A2: Análisis Estadístico I

Estadística Avanzada, Universitat Oberta de Catalunya

Paula Muñoz Lago

19 noviembre 2019

Contents

1. Analítica descriptiva

1.1 Lectura del fichero

Leer el fichero ESTRADL_clean.csv. Validar que los datos leídos son correctos. Si no es así, realizar las conversiones oportunas.

En primer lugar, realizamos la carga de datos.

```
current_working_directory <- getwd()
data <- read.csv(paste(current_working_directory,"/ESTRADL_clean.csv", sep = ""))
```

Para comprobar que los datos leídos son correctos, imprimiremos sus clases. Veremos que la edad de la primera menarquía es de tipo numérico, mientras que una edad es más correcto que sea de tipo entero, por lo que procedemos a la conversión del tipo.

```
sapply(data, class)
```

```
##        Id   Estradl    Ethnic    Entage  Numchild    Agefbo   Anykids
## "integer" "numeric"  "factor" "integer" "integer" "integer"  "factor"
##  Agemenar       BMI       WHR      Area
## "numeric" "numeric" "numeric"  "factor"
```

```
data$Agemenar <- as.integer(data$Agemenar)
```

1.2 Análisis descriptivo visual

Representar de forma visual las variables del conjunto de datos y las distribuciones de sus valores. Escoged la representación más apropiada en cada caso.

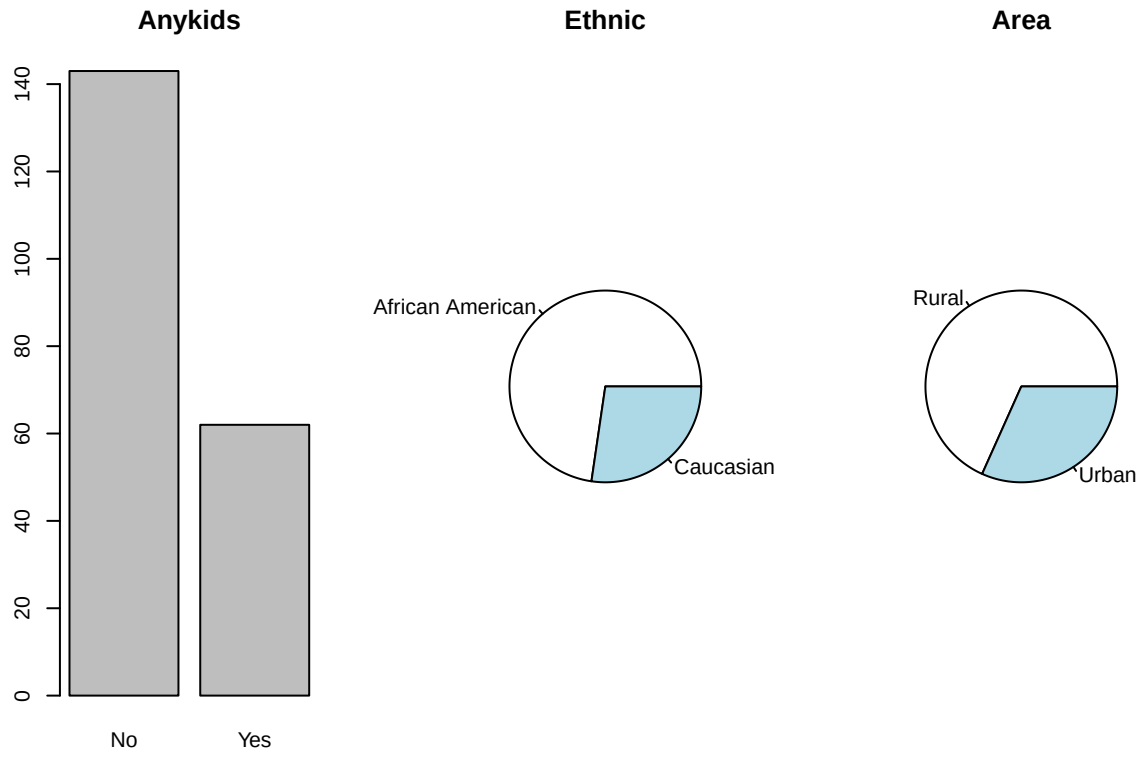
- **Anykids, Ethnic y Area**

```
par( mfrow=c(1,3))

barplot(table(data$Anykids), main = "Anykids")

pie(table(data$Ethnic), main = "Ethnic")

pie(table(data$Area), main="Area")
```




```
summary(data$Anykids)
```

```
##  No Yes
```

```
## 143  62
```

```
summary(data$Ethnic)
```

