Estadística descriptiva

Técnicas estadísticas avanzadas para la conservación de la biodiversidad - Universidad de Huelva

---

David García Callejas

01/2021

Estadística descriptiva

- Poblaciones y muestras

Estadística descriptiva

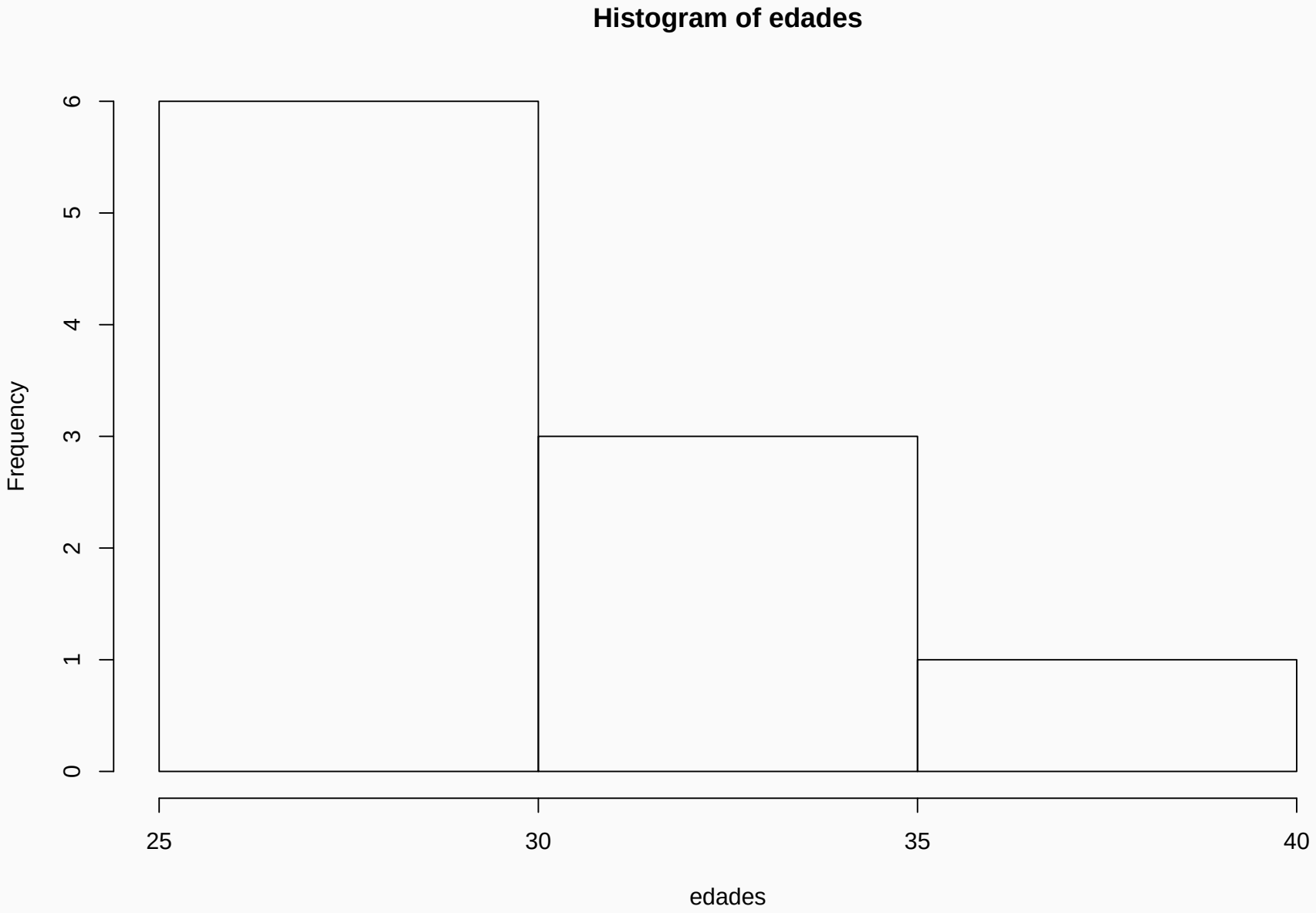
- Poblaciones y muestras
- Representaciones gráficas

Poblaciones y muestras

- Población: todos los alumnos del máster
- Muestra poblacional: 3 alumnos al azar de la población

```
edades <- sample(25:40,size = 10,replace = TRUE)
hist(edades)
```

