Estadística descriptiva

Bioestadística

Marta Coronado (marta.coronado@uab.cat)

Grado en Genética | Curso 2025/26

Facultat Biociències UAB

Outline

1. Variables

- ¿Qué es una variable?
- Tipos de variables
- Variables y escalas

2. Estadística descriptiva

- Métodos
- Parámetros de tendencia central
- Parámetros de posición
- Parámetros de dispersión
- En R

01

Variables

Variables

Una **variable** representa **información** acerca de diferentes **características** de un sistema de estudio (**población**).

Muestra	Variable	Unidad de observación
150 bebés nacidos en un determinado hospital	Peso al nacer (kg)	Un bebé
73 polillas *Cecropia* capturadas en una trampa	Sexo	Una polilla
81 plantas descendientes de un cruce parental único	Color de la flor	Una planta
Colonias bacterianas en cada uno de seis platos de Petri	Número de colonias	Un plato de Petri

Variables

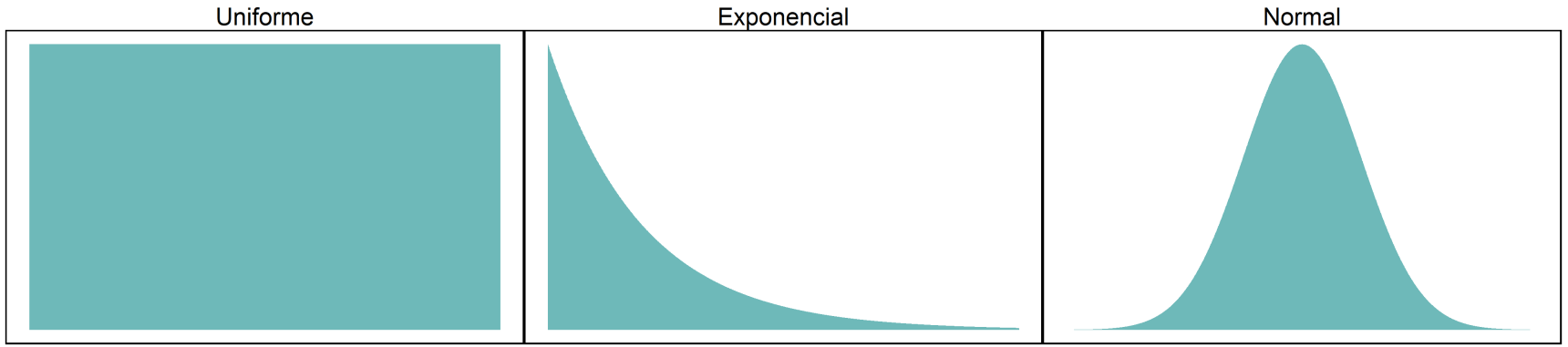
Una **variable** representa **información** acerca de diferentes **características** de un sistema de estudio (**población**).

- El muestreo siempre es **limitado**: solo podemos observar una fracción de todos los valores posibles que existen para una variable.
- Cada variable sigue una cierta **distribución** de valores, y cada medición que realizamos se espera que refleje, en mayor o menor grado, esa **distribución global**.

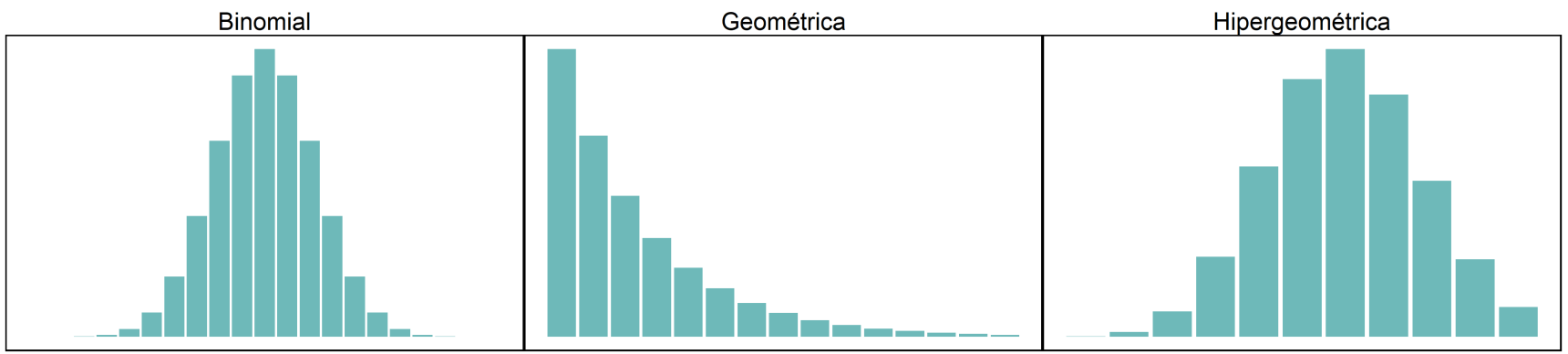
Muestra	Variable	Unidad de observación
150 bebés nacidos en un determinado hospital	Peso al nacer (kg)	Un bebé
73 polillas *Cecropia* capturadas en una trampa	Sexo	Una polilla
81 plantas descendientes de un cruce parental único	Color de la flor	Una planta
Colonias bacterianas en cada uno de seis platos de Petri	Número de colonias	Un plato de Petri

Distribuciones según el tipo de variable

Variables cuantitativas



Variables cualitativas



Tipos de variables y sus escalas

Variables categóricas o cualitativas

- Las escalas pueden ser:
  - Nominales
  - Ordinales

Variables continuas o cuantitativas

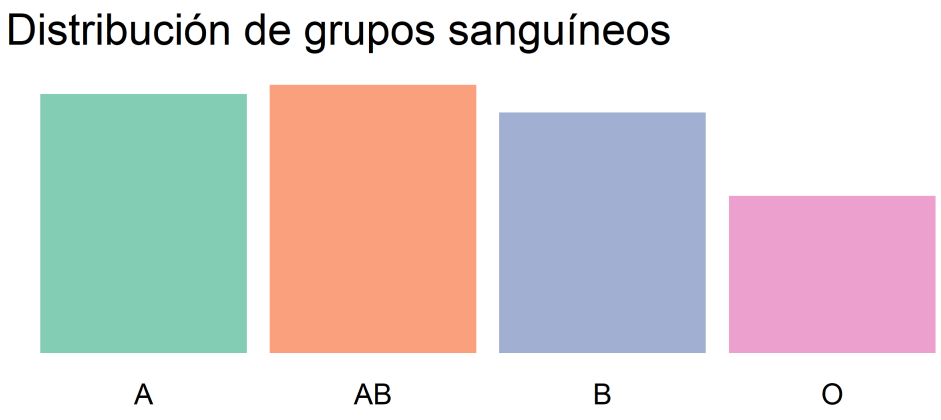
- Las escalas pueden ser:
  - Discretas
  - Continuas

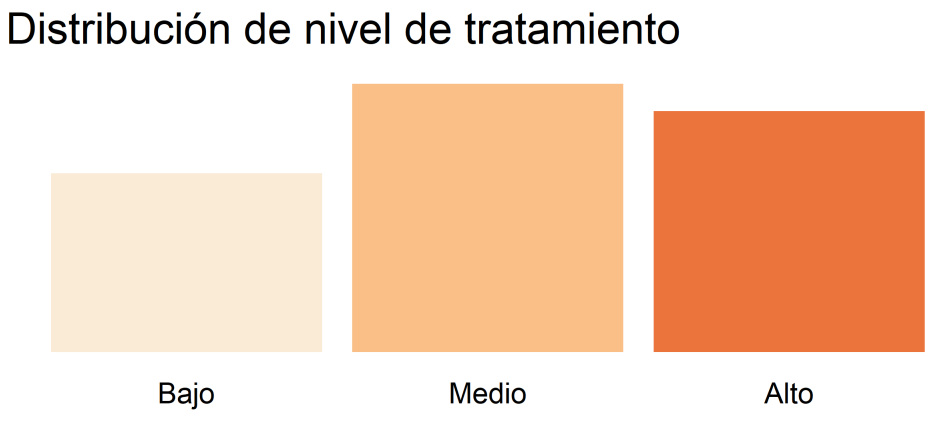
Nivel de medición	Nominal	Ordinal	Intervalo	Ratio
Categoriza las variables	✓	✓	✓	✓
Ordena las categorías en un orden		✓	✓	✓
Tiene intervalos conocidos			✓	✓
Tiene un 0 real o con significado				✓

Variables categóricas (cualitativas)

Las variables categóricas son aquellas que permiten clasificar la información en grupos o categorías claramente diferenciados.

