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

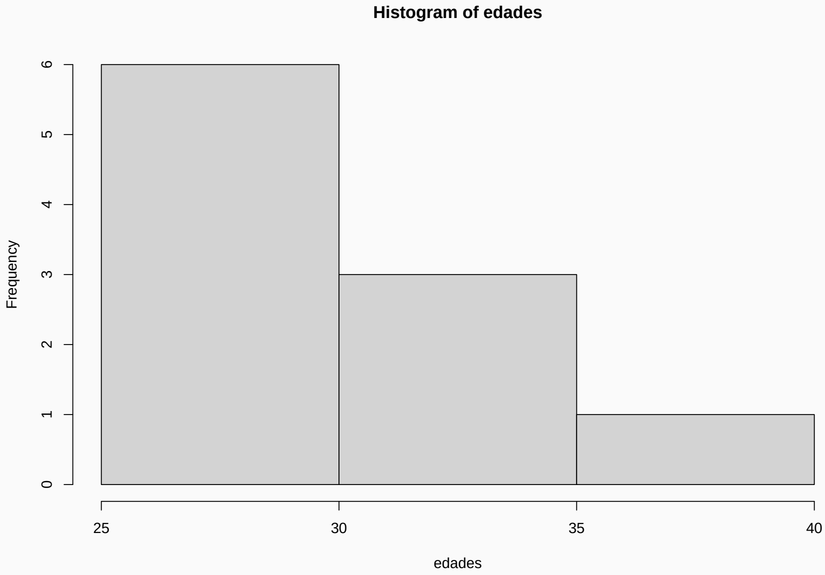
- Poblaciones y muestras
- Representaciones gráficas

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)
```

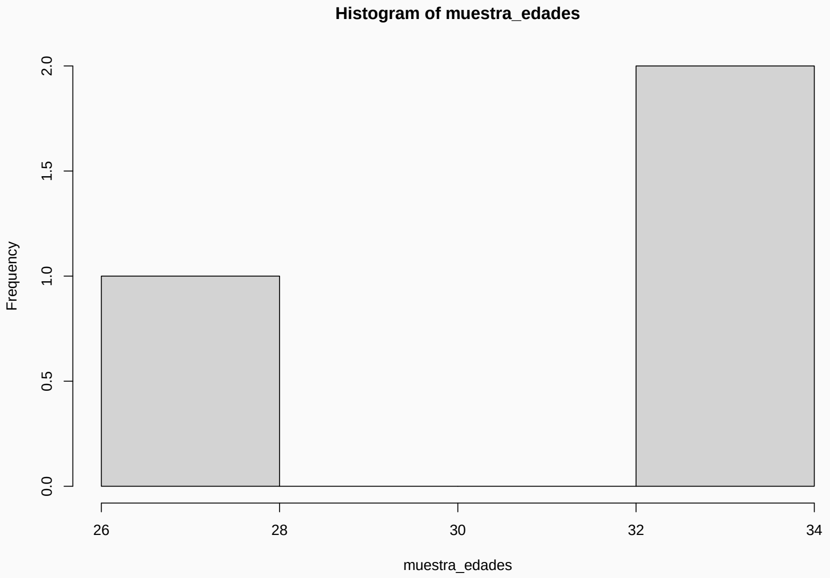
Poblaciones y muestras



Poblaciones y muestras