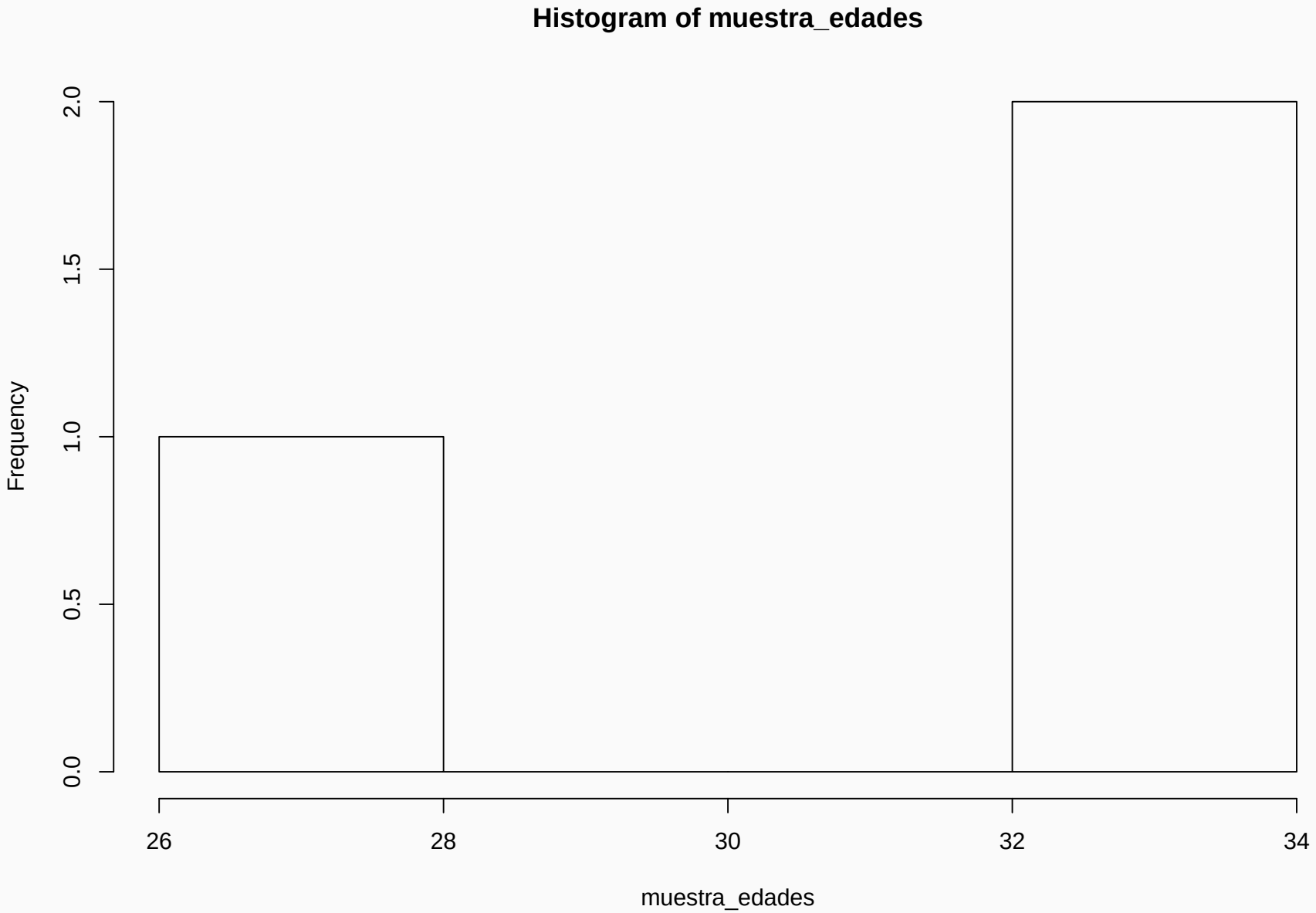
Poblaciones y muestras



Poblaciones y muestras

```
muestra_edades <- sample(edades,
                         size = 3,
                         replace = FALSE)
hist(muestra_edades)
```

Poblaciones y muestras



Poblaciones y muestras

Leer datos de una población

```
pob <- read.csv2(here::here("datasets",
                            "starwars_info_personajes.csv"))
alturas <- pob$height
```