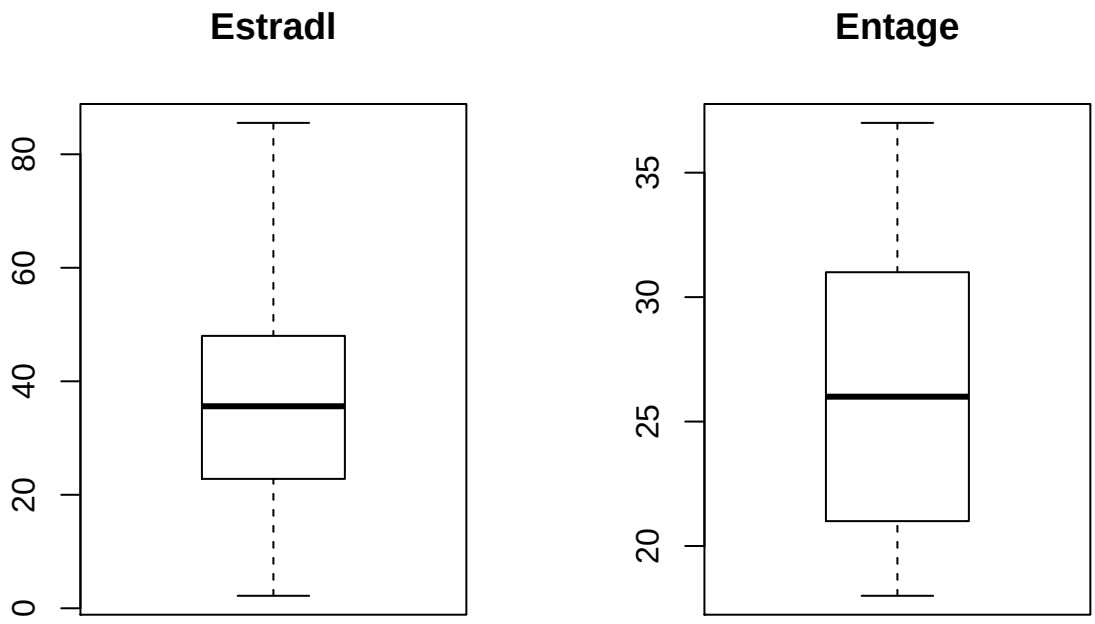
```
## African American        Caucasian
##              149               56
```

```
summary(data$Area)
```

```
## Rural Urban
##   140    65
```

- **Estradiol y Entage**

```
par( mfrow=c(1,2))
boxplot(data$Estradl, main = "Estradl")
boxplot(data$Entage, main = "Entage")
```




```
summary(data$Estradl)
```