- Solo pueden tomar un número limitado de valores
- Cada elemento de la población se asigna a uno de esos grupos finitos
- En algunos casos, esas categorías pueden tener un orden lógico (ordinales)





Variables categóricas (cualitativas)

Nominales

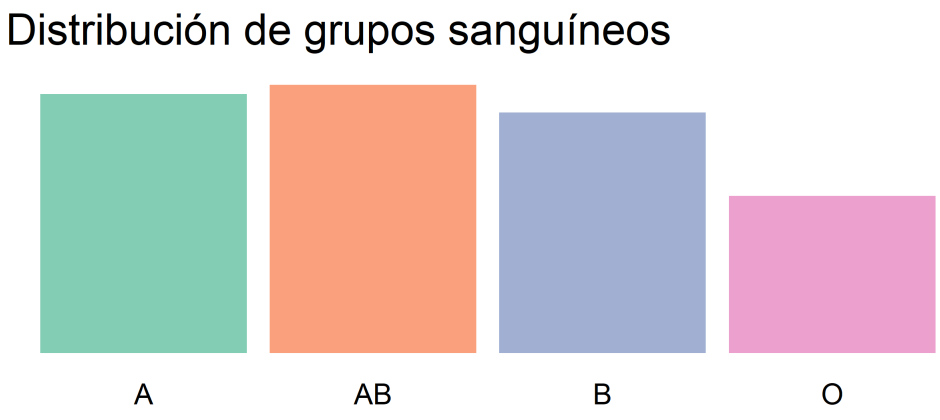
Los valores no se pueden ordenar.

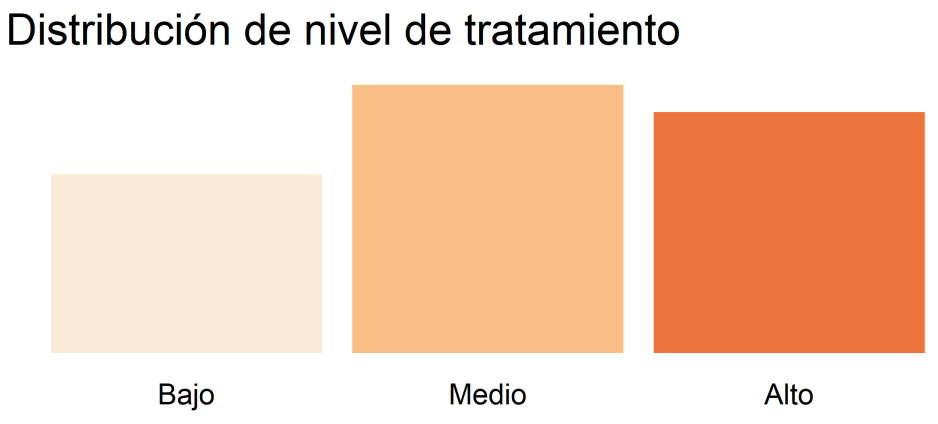
- Tipo sanguíneo (A, B, AB, O)
- Sexo (macho, hembra)
- Tipo de suelo (arenoso, arcilloso, permafrost, etc.)
- Forma de una semilla (arrugada, lisa, etc.)

Ordinales

Los valores se pueden ordenar.

- Intensidad de dolor (bajo, medio, alto)
- Respuesta a un tratamiento (ninguno, parcial, completo)





En R, este tipo de variables suelen representarse como `factor` o como `character`.

Variables categóricas (cualitativas)

En R, este tipo de variables suelen representarse como `factor` o como `character`.

```
# Variable categórica nominal (no tiene orden)
color_flores <- factor(c("rojo", "blanco", "rosa", "rojo", "blanco"))
levels(color_flores)
```

```
## [1] "blanco" "rojo"   "rosa"
```

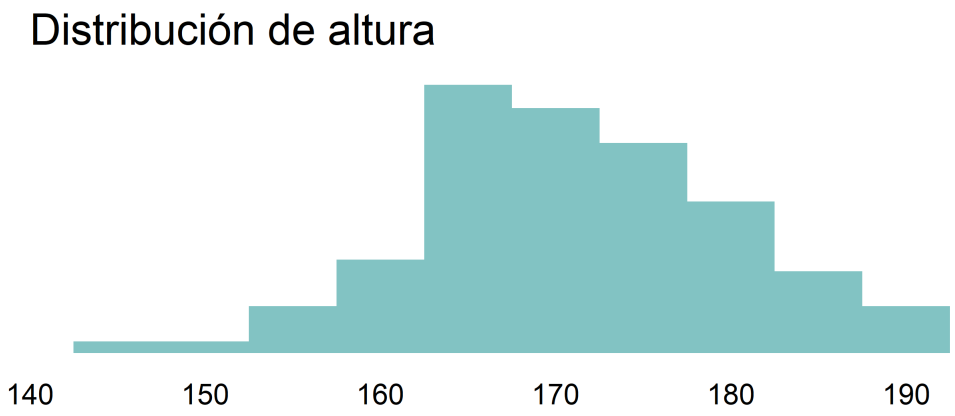
```
# Variable categórica ordinal (tiene un orden lógico)
dolor <- factor(
  c("bajo", "medio", "alto", "bajo", "alto"),
  levels = c("bajo", "medio", "alto"), ordered = TRUE
)
levels(dolor)
```

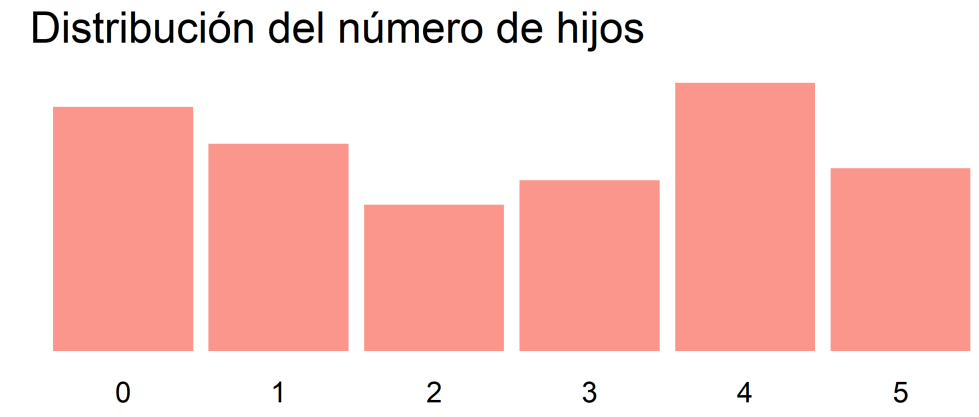
```
## [1] "bajo"  "medio" "alto"
```

Variables continuas (cuantitativas)

Las variables continuas son aquellas que pueden tomar un rango de valores posibles, no solo valores específicos.

- Pueden tener infinitos valores posibles dentro de ese rango.
- Pueden variar de manera gradual, pudiendo tener un valor diferente para cada unidad o medida.
- Siempre se pueden ordenar de menor a mayor.





Variables continuas (cuantitativas)

Continua

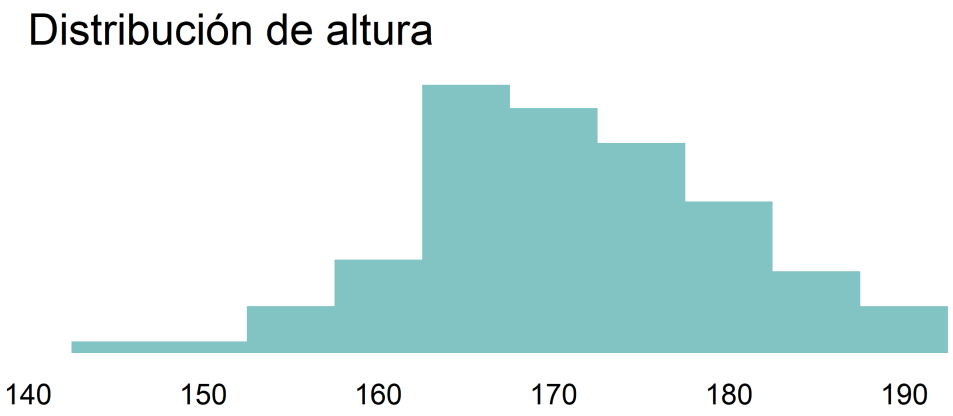
Variable numérica que se mide en una escala continua (incluyendo decimales).

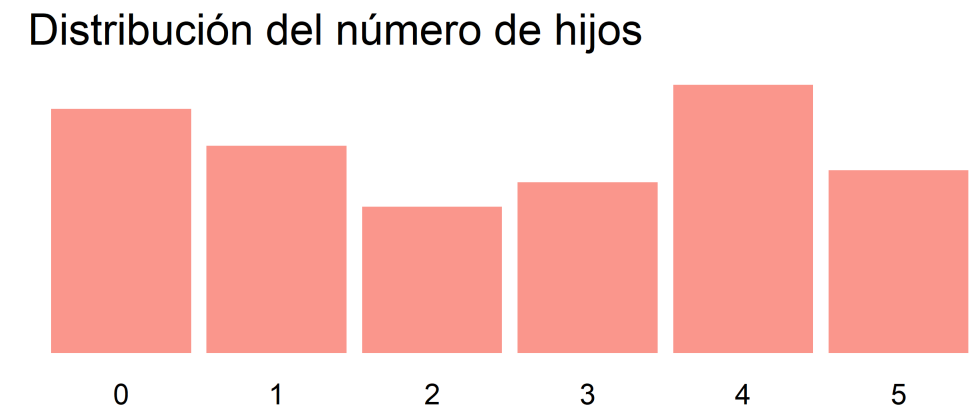
- Peso de un bebé
- Concentración del colesterol en sangre

Discreta

Variable numérica para la cual podemos enumerar todos los posibles valores (no incluye decimales).