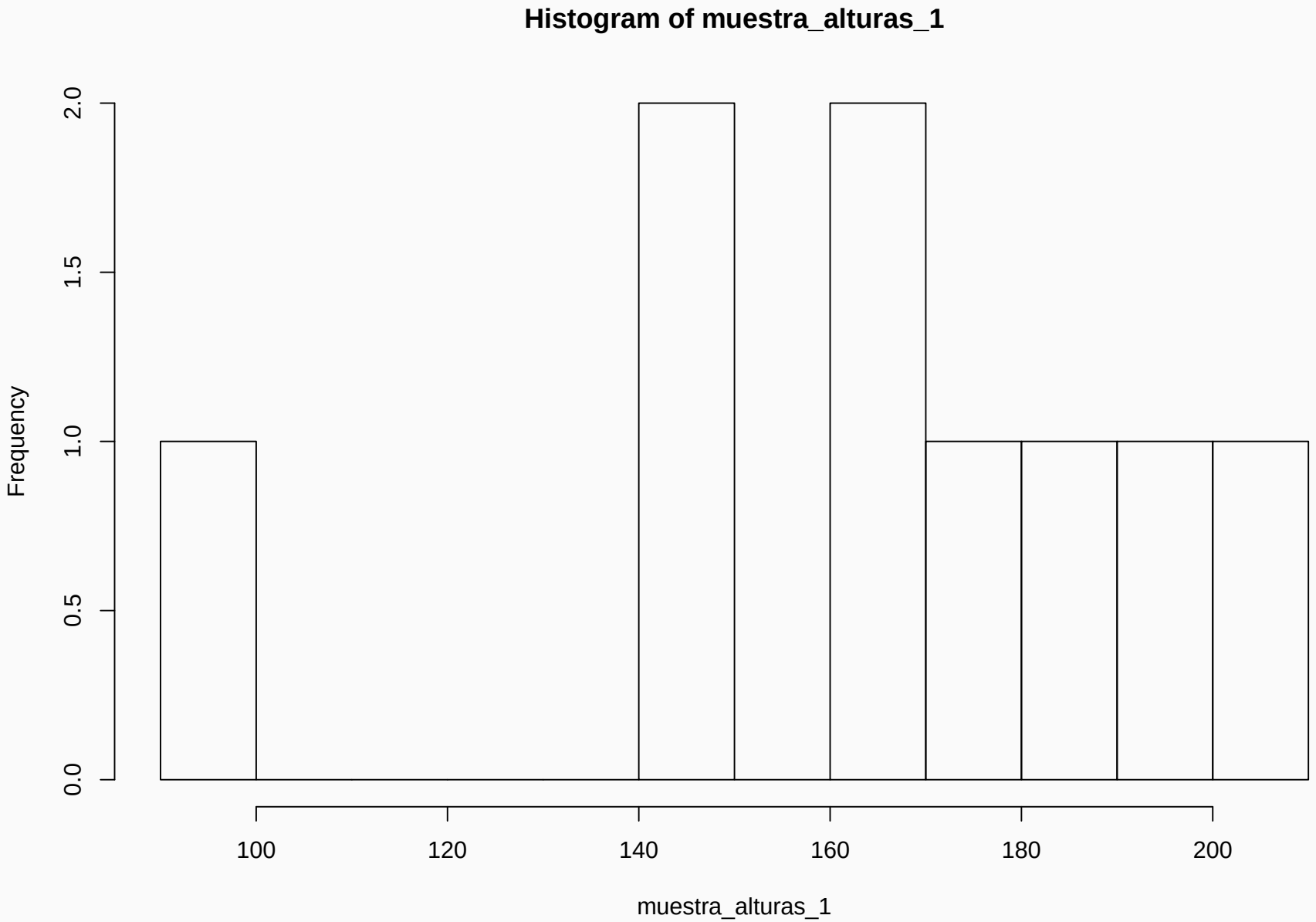
Obtener una muestra de esa población

```
muestra_alturas_1 <- sample(alturas,
                            size = 10,
                            replace = FALSE)
muestra_alturas_2 <- alturas[1:10]
```