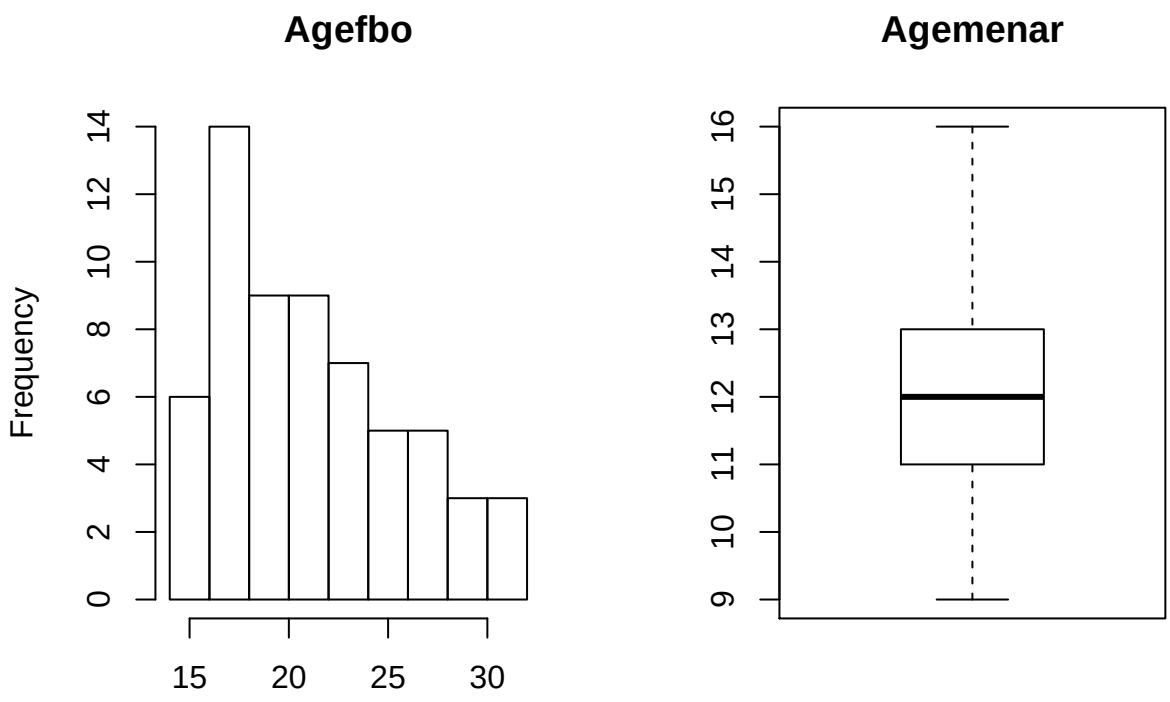
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.20   22.80   35.60   36.26   48.00   85.53
```

```
summary(data$Entage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   21.00   26.00   26.05   31.00   37.00
```

- **Agefbo y Agemenar**

```r
par( mfrow=c(1,2))
hist(data$Agefbo[which(data$Agefbo > 0)], main="Agefbo")
boxplot(data$Agemenar, main = "Agemenar")
```




```r
summary(data$Agefbo)
```

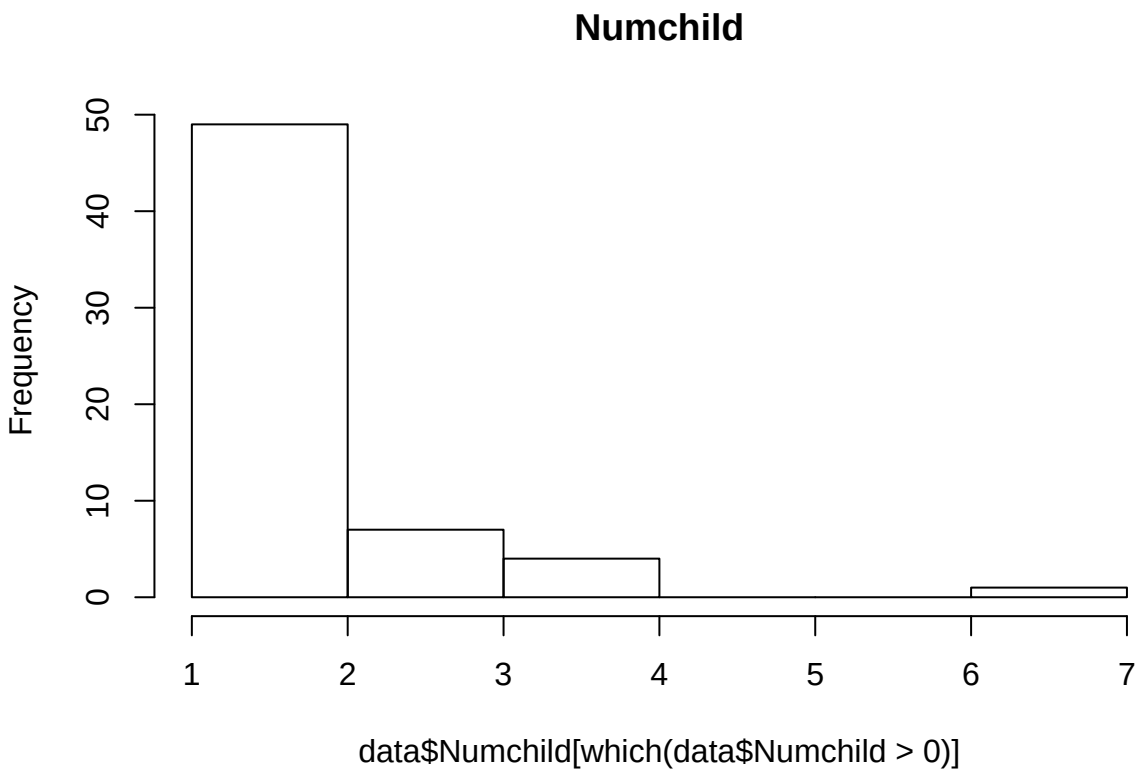
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   6.444  17.000  32.000
```

```r
summary(data$Agemenar)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00   11.00   12.00   12.31   13.00   16.00
```