- Número de colonias de bacterias en una placa de Petri
- Número de ganglios linfáticos cancerosos detectados en un paciente





En R, las variables continuas normalmente se almacenan como numéricas (`numeric` o `integer`).

Variables continuas (cuantitativas)

En R, las variables continuas normalmente se almacenan como numéricas (`numeric` o `integer`).

```
# Crear un vector de temperatura (con decimales)
temperatures <- c(23.5, 30.2, 45.8, 50.1, 60.0)

# Comprobar tipo de objeto
class(temperatures)
```

```
## [1] "numeric"
```

```
# Crear un vector de mutaciones observadas en distintas secuencias
dna_mutations <- c(0L, 3L, 12L, 1L, 8L)

# Comprobar tipo de objeto
class(dna_mutations)
```

```
## [1] "integer"
```

```
dna_mutations
```

```
## [1]  0  3 12  1  8
```

Conversión de variables continuas a categóricas

Las **variables continuas** se pueden transformar en **variables categóricas** mediante un proceso llamado *binning*:

- Consiste en dividir el **rango de valores de la variable** en varios "intervalos" o **categorías** (*bins*).
- Se puede decidir cuántos intervalos queremos crear según la necesidad del análisis.

Ejemplo:
Supongamos que tenemos un rango de temperaturas de −2.15°C a 17.85°C, y queremos dividirlo en 4 intervalos de igual tamaño:

- Intervalo 1: $-2.15°C \leq X \leq 2.85°C$
- Intervalo 2: $2.85°C < X \leq 7.85°C$
- Intervalo 3: $7.85°C < X \leq 12.85°C$
- Intervalo 4: $12.85°C < X \leq 17.85°C$

> Nota: Una variable continua puede representarse tanto como **continua** como **categórica**, mientras que una variable categórica **solo puede ser categórica**.

Variables según su escala

Otra forma de clasificar las variables es según la **escala** en la que se representan.

> **Cada escala de variable** requiere **procedimientos estadísticos distintos** para su análisis. Las principales escalas de variables son:

- Nominal
- Binaria
- Ordinal
- De intervalo (*Interval*)
- De razón o relación (*Ratio*)

Nivel de medición	Nominal	Ordinal	Intervalo	Ratio
Categoriza las variables	✓	✓	✓	✓
Ordena las categorías en un orden		✓	✓	✓
Tiene intervalos conocidos			✓	✓
Tiene un 0 real o con significado				✓

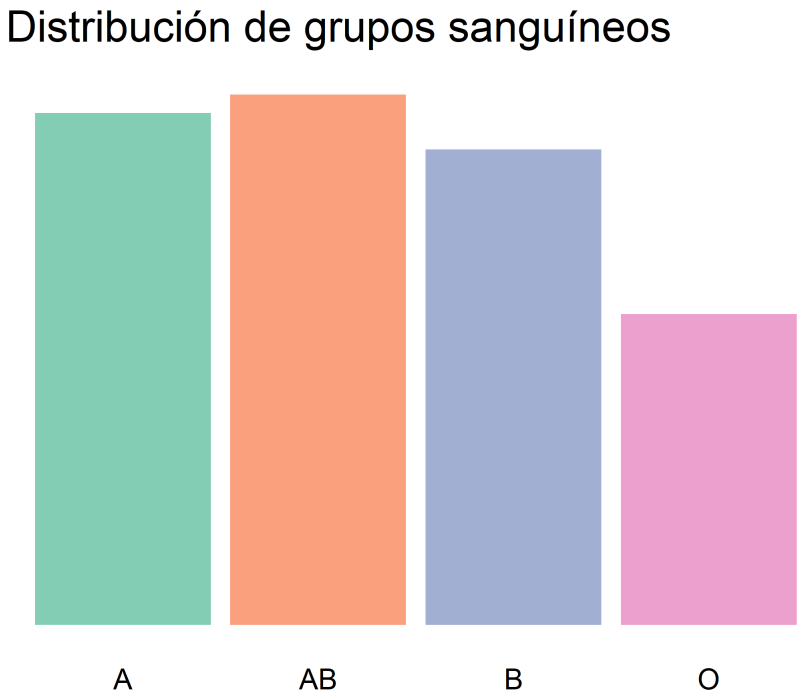
Escalas nominal y binaria

Las **escalas nominales** de variables corresponden a *variables categóricas* que no se pueden ordenar de una manera lógica/con significado.

Ejemplos:

- Color del pétalo (rojo, verde, azul, etc.)
- Identificador

> Las **escalas binarias** son un caso especial de las nominales, solo pueden tomar dos valores posibles: 0 y 1.

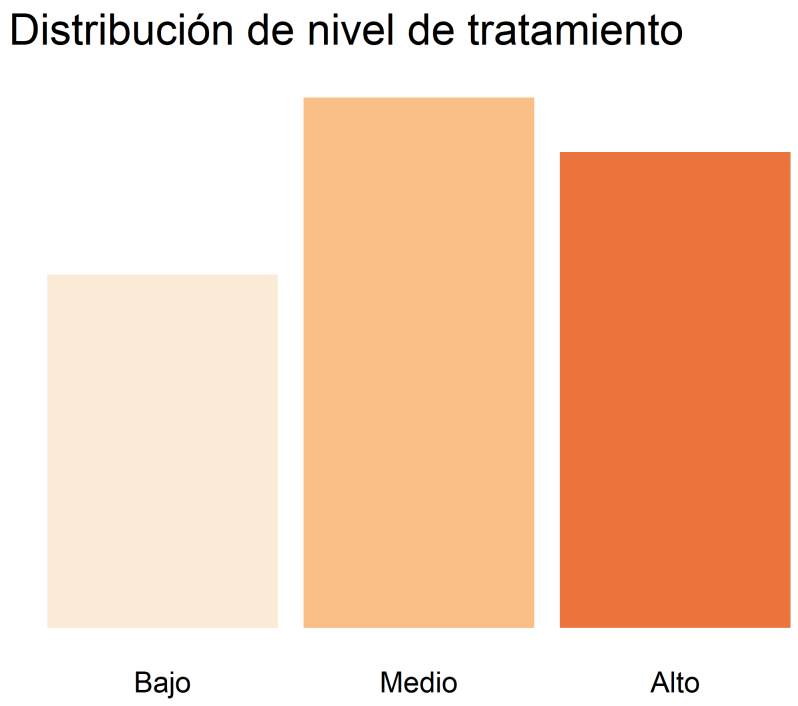


Escala ordinal

La escala ordinal de variables corresponden a *variables categóricas* que se pueden ordenar de una manera lógica/con significado.

Ejemplos:

- Tamaño (pequeño, mediano, grande, etc.)
- Variables continuas *binneadas*



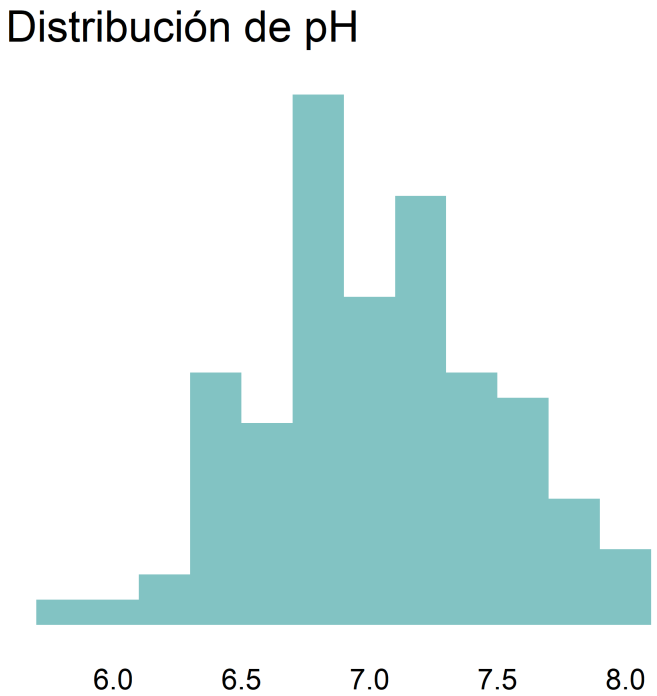
Escala de intervalo

Las escalas de intervalo corresponden a una mezcla de variables continuas.