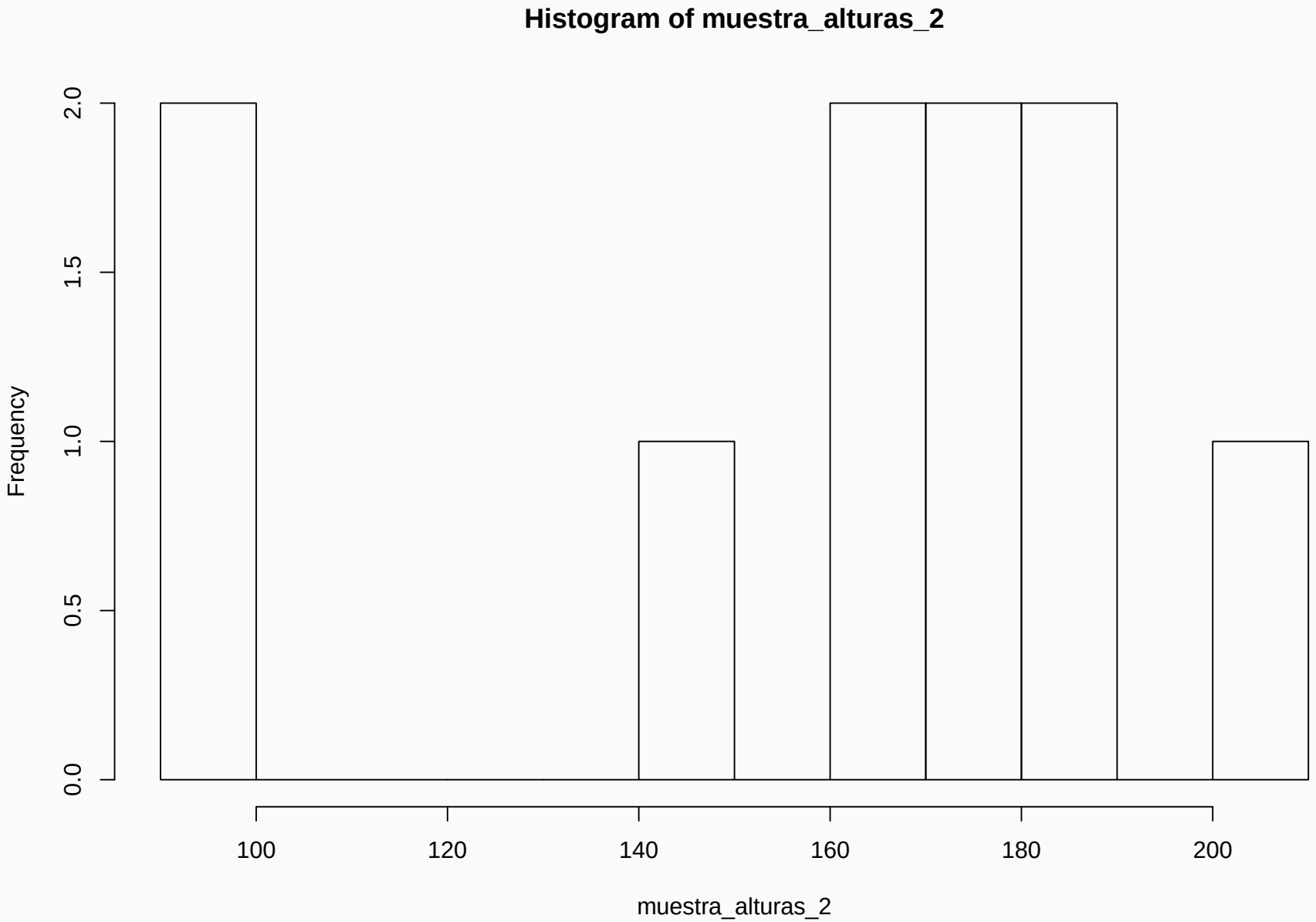
Estimar la representatividad de esa muestra

```
hist(muestra_alturas_1,breaks = 10)
hist(muestra_alturas_2,breaks = 10)
hist(alturas,breaks = 10)
```