- **Numchild**

```r
hist(data$Numchild[which(data$Numchild > 0)], main="Numchild")
```

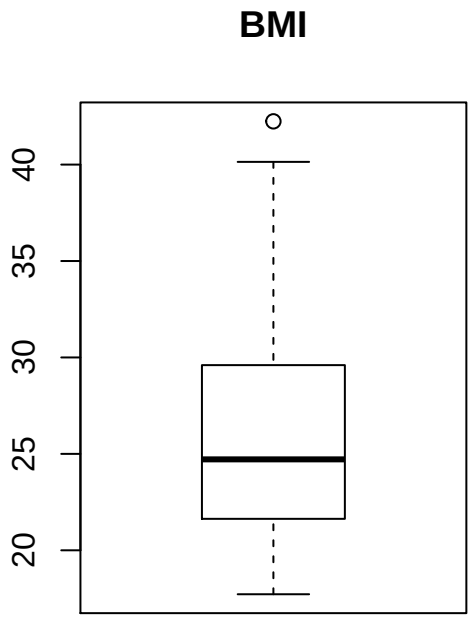


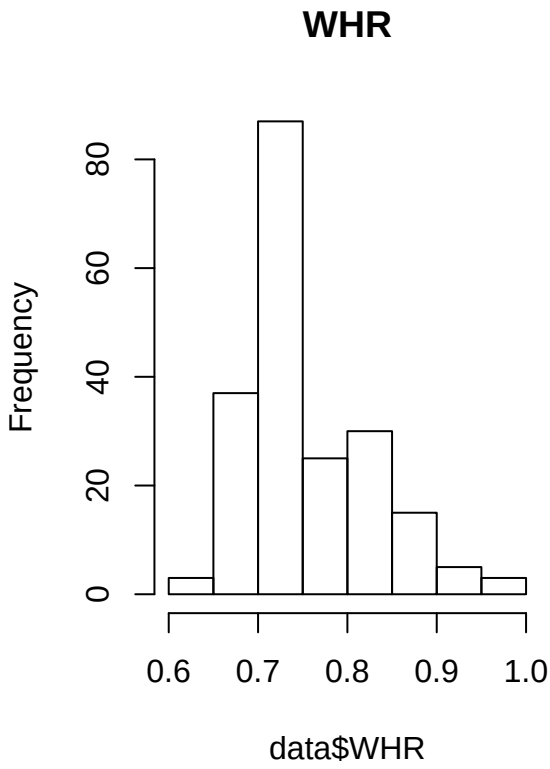

```
summary(data$Numchild)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.5268  1.0000  7.0000
```

- **BMI y WHR**

```
par( mfrow=c(1,2))
boxplot(data$BMI, main = "BMI")
hist(data$WHR, main = "WHR")
```






```
summary(data$BMI)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.72   21.63   24.72   25.91   29.60   42.24
```

```
summary(data$WHR)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.6200  0.7100  0.7400  0.7596  0.8100  0.9800
```

2. Edad de menarquía media de la población

A partir de los datos de la muestra, se desea estimar la edad de menarquia media de las mujeres. En concreto, se desea estudiar si la edad de menarquia media de la población es de 14 años, o bien es inferior a 14 años. Para ello, realizar un contraste estadístico con un nivel de confianza del 98 %.