```r
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
```

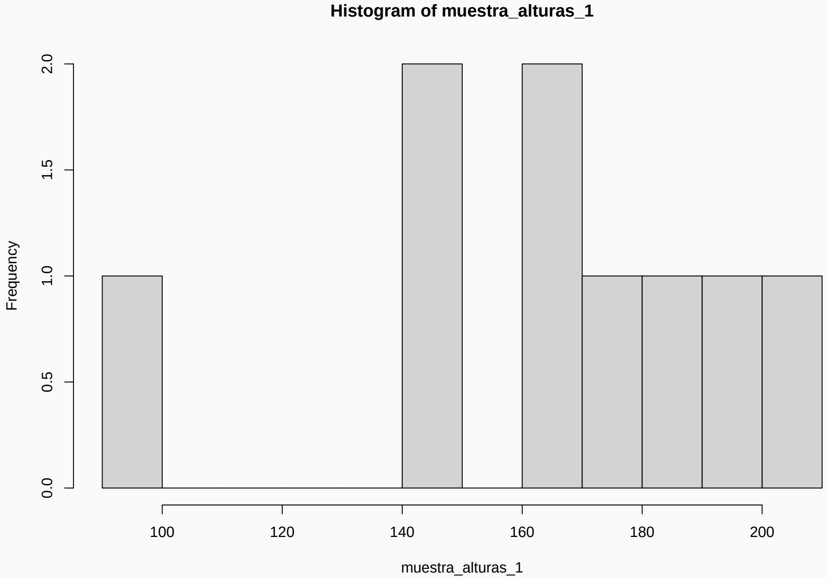
Obtener una muestra de esa población

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

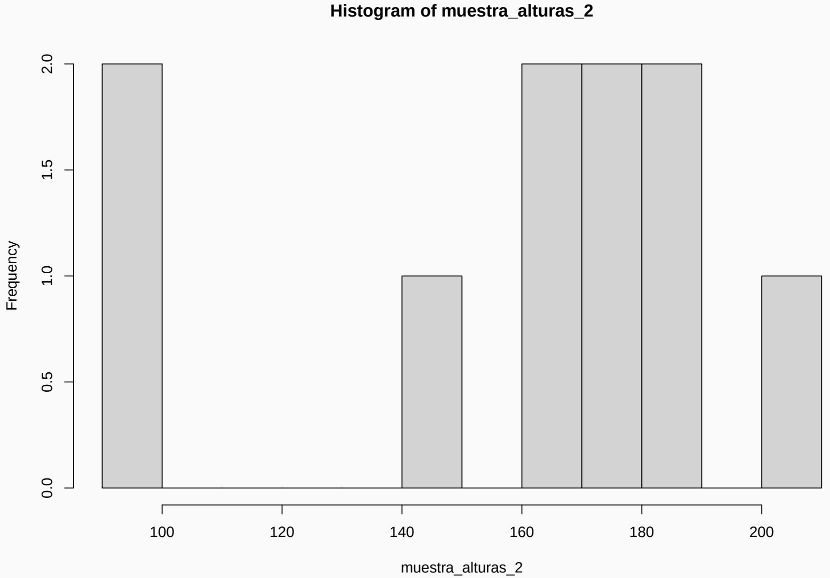
Estimar la representatividad de esa muestra

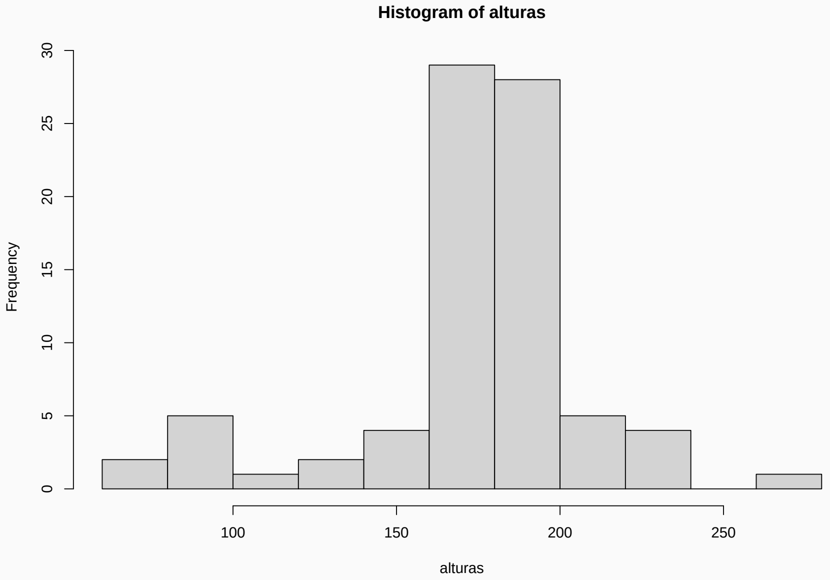
```
hist(muestra_alturas_1,breaks = 10)
hist(muestra_alturas_2,breaks = 10)
hist(alturas,breaks = 10)
```

Poblaciones y muestras



Poblaciones y muestras




Histogram of alturas
Frequency
0
5
10
15
20
25
30
100
150
200
250
alturas

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

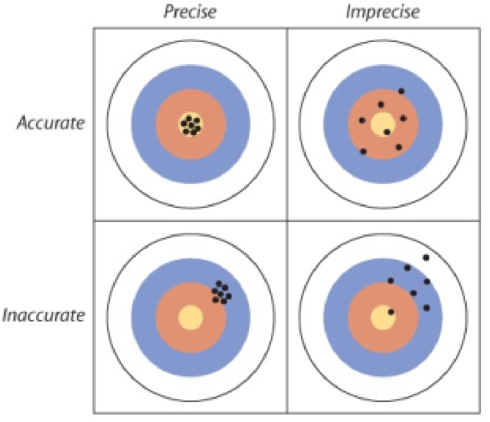
- Las propiedades de una medida en una población se describen con una serie de medidas:
  - de centralidad: media, mediana, moda
  - de dispersión: varianza, desviación típica

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:

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

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

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

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