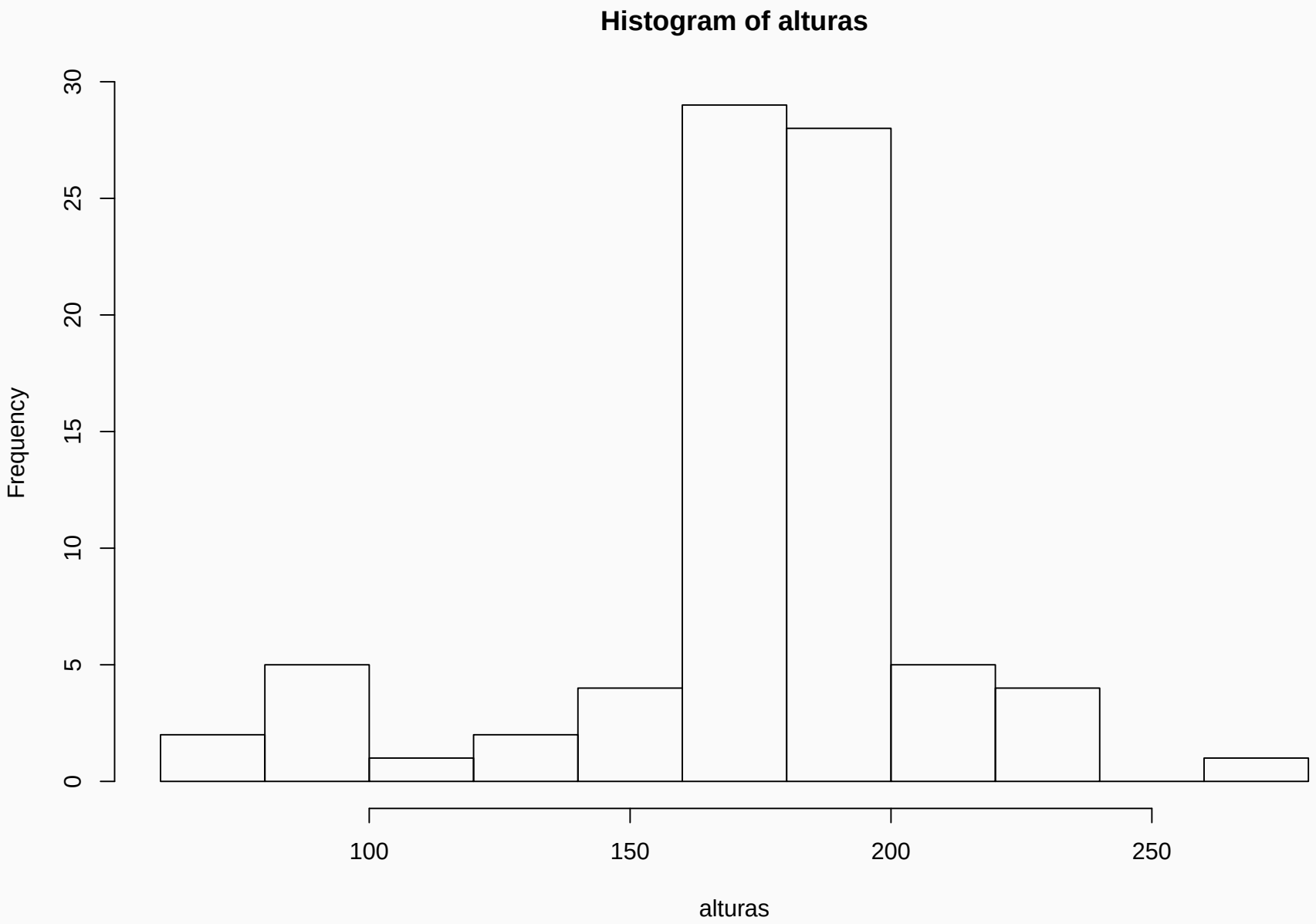
Poblaciones y muestras



Poblaciones y muestras



Poblaciones y muestras



Poblaciones y muestras

- Las propiedades de una medida en una población se describen con una serie de medidas:

Poblaciones y muestras

- Las propiedades de una medida en una población se describen con una serie de medidas:
  - de centralidad: media, mediana, moda

Poblaciones y muestras

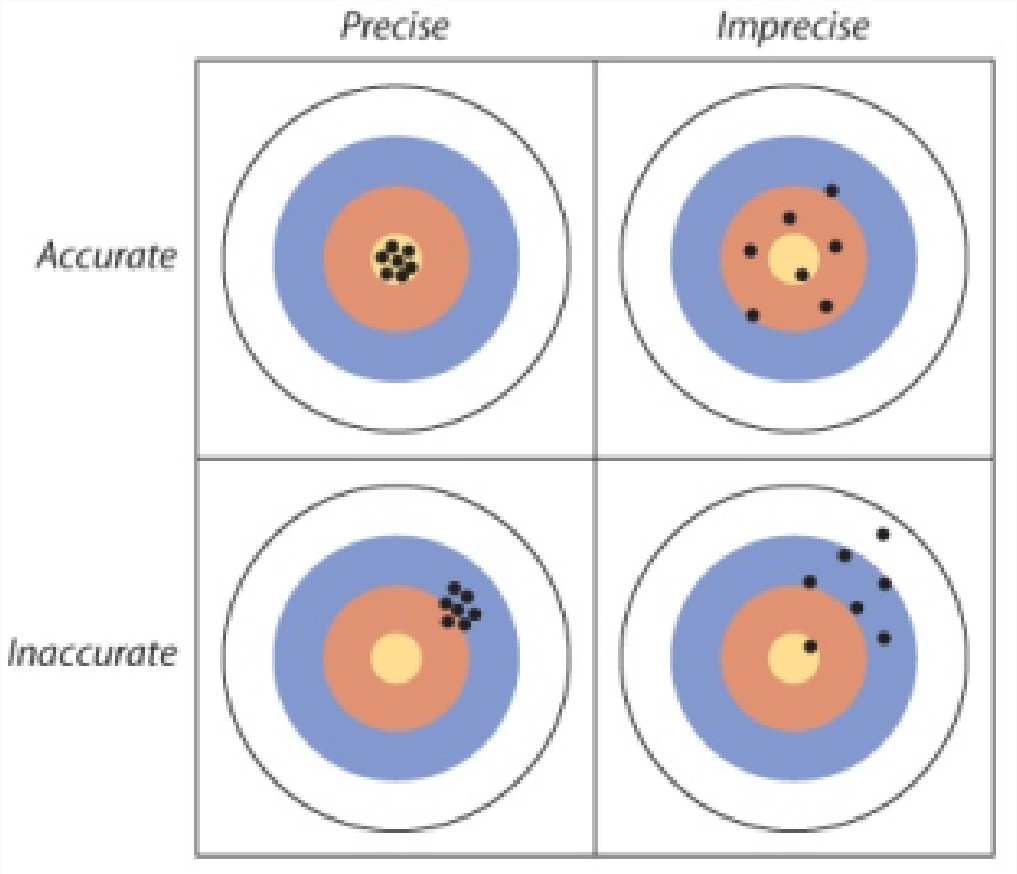
- Las propiedades de una medida en una población se describen con una serie de medidas:
  - de centralidad: media, mediana, moda
  - de dispersión: varianza, desviación típica

