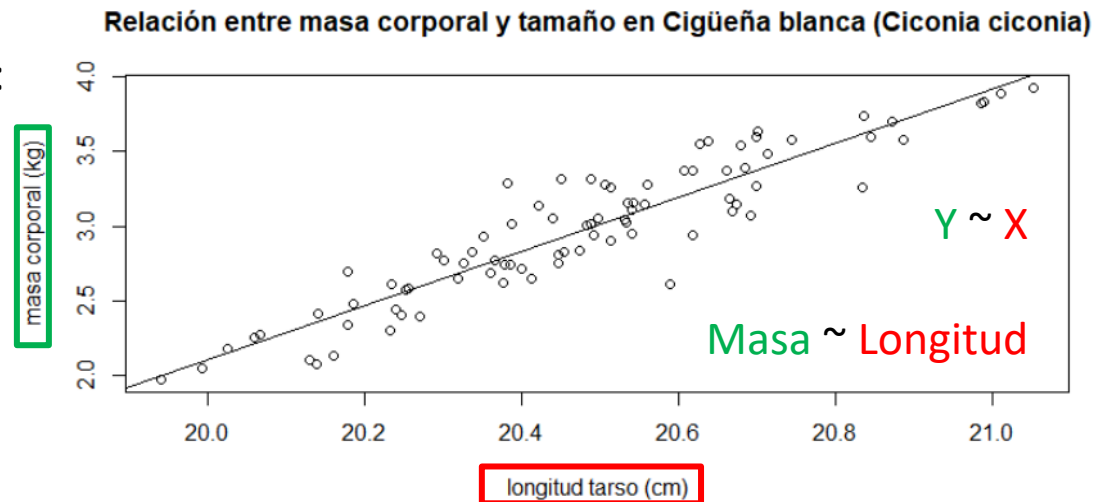
Ecuación matemática que relaciona variables.

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 influyen en variable respuesta.
- **Variable respuesta** (variable dependiente; y): aquella que queremos describir



1. Conceptos básicos

Ejercicios: 1.Ejer_ConceptosBasicos