```r
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)
```

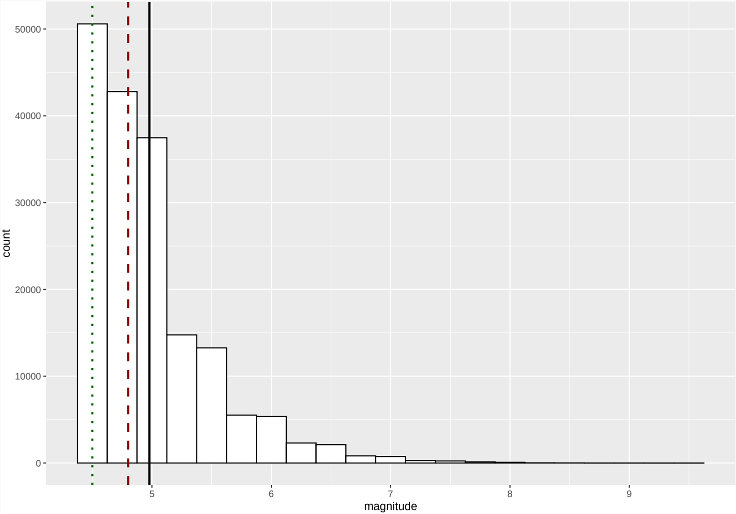
```
## [1] 4.5
```

Poblaciones y muestras

```r
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)
```

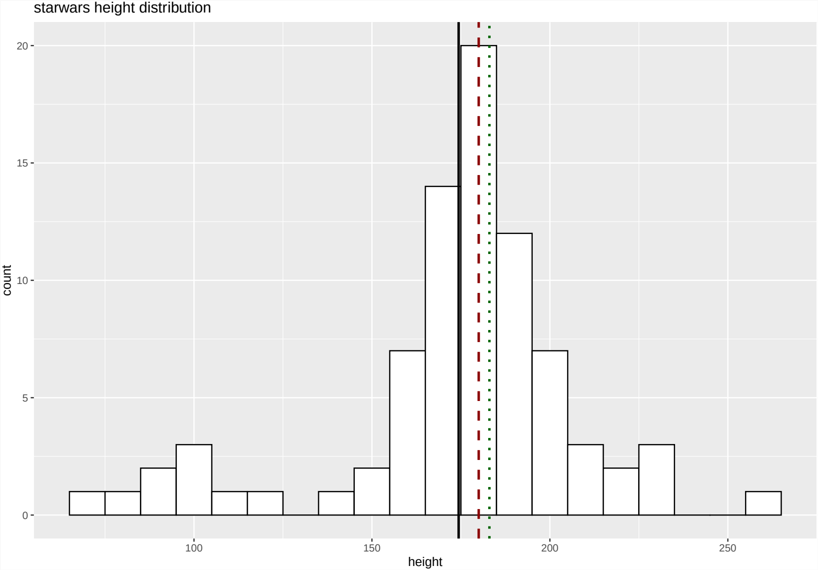
Poblaciones y muestras

- 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
  - Desviación típica

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

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

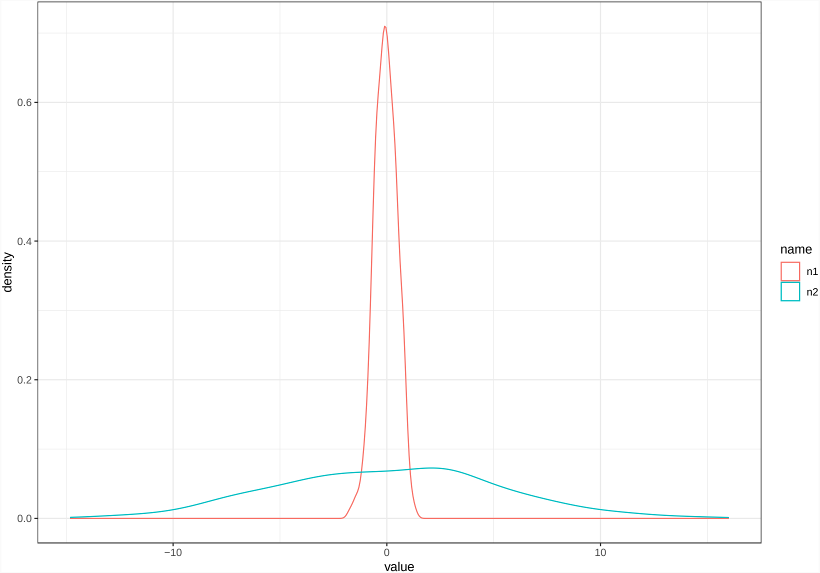
```
## [1] 34.77043
```

Poblaciones y muestras