Poblaciones y muestras

- Las propiedades de una medida en una población se describen con una serie de medidas:
  - de centralidad: media, mediana, moda
  - de dispersión: varianza, desviación típica
- En estadística, aplicamos estas medidas a las muestras como estimaciones de la población total.

Poblaciones y muestras

Las medidas muestrales están influenciadas por el error de muestreo. Esto provoca errores de exactitud (sesgos o *bias*) y de precisión (*variance*):



Poblaciones y muestras

- Media de una población o muestra:

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} \tag{1}$$

Poblaciones y muestras

Ejercicio

Usando los datos *earthquakes.csv*, calcula:

1. La magnitud media de los terremotos incluidos.
2. La magnitud media de una muestra de 10 terremotos.
3. La diferencia entre la media poblacional y la media muestral.

Poblaciones y muestras

```
eq <- read.csv2(here::here("datasets",
                           "earthquakes.csv"))
pop.mean <- mean(eq$magnitude)
pop.mean
```

```
## [1] 4.978541
```

```
sample.eq <- sample(eq$magnitude,
                    size = 10,
                    replace = FALSE)
sample.mean <- mean(sample.eq)
sample.mean
```

```
## [1] 4.96
```

La diferencia entre la media poblacional y la media muestral es de 0.0185405

Poblaciones y muestras

- Mediana: el valor que deja a cada lado el 50% de los datos

```
median(eq$magnitude)
```

```
## [1] 4.8
```

- Moda: el valor más repetido

```
Mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}

Mode(eq$magnitude)
```