- Las variables en escalas de intervalo se miden en intervalos iguales a partir de un punto de origen definido.
- El punto de origen (por ejemplo, 0) no significa ausencia de la característica que se está midiendo.

Ejemplos:

- Temperatura (°C)
- pH



Escala de relación/proporción

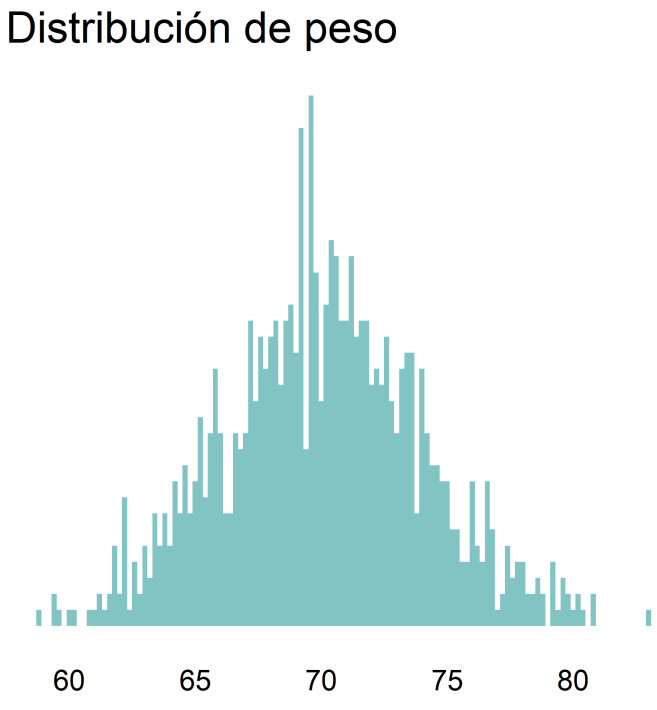
Las escalas de razón/proporción de las variables corresponden a variables continuas.

- Las variables en escalas de razón/proporción se miden en intervalos iguales a partir de un punto de origen definido.
- El punto de origen sí indica ausencia de la característica que se está midiendo.

> Las escalas enteras son un caso especial de las escalas de razón, ya que solo permiten números enteros.

Ejemplos:

- Temperatura (K)
- Peso



Temperatura en Celsius *vs.* Kelvin

Un cambio de 1 grado Celsius es equivalente a un cambio de 1 Kelvin, porque ambos comparten la misma magnitud de intervalo, **pero las escalas tienen puntos cero diferentes**:

Celsius como escala de intervalo

- **Cero arbitrario:** 0°C es el punto de congelación del agua, un punto de referencia arbitrario.
- **Sin cero verdadero:** Como no tiene un cero verdadero, no puedes decir que 30°C es el doble de 15°C.
- **Intervalos iguales:** La diferencia entre 1°C y 2°C es la misma que la diferencia entre 25°C y 26°C.

Kelvin como escala de razón

- **Cero verdadero:** 0K representa el cero absoluto, la temperatura más baja posible, donde hay ausencia completa de energía térmica.
- **Relaciones significativas:** Gracias al cero verdadero, los ratios son significativos, así que 20K es el doble de caliente que 10K.
- **Mismos intervalos que Celsius:** Un cambio de 1K representa la misma diferencia de temperatura que 1°C.

02

Estadística descriptiva

Introducción a la estadística descriptiva

La estadística descriptiva se utiliza para **resumir los datos**.

- **Objetivo:** Describir un conjunto de datos dado (n) respecto a una variable (p_j) o a un conjunto de variables (p).
- **Procedimiento:** Utilizar un conjunto de métodos adecuadamente seleccionados para resumir o visualizar los datos disponibles.

Las características de las variables se expresan con frecuencia mediante **parámetros**.

Métodos y particularidades

En R, solemos trabajar con tablas (`data.frames`) de $n \times p$ (número de filas × número de columnas).

Esta información se utiliza para calcular parámetros informativos:

Medidas de tendencia central

- Media aritmética
- Mediana
- Moda

Parámetros de dispersión (medidas de dispersión)

- Rango (mínimo-máximo)
- Varianza
- Desviación estándar
- Rango intercuartílico

Parámetros de posición

- Cuartiles, deciles, percentiles, ...

Parámetros y su significado

¿Qué es un parámetro?

En estadística descriptiva, un parámetro proporciona información sobre la **forma de la distribución** de los valores de una determinada variable.

¿Por qué es importante?

Los parámetros se pueden utilizar para **resumir las propiedades de los datos** y hacer que grandes conjuntos de datos, con múltiples valores por variable, sean más accesibles.

Entonces...

Para saber qué parámetros usar, es necesario **conocer cuáles existen y cómo calcularlos**.

> En esencia, los parámetros son datos **ya digeridos o resumidos**.

Vamos a generar unos datos

Para el cálculo de distintos parámetros de estadística descriptiva, necesitaremos datos:

```
set.seed(42) # para hacer el código reproducible
data_vec <- rexp(100, rate = 0.1) + 10
```

```
##          [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]   [,8]   [,9]  [,10]
##  [1,] 11.983 23.447 58.628 15.694 17.783 11.632 14.682 67.089 23.357 35.117
##  [2,] 16.609 34.087 16.658 40.186 16.563 26.495 20.254 12.179 11.470 25.739
##  [3,] 12.835 10.961 59.960 14.976 13.864 21.293 10.301 16.736 11.794 33.637
##  [4,] 10.382 10.571 12.236 13.546 24.161 12.913 14.879 15.030 14.393 22.016
##  [5,] 14.732 22.540 22.114 27.564 25.231 15.516 12.383 20.645 22.730 16.345
##  [6,] 24.636 13.098 17.191 17.375 14.387 11.273 12.525 76.523 11.619 27.476
##  [7,] 13.140 14.752 23.085 10.288 51.681 24.456 25.364 21.158 14.490 11.033
##  [8,] 14.101 16.221 12.800 13.512 21.951 19.140 15.128 12.243 14.183 14.388
##  [9,] 21.916 22.456 12.311 13.927 10.288 13.349 14.665 16.579 14.019 19.895
## [10,] 17.149 13.703 22.871 29.021 78.466 94.623 10.715 21.200 16.428 15.576
```

Media aritmética

Definición:
También llamada **promedio**, esta métrica es el **promedio matemático** de los valores de los datos proporcionados.

> No es resistente a valores atípicos (*outliers*) ni a distribuciones asimétricas.

Cálculo:

$$\bar{x} = \mu = \frac{1}{n} \sum_{i=1}^{n} x_i$$

donde:

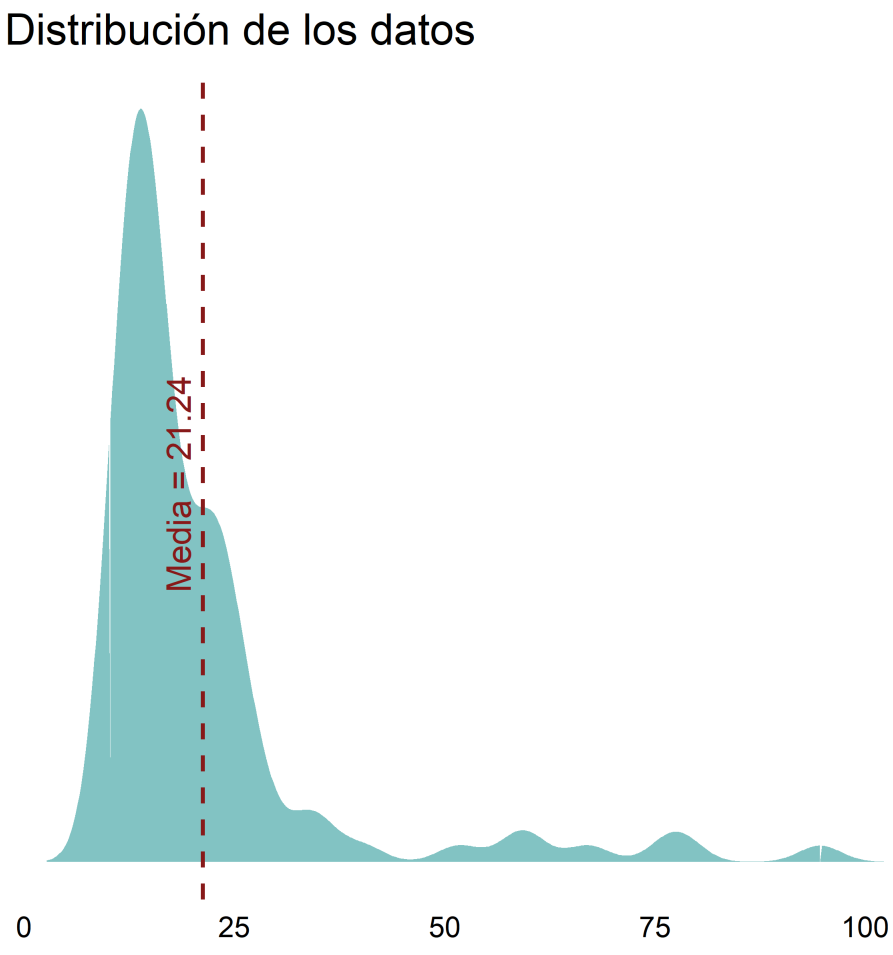
- $\bar{x} = \mu$: media aritmética
- n : número de muestras (igual al número de valores de la variable en cuestión)
- i : índice de los valores de la variable ($i = 1, 2, \ldots, n$)
- x_i : valor i-ésimo de la variable x

Media aritmética

La media aritmética se calcula aplicando la función `mean()` de `R base`.

```
# calculo
mean(data_vec)
```

```
## [1] 21.2431
```



Mediana

Definición:
La **mediana** es el valor que separa la mitad superior de los datos de la mitad inferior.