```r
n1 <- rnorm(500,0,0.5)
n2 <- rnorm(500,0,5)

compara <- data.frame(n1 = n1, n2 = n2)
compara.long <- pivot_longer(compara,cols = n1:n2)
compara.plot <- ggplot(compara.long,aes(x = value, color = name)) +
  geom_density() +
  theme_bw()
```

Poblaciones y muestras



Poblaciones y muestras

- Medidas de dispersión
  - error estándar asociado a la media:

  Mientras que la desviación típica cuantifica la dispersión de una población, el error estándar mide la incertidumbre de la media asociada a una muestra:

$$SE = \frac{\sigma}{\sqrt{N}} \tag{3}$$

Poblaciones y muestras

Población:

```
head(alturas)
```

```
## [1] 172 167  96 202 150 178
```

Para una muestra determinada, el error estándar será más alto cuanto menor sea el tamaño muestral

```
muestra_peq <- sample(alturas, size = 5)
sd(muestra_peq,na.rm = TRUE)/sqrt(5)
```

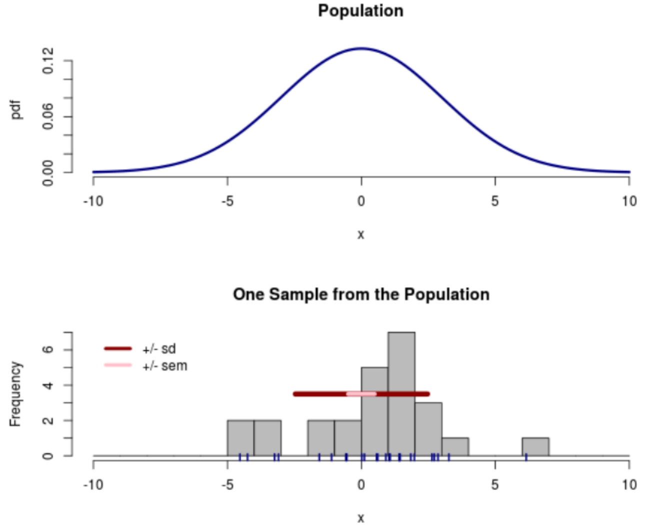
```
## [1] 7.499333
```

```
muestra_gran <- sample(alturas, size = 50)
sd(muestra_gran,na.rm = TRUE)/sqrt(50)
```