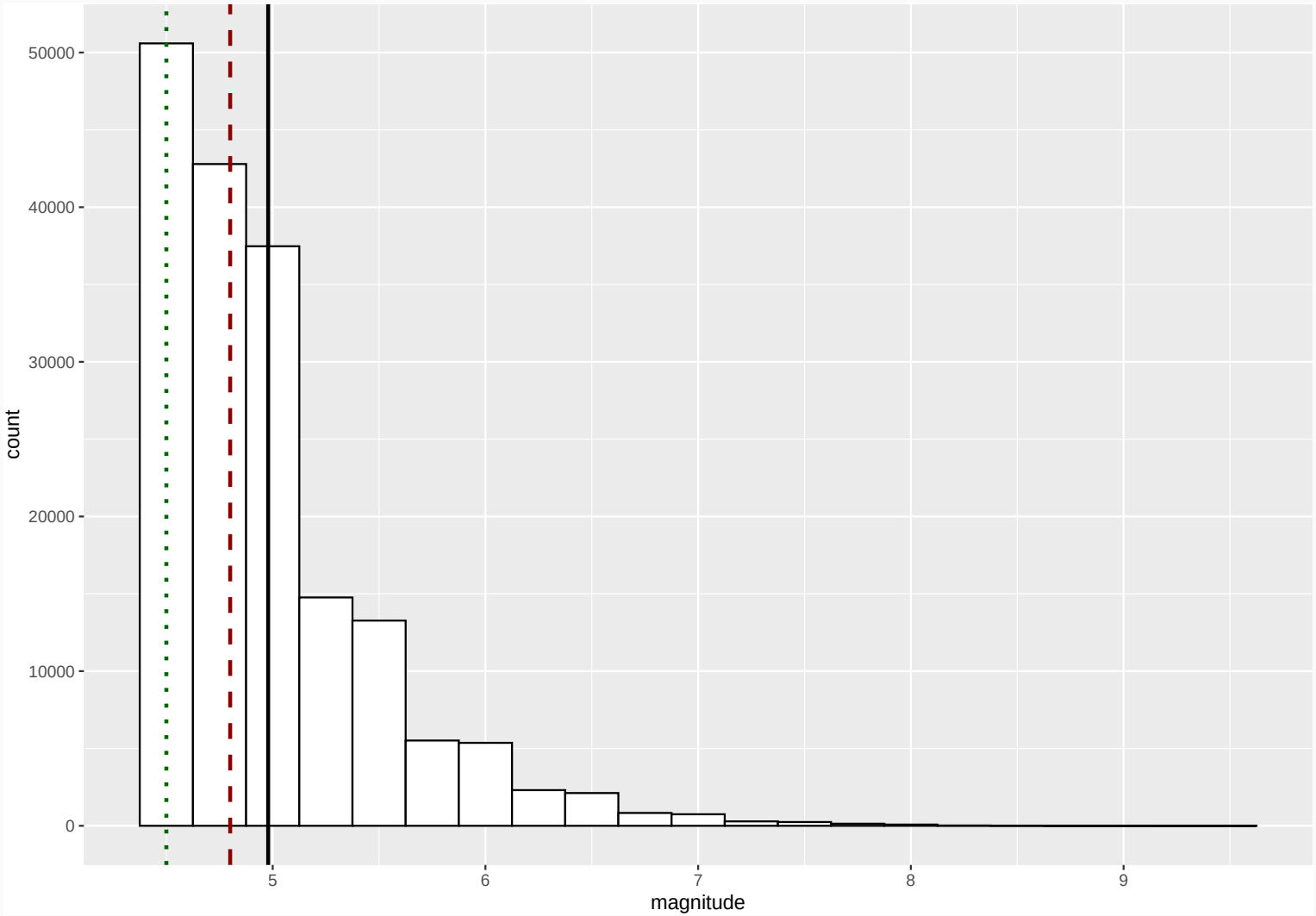
```
## [1] 4.5
```

Poblaciones y muestras

```
ggplot(eq, aes(x=magnitude)) +
    geom_histogram(binwidth=.25, colour="black", fill="white") +
    geom_vline(aes(xintercept=mean(magnitude, na.rm=T)),
               color="black", size=1) +
    geom_vline(aes(xintercept=median(magnitude, na.rm=T)),
               color="darkred", linetype="dashed", size=1) +
    geom_vline(aes(xintercept=Mode(magnitude)),
               color="darkgreen", linetype="dotted", size=1)
```