> Es resistente a **valores atípicos (*outliers*)** y a distribuciones asimétricas.

Cálculo

- Para un número **impar** de observaciones (n):

$$\text{mediana}(x) = \left(\frac{n+1}{2}\right)^{\text{th}}$$

- Para un número **par** de observaciones:

$$\text{mediana}(x) = \frac{\left(\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th}\right)}{2}$$

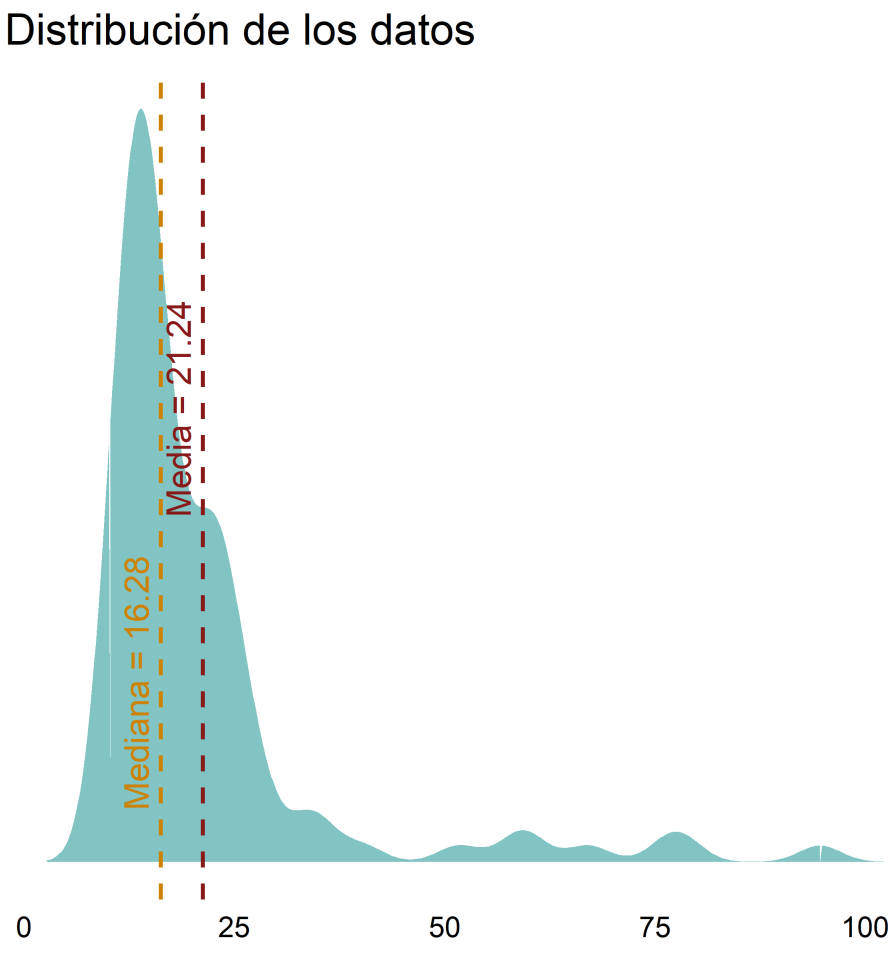
- $\text{mediana}(x)$: mediana de los valores disponibles de la variable x.
- n: número de observaciones disponibles para x.

Mediana

La mediana se calcula utilizando la función `median()` de `R base`.

```
# calculo
median(data_vec)
```

```
## [1] 16.28286
```



Moda

Definición:
La **moda** de un conjunto de datos es el valor que aparece con mayor frecuencia.

> Es resistente a **valores atípicos (outliers)**, aunque la **forma de la distribución** puede ser crucial.

Cálculo

$$mode(x) = max_{k=1}\left(I_{i=1}(x_i = x_k)\right)$$

donde

- $\text{moda}(x)$: moda de los valores disponibles de la variable x.
- $\text{max}_{k=1}()$: argumento que maximiza sobre k en 1 a p.
- $I_i()$: función indicadora que devuelve 1 si la condición interna es cierta para i en 1 a n.

Moda

Una forma de calcular la moda en R es usando las funciones `max()` y `table()` de base R.

```
# cuentas de valores en un vector redondeado (0 decimales)
table <- table(round(data_vec))
table
```

```
##
## 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 34 35 40 52 59 60
##  4  6  9  7 12  8  6  8  1  1  2  4  5  6  2  3  2  1  1  1  2  1  1  1  1  1
## 67 77 78 95
##  1  1  1  1
```

```
# número máximo de apariciones de un mismo valor
max <- max(table)
max
```

```
## [1] 12
```

Moda

Una forma de calcular la moda en R es usando las funciones `max()` y `table()` de base R.

```
# posición del valor máximo en la tabla
pos <- which(table == max)
pos
```

```
## 14
##  5
```

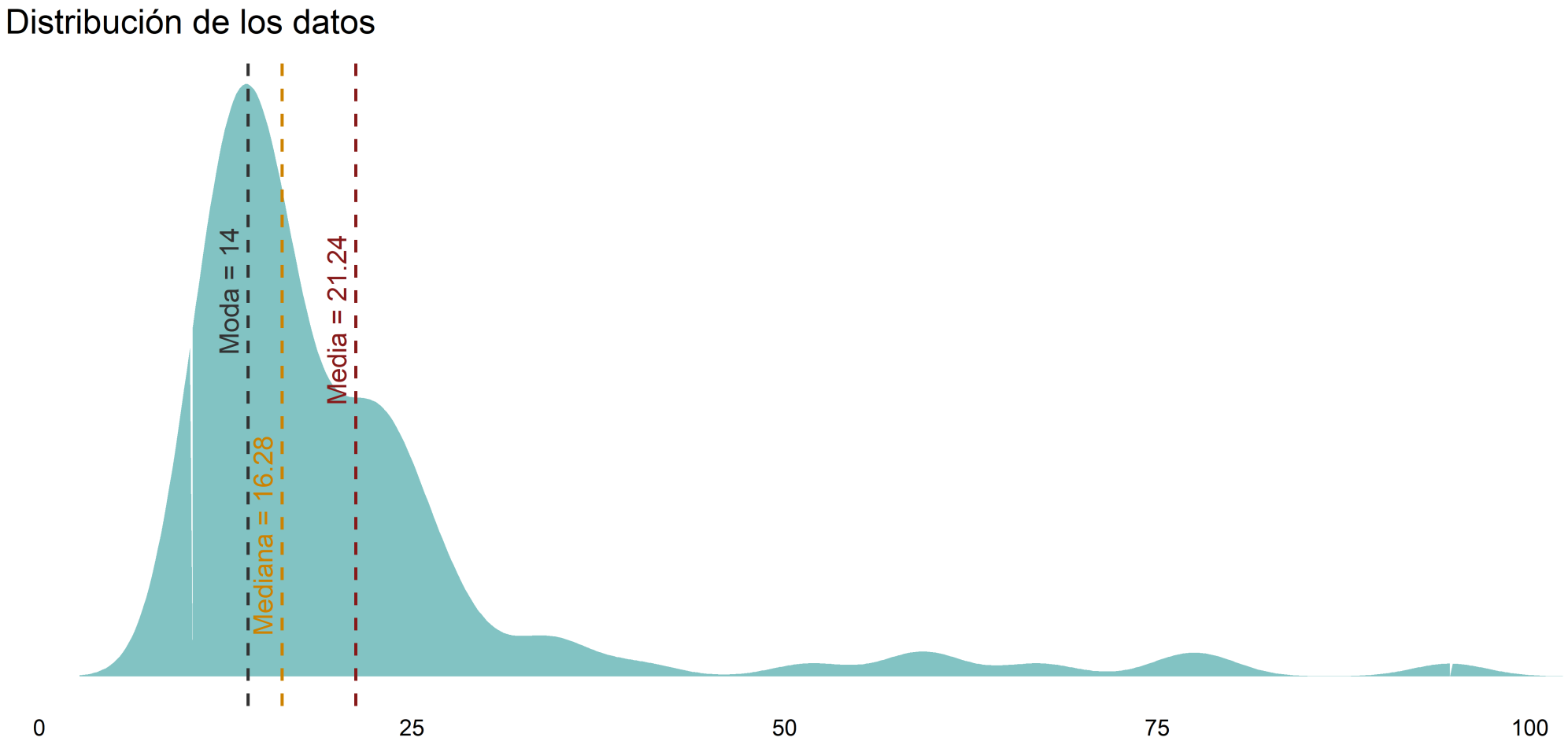
```
# valor en la posición máxima
mode <- names(table)[pos]
as.numeric(mode) # moda
```

```
## [1] 14
```

> 🚀 Reto: haz una función que se llame "`moda()`" en R con el código anterior.

```
moda(data_vec)
```

Moda


Distribución de los datos
Moda = 14
Mediana = 16.28
Media = 21.24
0
25
50
75
100

¿Qué parámetro de posición usar?