2.1 Hipótesis nula y alternativa

La hipótesis nula es de la que partimos, la indicada en el enunciado, mientras que la hipótesis alternativa tiene que representar un caso diferente, siendo así una alternativa unilateral.

$$\begin{cases} H_0: & \mu = 14 \\ H_1: & \mu < 14 \end{cases}$$

2.2 Método

Una vez planteadas las hipótesis, debemos tomar una decisión. Aceptar o rechazar H_0.

El nivel de significación en este caso será $\alpha = 0.02$. Es decir, podremos rechazar la hipótesis nula de forma equivocada 2 de cada 100 veces. Se trata de una distribución normal $N(\mu,\sigma^2)$ y puesto que no disponemos de información sobre la varianza σ, utilizaremos una distribución t-Student para aproximar una variable S a σ.

2.3 Cálculos

T de Student

Para decidir si rechazamos la hipótesis nula o no, calcularemos el estadístico de contraste con la siguiente formula.

$$t = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$

Seguiremos la ley de t de Student con n-1 grados de libertad, dado que no conocemos la varianza (σ). Puesto que N>30, es decir, consideramos que el tamaño de la muestra es grande, podremos aproximarlo a una distribución normal.

En nuestro caso, $\mu = 14$, n = 205 y $\bar{X} = 12.3122$, y, al desconocer σ, debemos calcular S (desviación típica muestral).

- **S**

$$S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = 1.3756$$

El calculo se ha realizado en R con el siguiente código:

```
# (xi-xmean)^2
sum = 0
m = mean(data$Agemenar)
for(a in data$Agemenar){
  x = (a - m)^2
  sum = sum + x
}
#sqrt
S = sqrt(sum / (length(data$Agemenar) - 1))
S
```

```
## [1] 1.375592
```

- **t**

Una vez obtenida la variable S, procedemos a obtener t, con el siguiente código.

```
t = (m - 14) / (S / sqrt(length(data$Agemenar)))
t
```

```
## [1] -17.56749
```

De esta forma, obtenemos que t = -17.56 con 204 grados de libertad

Valor crítico

El valor crítico, dado el nivel de confianza del 98%, al ser un estudio unilateral, obtenemos que es:

$$\mu <= \bar{X} + (t_{0.02} * \frac{S}{\sqrt{N}})$$

$$|t_{0.02}| = 2.066964$$

```
talpha = qt(p=0.02, df = 204)
margen_error = (talpha * S)/sqrt(205)
limite_superior = m + margen_error
limite_inferior = m - margen_error
```

El intervalo de confianza obtenido a raíz del valor crítico es $(12.1 < \mu < 12.5)$.

P-value

$$P(|t_{n-1}| < t) = P(|t_{204}| < -17.56) = 1.093513 * 10^{-42} \simeq 0$$

```
p_value = 2*pt(t, df = length(data$Agemenar) - 1)
p_value
```

```
## [1] 1.093513e-42
```

2.4 Interpretación

Rechazaremos la hipótesis nula, dado que el p-valor $< \alpha$ y $14 \not\subset 12.11, 12.51$.

3. Intervalo de confianza de Estradiol

3.1 Calcular el intervao de confianza del 95% de la variable Estradiol

Dado que se trata de una muestra normal, el intervalo de confianza es aquél que se encuentra entre unos márgenes en la campana de la normal. Siendo la confianza del 95%, la parte de la campana que queda fuera del intervalo de confianza sería un 5% que se distribuye en dos partes, de 2.5% cada una.

Para comenzar con los cálculos, necesitamos los siguientes datos:

$\bar{X}$ (Media del valor del estradiol en la muestra) = 36.26

```
X = mean(data$Estradl)
X
```

```
## [1] 36.26298
```

S (desviación típica para una distribución de Student) = $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x_i})^2} = 17.46537$

Dado que en este ejercicio necesitaremos realizar el cálculo de S varias veces, crearemos una función.

```
compute_S <- function(vector){
  # S
  # (xi-xmean)^2
  sum = 0
  m = mean(vector)
```

```
  for(a in vector){
    x = (a - m)^2
    sum = sum + x
  }
  #1/n - 1
  y = 1 / (length(vector) - 1)
  #sqrt
  S = sqrt(sum / (length(vector) - 1))
}
```

```
S <- compute_S(data$Estradl)
S
```

```
## [1] 17.46537
```