```
## [1] 4.998448
```

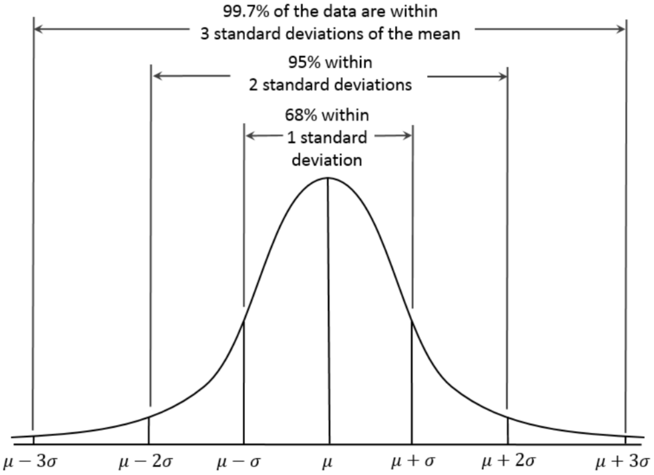
Poblaciones y muestras

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



Poblaciones y muestras

En una distribución normal:



Poblaciones y muestras

- Medidas de dispersión
    - intervalos de confianza:

Dan una estima del rango de valores plausibles para una medida muestral, generalmente la media. Vienen asociados a un valor de "confianza", que suele ser el 95% (por razones históricas). La interpretación de un intervalo de confianza es poco intuitiva:

El 95% de intervalos de confianza calculados a partir de muestras de la población contendrán el valor real de la media

- Esto NO quiere decir que la media real de la población esté con un 95% de probabilidad en un intervalo determinado.

Poblaciones y muestras

https://rpsychologist.com/d3/ci/

Poblaciones y muestras

¿Cómo se calcula un intervalo de confianza para una media muestral?

1. Número de observaciones n
2. Media muestral $\bar{x}$
3. Desviación típica de la muestra σ
4. Nivel de confianza (p.ej. 95%)

Con estos ingredientes, podemos calcular el intervalo alrededor de la media:

$$CI = \bar{x} \pm Z \frac{\sigma}{\sqrt{n}} \tag{4}$$

Poblaciones y muestras

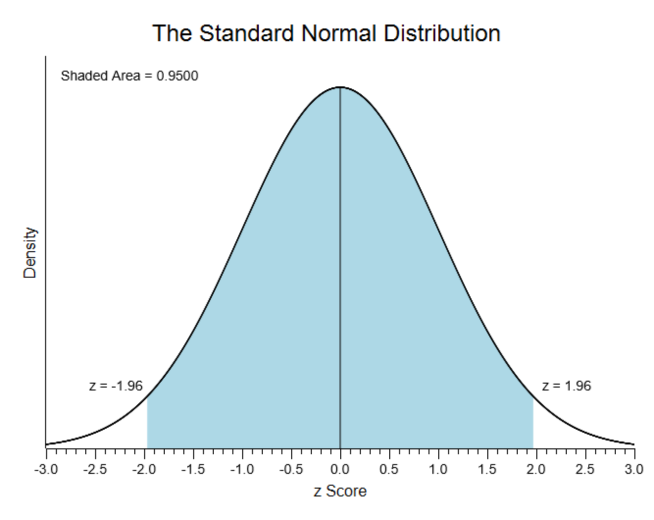
Ejemplo:

$$n = 50$$
$$\bar{x} = 4.3$$
$$\sigma = 0.6$$
$$CI = 4.3 \pm 1.96 \frac{0.6}{\sqrt{50}} = [4.133, 4.466]$$

Todo bien, pero... ¿de dónde sale ese 1.96?

Poblaciones y muestras

Para una distribución normal estándar ($\mu = 0, \sigma = 1$), el 95% de los datos está comprendido entre -1.96 y 1.96.



El valor de Z para cualquier porcentaje se puede consultar en tablas estándar (e.g. aquí). Hoy día, afortunadamente, no es necesario calcular intervalos de confianza a mano.

Poblaciones y muestras

Calcular intervalos de confianza de la media de una muestra en R:

```r
t.test(alturas,conf.level = .95)
```

```
##
##  One Sample t-test
##
## data:  alturas
## t = 45.131, df = 80, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  166.6697 182.0464
## sample estimates:
## mean of x
##   174.358
```

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*