Todas las medidas de tendencia central describen la posición central de la distribución de frecuencias de los valores de una variable en el conjunto de datos.

- La **media aritmética** solo es realmente útil cuando la distribución de los datos es **simétrica**.
- La **mediana** se comporta de manera **robusta** frente a distribuciones **asimétricas** de los datos.
- La **moda** es más aplicable en **contextos de clasificación** y se usa con poca frecuencia (se utiliza con variables cualitativas).

> En general, la **mediana suele ser suficiente** en la mayoría de los análisis descriptivos.

Rango (mínimo-máximo)

Máximo: el valor más alto disponible para una variable dada.

```
max(data_vec)
```

```
## [1] 94.62336
```

Mínimo: el valor más bajo disponible para una variable dada.

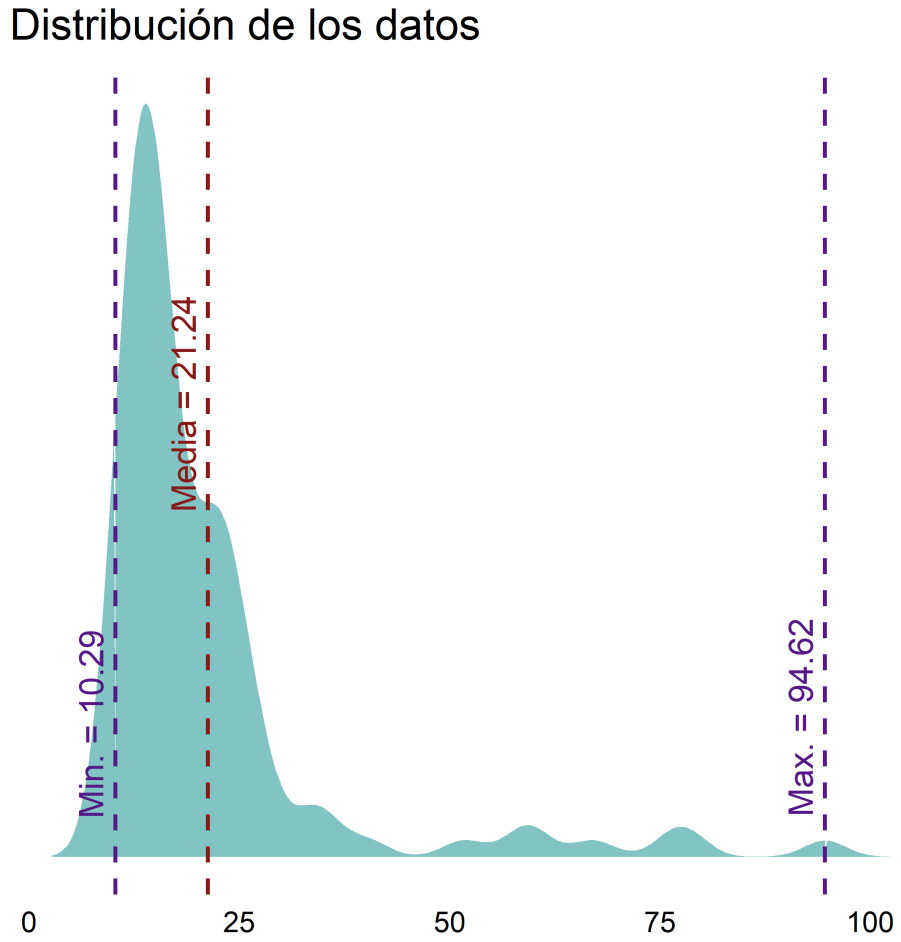
```
min(data_vec)
```

```
## [1] 10.28815
```

Rango: la amplitud de valores que abarca la distribución de los datos.

```
range(data_vec)
```

```
## [1] 10.28815 94.62336
```



Varianza

Definición:
La varianza mide **qué tanto se dispersan los valores de los datos respecto a su valor promedio**.
No es resistente a **valores atípicos (*outliers*)** ni a **distribuciones asimétricas**.

Cálculo:

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

donde

- s^2: varianza
- n: número de muestras (igual al número de valores de la variable en cuestión)
- i: índice de los valores de la variable ($i = 1, 2, \ldots, n$)
- x_i: valor i-ésimo de la variable x
- $\bar{x}$: media aritmética

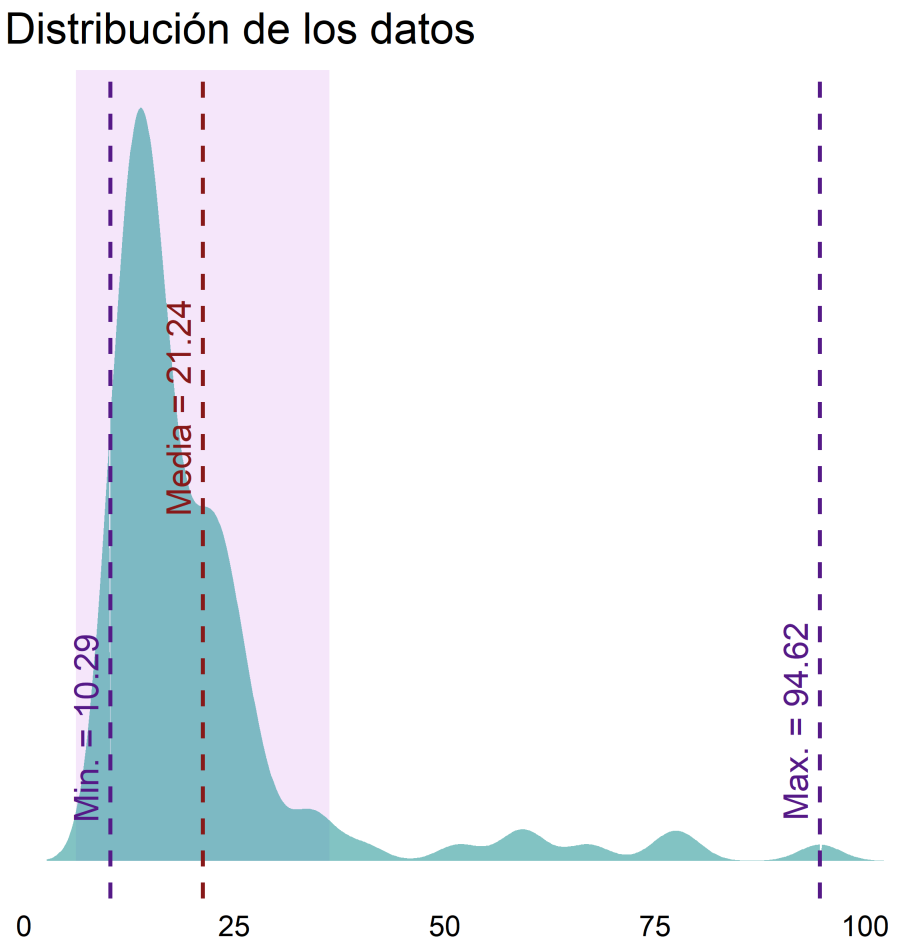
Varianza

La varianza se calcula usando la función `var()` de `base R`.

```
var(data_vec)
```

```
## [1] 226.5367
```

La varianza se mide en unidades cuadráticas.



Desviación estándar

Definición:
La desviación estándar cuantifica la **cantidad de variación o dispersión** de un conjunto de valores de datos.

> No es resistente a **valores atípicos (outliers)** ni a **distribuciones asimétricas**.