$\alpha = 0.05$ (indicado en el enunciado)

Para conocer el intervalo de confianza, deberemos realizar la siguiente operación:

$$P(-|t_{n-1}| <= \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} <= |t_{n-1}|) = 0.95$$

```
tn1 = qt(p=(1 - 0.95), df = 204)
tn1
```

```
## [1] -1.652357
```

$$P(-1.97 <= \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} <= 1.97) = 0.95$$

$$P(\bar{X} - 1.97\frac{S}{\sqrt{n}} <= \mu <= \bar{X} + 1.97\frac{S}{\sqrt{n}}) = 0.95$$

```
margen_error = abs((tn1 * S)/sqrt(205))
limite_superior = X + margen_error
limite_inferior = X - margen_error
limite_inferior
```

```
## [1] 34.24737
```

```
limite_superior
```

```
## [1] 38.27858
```

Por tanto, obtenemos como intervalo de confianza: (34.24, 38.27).

3.2 Interpretar el resultado

A partir del resultado obtenido del intervalo de confianza, explicar la interpretación del mismo en cuanto al valor de estradiol en las mujeres.

En nuestra muestra, la media de los niveles de estradiol en mujeres abarca desde 33.85 hasta 38.66 con una confianza del 95%.

3.3 Comparar intervalos

Si calculáramos el intervalo de confianza del 97 %, ¿cómo sería el intervalo de confianza en relación al calculado previamente? Justificar la respuesta. No es necesario realizar los cálculos.

Dado que en este caso el nivel de confianza es superior, el porcentaje de rechazo, α, es inferior. $\alpha = 0.03$, por lo que la región de aceptación será mayor y los valores críticos distarán más entre ellos. El límite inferior será levemente menor que 33.85, como en el caso anterior, y el límite superior el valor crítico será ligeramente superior a 38.66.

4. Diferencias en el nivel de estradiol según etnia

4.1 Hipótesis nula y alternativa

$$\begin{cases} H_0: & \mu_0 = \mu_1 \\ H_1: & \mu_0 \neq \mu_1 \end{cases}$$

Siendo μ_0 la media del nivel de estradiol en mujeres caucásicas y μ_1 en mujeres negras. Proponemos como H_0 que las medias del nivel de estradiol en mujeres caucásicas y negras es la misma, y como hipótesis alternativa, que son distintas.

4.2 Método

En función de las características de la muestra, decidir qué método aplicar para validar la hipótesis planteada. Para ello, debéis especificar como mínimo: a) si es un contraste de una muestra o de dos muestras (en caso de dos muestras, si éstas son independientes o están relacionadas), b) si podéis asumir normalidad y por qué, c) si el test es paramétrico o no paramétrico, d) si el test es bilateral o unilateral.

Se trata de un contraste bilateral entre dos muestras independientes. Estas muestras tienen una distribución normal, ya que los datos pueden organizarse en forma de campana gaussiana y podremos calcular la probabilidad de que varios valores ocurran en un cierto intervalo dada una confianza, por lo que también podemos afirmar que el que vamos a realizar se trata de un test paramétrico. Puesto que no conocemos la distribución, utilizaremos la t-student.

4.3 Cálculos

Realizar los cálculos para validar o rechazar la hipótesis de la investigación, con un nivel de confianza del 95 %. Calcular: el estadístico de contraste, el valor crítico y el valor p.

Dado que no conocemos la desviación típica, la calcularemos con la siguiente fórmula:

$$S = \sqrt{\frac{\sum_{i=1}^{n_1}(x_{i1} - \bar{x_1})^2 + \sum_{i=1}^{n_2}(x_{i2} - \bar{x_2})^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Para ello, necesitamos calcular los siguientes datos

- $\mathbf{X_1}$ **y** $\mathbf{X_2}$

```
X1 = mean(data$Estradl[which(data$Ethnic == "Caucasian")])
X1
```

```
## [1] 42.99893
```

```r
X2 = mean(data$Estradl[which(data$Ethnic == "African American")])
X2
```

```
## [1] 33.73134
```

- **S_1 y S_2**

```r
S1 <- compute_S(data$Estradl[which(data$Ethnic == "Caucasian")])
S1
```

```
## [1] 18.26891
```

```r
S2 <- compute_S(data$Estradl[which(data$Ethnic == "African American")])
S2
```

```
## [1] 16.51693
```