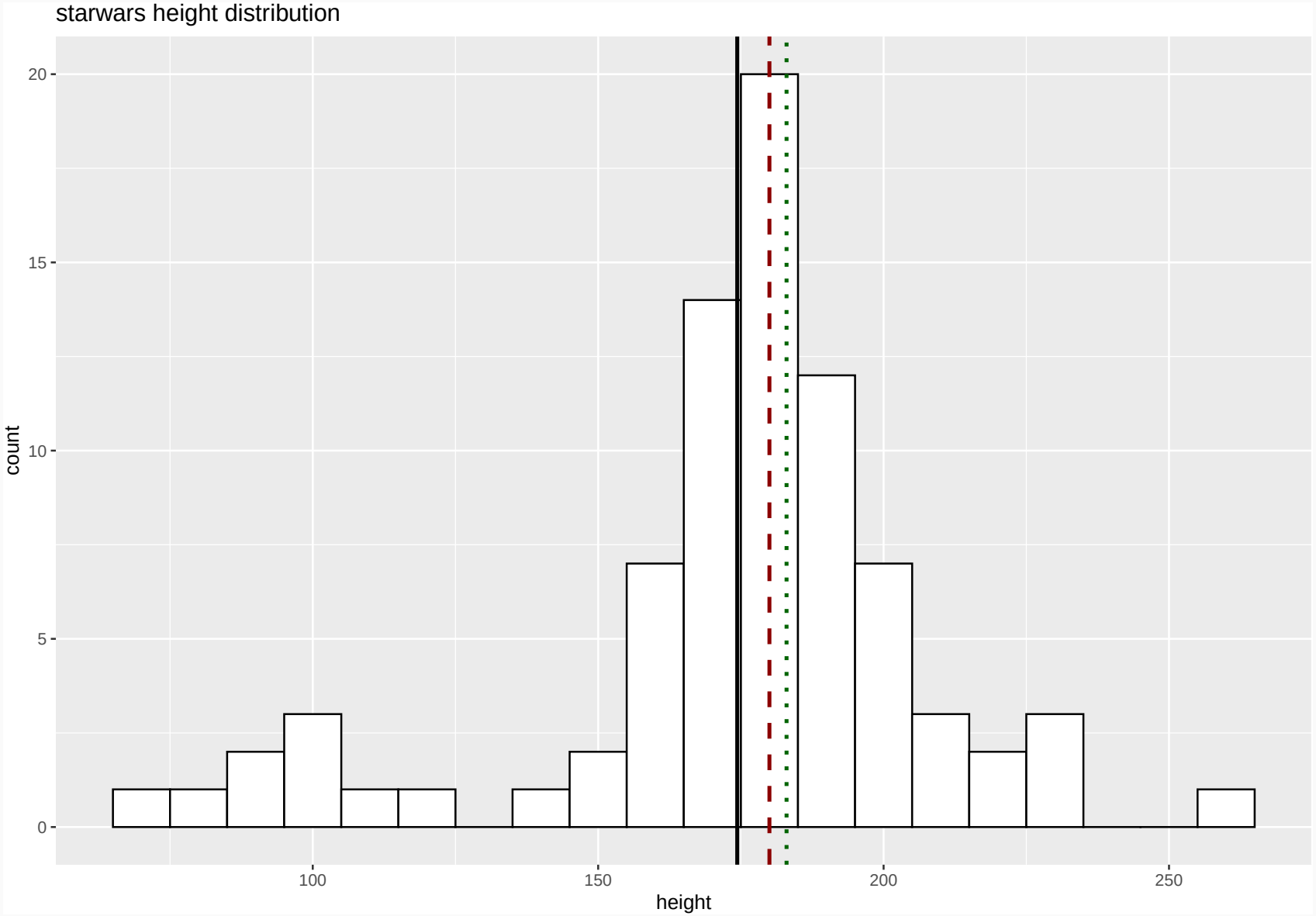
Poblaciones y muestras

- Media: negro
- Mediana: rojo
- Moda: verde



Poblaciones y muestras

¿Qué tal se ven otro tipo de datos?



Poblaciones y muestras

- Medidas de dispersión
  - Valores mínimos, máximos, cuantiles

```
summary(pob$height)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    66.0   167.0   180.0   174.4   191.0   264.0       6
```

Poblaciones y muestras

- Medidas de dispersión
  - Varianza y desviación típica

$$SD = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)}{N-1}} \tag{2}$$

```
sd(pob$height,na.rm = TRUE)
```

```
## [1] 34.77043
```

Poblaciones y muestras

- Medidas de dispersión
  - error estándar asociado a la media

Poblaciones y muestras

https://gallery.shinyapps.io/sampling_and_stderr/

Medidas muestrales

- tendencias centrales: media, mediana, moda (R)

Medidas muestrales

- tendencias centrales: media, mediana, moda (R)
- dispersión: min/max, cuantiles, desviación típica, error estándar, coeficiente de variación, intervalos de confianza (R)

Medidas muestrales

- tendencias centrales: media, mediana, moda (R)
- dispersión: min/max, cuantiles, desviación típica, error estándar, coeficiente de variación, intervalos de confianza (R)
- desviación típica y error estándar (R)

Medidas muestrales

- tendencias centrales: media, mediana, moda (R)
- dispersión: min/max, cuantiles, desviación típica, error estándar, coeficiente de variación, intervalos de confianza (R)
- desviación típica y error estándar (R)
- intervalos de confianza (R)

Medidas muestrales

- tendencias centrales: media, mediana, moda (R)
- dispersión: min/max, cuantiles, desviación típica, error estándar, coeficiente de variación, intervalos de confianza (R)
- desviación típica y error estándar (R)
- intervalos de confianza (R)
- *datos: altura alumnos, earthquakes, starwars height/mass*

Representaciones gráficas

- histogramas/distribuciones (distribucion de frecuencias)

Representaciones gráficas

- histogramas/distribuciones (distribucion de frecuencias)
- boxplots

Representaciones gráficas

- histogramas/distribuciones (distribucion de frecuencias)
- boxplots
- correlaciones según tipo de variables (WS)

Representaciones gráficas

- histogramas/distribuciones (distribucion de frecuencias)
- boxplots
- correlaciones según tipo de variables (WS)
- *datos: altura alumnos, earthquakes, starwars height/mass, happiness-sunshine*