Cálculo:

$$SD = s = \sqrt{s^2}$$

donde

- $SD = s$: desviación estándar
- s^2: varianza

Desviación estándar

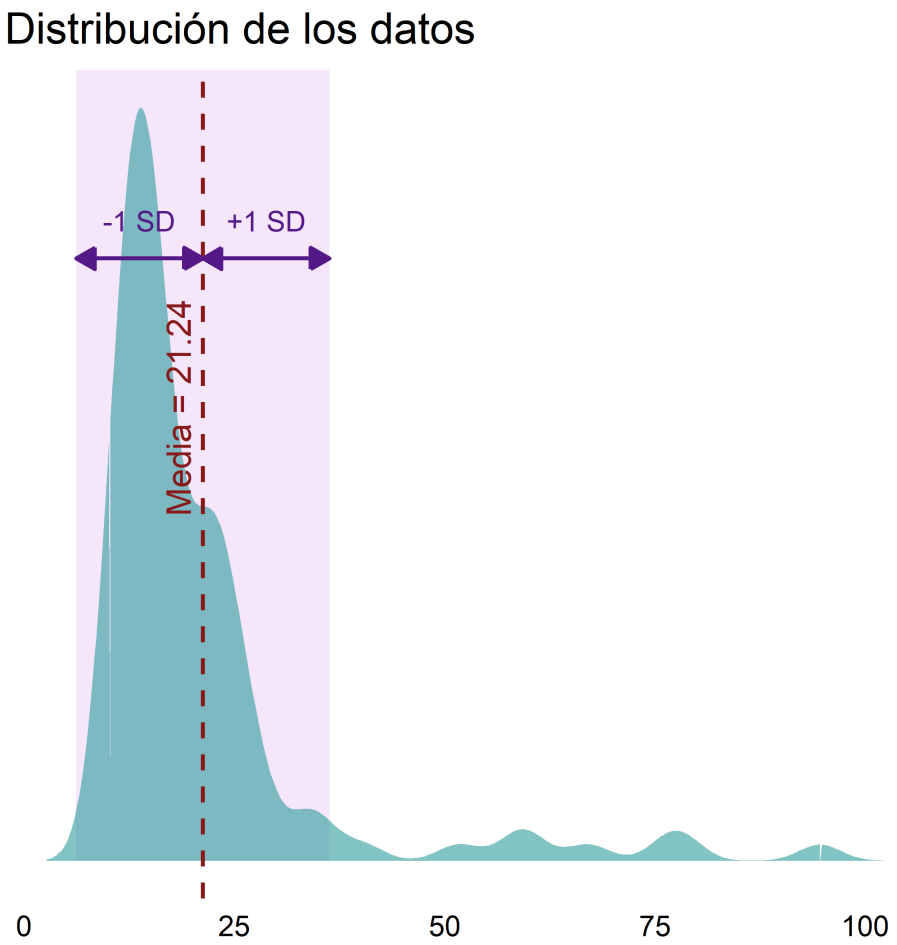
La desviación estándar se calcula usando la función `sd()` de `base R`.

```
sd(data_vec)
```

```
## [1] 15.05114
```

La desviación estándar sí está en la misma unidad que los datos.

El gráfico muestra el intervalo de una desviación estándar por encima y por debajo de la media.



Rango intercuartílico

Definición:
Los cuantiles son **puntos de corte** que dividen el rango de una distribución de valores en intervalos adyacentes con **igual probabilidad**.

Al calcular cuantiles, siempre se obtiene **un punto de corte menos** que el número de cuantiles producidos.

> Son resistentes a **valores atípicos (outliers)** y a **distribuciones asimétricas**.

Cuantiles más utilizados:

- **Cuantil 50**: corresponde a la **mediana**.
- **Cuantil 25 y 75**: también conocidos como **cuartiles**.

Rango intercuartílico

Los cuantiles se calculan usando la función `quantile()` de **base R**.

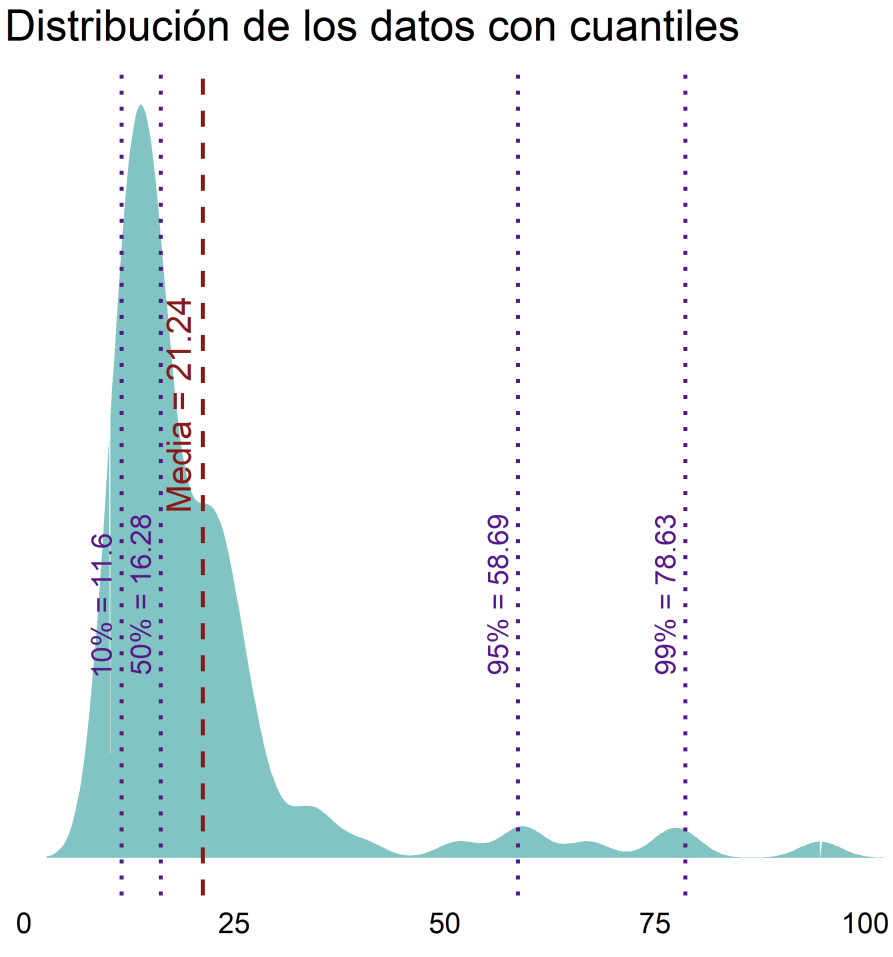
Se puede indicar un **segundo argumento** para especificar los cuantiles que queremos obtener.

```
# Cuantiles que queremos calcular
q <- c(0.25, 0.5, 0.95, 0.99)

# Cálculo de los cuantiles
quantile(data_vec, q)
```

```
##      25%      50%      95%      99%
## 13.29636 16.28286 58.69463 78.62730
```

```
median(data_vec)
```

```
## [1] 16.28286
```



¿Qué parámetro de dispersión debo usar?

Todas las medidas de dispersión describen la **extensión de la distribución de frecuencias** de los valores de una variable en el conjunto de datos.

- La **varianza** solo es realmente útil cuando la distribución de los datos es **simétrica**.
- La **desviación estándar** también es útil principalmente para distribuciones **simétricas**.
- Los **cuantiles** se comportan de manera **robusta** frente a distribuciones **asimétricas**.

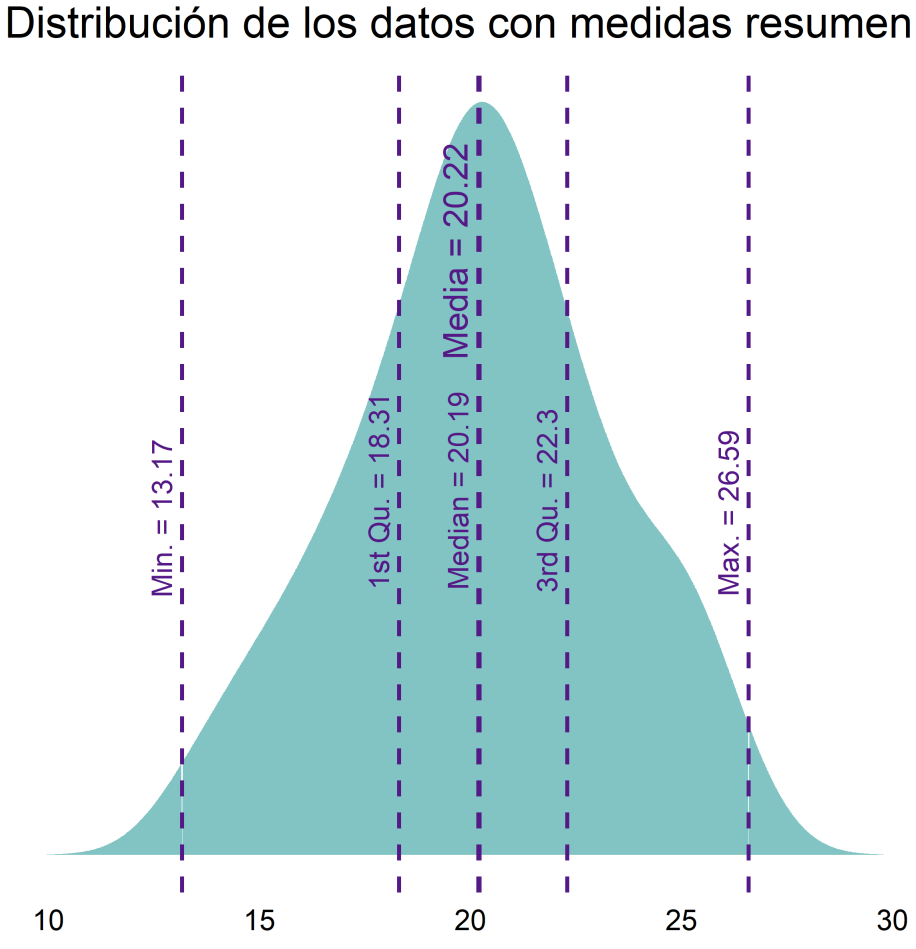
> En la práctica, el **rango intercuartílico** suele ser suficiente para describir la dispersión de los datos, salvo que se busque una interpretación ligada a la media (donde se prefiere la **desviación estándar**).

La función `summary()`

La función `summary()` se puede usar sobre un vector en R para obtener información rápida sobre medidas de tendencia central y medidas de dispersión.

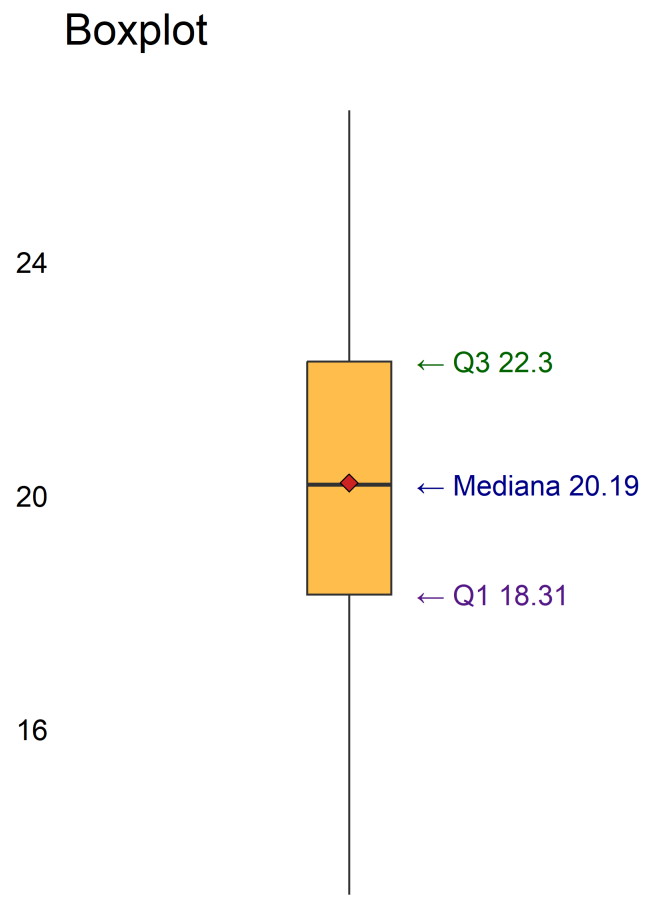
```
# Cálculo de resumen
set.seed(21)
data_vec <- rnorm(100, mean = 20, sd = 3)

summary(data_vec)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.17   18.31   20.19   20.22   22.30   26.59
```



Diagramas de Tukey (boxplot)

- Resumen de la información con **5 números**:
  - Mínimo, máximo y **cuartiles**.
  - Suelen dar una **buena idea de la distribución** de los datos.
- La **zona central**, llamada “caja” (**box**), contiene el **50% central de las observaciones**.
  - Su tamaño se llama **rango intercuartílico**.



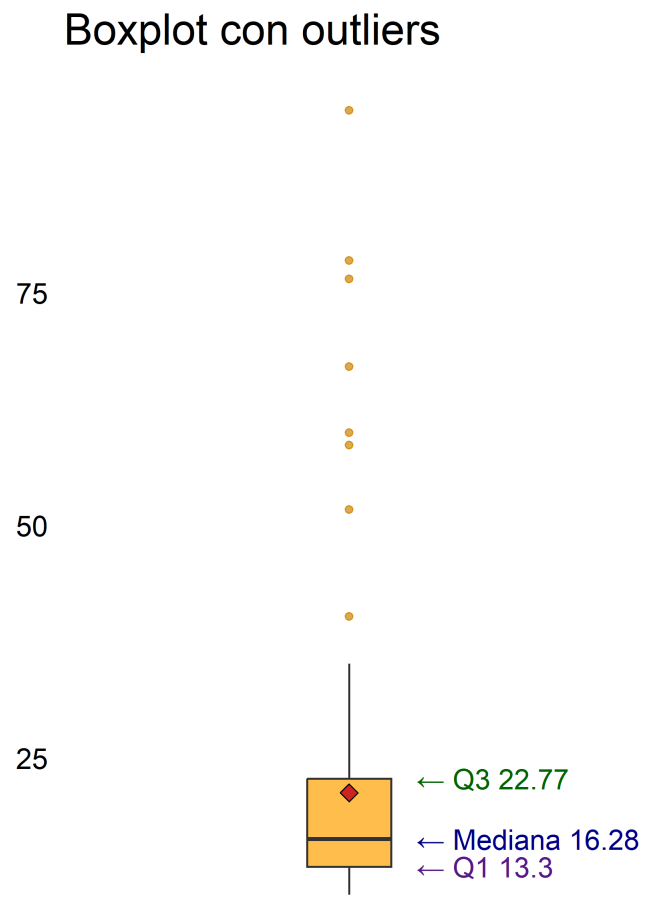
Valores atípicos (*outliers*)

- A veces, un **dato** difiere tanto del resto que parece no pertenecer al mismo conjunto de datos.
- Un *outlier* puede aparecer por **errores de registro**, fallos en el experimento u otras causas.
- Los **outliers** se pueden definir como **valores que exceden 1.5 veces el rango intercuartílico**.
- Para mostrarlos en un **boxplot**, se calculan valores llamados **"cercas" (fences)**:

$$\text{Cerca inferior} = Q1 - 1.5 \times IQR$$

$$\text{Cerca superior} = Q3 + 1.5 \times IQR$$

- Los **outliers** son los valores que caen **fuera de estas cercas**.



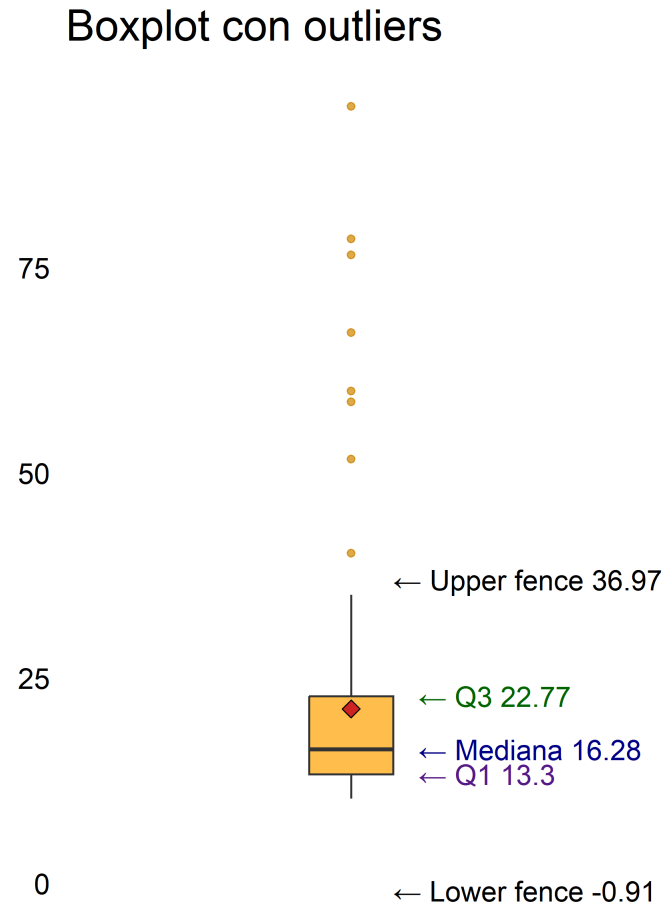
Valores atípicos (*outliers*)

```r
# Calcular estadísticas
Q1 <- quantile(df$valor, 0.25)
Q3 <- quantile(df$valor, 0.75)
IQR_val <- Q3 - Q1
lower_fence <- Q1 - 1.5 * IQR_val
upper_fence <- Q3 + 1.5 * IQR_val
lower_fence
```

```
##       25%
## -0.907378
```

```r
upper_fence
```

```
##      75%
## 36.96925
```



Contacto

Marta Coronado Zamora	David Castellano
marta.coronado@uab.cat	david.castellano@uab.cat
@geneticament.bsky.social	@castellanoed.bsky.social
Universitat Autònoma de Barcelona	University of Arizona