- **S**

```r
S = sqrt(((length(data$Estradl[which(data$Ethnic == "Caucasian")]) - 1) * S1^2 + (length(data$Estradl[wh
S
```

```
## [1] 17.00944
```

- **t**

Procedemos a calcular el estadístico de contraste t, con la siguiente fórmula.

$$t = \frac{\bar{X_1} - \bar{X_2}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = 3.47$$

```r
error_estandar = (S*sqrt((1 / length(data$Estradl[which(data$Ethnic == "Caucasian")])) + (1 / length(da
t = (X1 - X2) / error_estandar
t
```

```
## [1] 3.476057
```

El valor crítico será $t_{\alpha/2, n_1+n_2-2} = \pm 1.97$

```r
qt(p=0.025, df=203)
```

```
## [1] -1.971719
```

- **Valor crítico**

A continuación, calculamos el p valor. Dado que H_1 mantiene que $\mu_0 - \mu_1 \neq 0$, entonces el p-valor será:

$$p = 2P(t_{n_1+n_2-2} > |t|) = 2P(t_{203} > 3.476) = 2(1 - P(t_{203} < 3.476)) = 0.00062$$

- **P-valor**

```r
2*(1 - pt(t, df = 203))
```

```
## [1] 0.0006222197
```

4.4 Interpretar

Interpretar los resultados y concluir si se puede afirmar que existen diferencias significativas en el nivel de estradiol según la etnia.

Dado que p-valor $< \alpha$, rechazamos la hipótesis nula y concluimos que la media de los niveles de estradiol en mujeres negras y caucásicas son diferentes.

5. Nivel de estradiol según hijos

A continuación, se desea evaluar si existen diferencias en el nivel de estradiol de las mujeres según si han tenido hijos o no. Es decir, se podría afirmar que el nivel de estradiol es inferior en las mujeres que han tenido hijos, con un nivel de confianza del 95 %? ¿Y con un nivel de confianza del 90 %? Seguir los pasos que se indican a continuación para dar respuesta a esta hipótesis (que son análogos a la pregunta anteior). Recordad que se deben realizar los cálculos manualmente. No se pueden usar funciones de R que calculen directamente el contraste de hipótesis. En cambio, sí se pueden usar funciones como qnorm, pnorm, qt y pt.Especificad todos los pasos detalladamente e imprimid los resultados de las variables relevantes de este contraste, tal como se requiere en la sección anterior.

5.1 Hipótesis nula y alternativa

$$\begin{cases} H_0: & \mu_0 < \mu_1 \\ H_1: & \mu_0 \geq \mu_1 \end{cases}$$

Siendo μ_0 las mujeres que han tenido hijos y μ_1 las que no los han tenido.

5.2 Método

Se trata de un contraste bilateral entre dos muestras independientes. Estas muestras tienen una distribución normal, ya que los datos pueden organizarse en forma de campana gaussiana y podremos calcular la probabilidad de que varios valores ocurran en un cierto intervalo dada una confianza, por lo que también podemos afirmar que se trata de un test paramétrico.

5.3 Cálculo

- **X_1 y X_2**

```
X1 = mean(data$Estradl[which(data$Anykids == "Yes")])
X2 = mean(data$Estradl[which(data$Anykids == "No")])
```

- **S_1 y S_2**

```
# S1
S1 <- compute_S(data$Estradl[which(data$Anykids == "Yes")])
S1
```

```
## [1] 17.12854
```

```
# S2
S2 <- compute_S(data$Estradl[which(data$Anykids == "No")])
S2
```

```
## [1] 17.64934
```

- **S**

```
# S
S = sqrt(((length(data$Estradl[which(data$Anykids == "Yes")]) - 1) * S1^2 + (length(data$Estradl[which(
S
```

```
## [1] 17.49447
```

- **t**

```r
# T
error_estandar = S*sqrt((1 / length(data$Estradl[which(data$Anykids == "Yes")])) + (1 / length(data$Esti
t = (X1 - X2) / error_estandar
t
```

```
## [1] 0.5673649
```

- **P-valor**

$$pvalor = P(t_{n_1+n_2-2} > t) = 1 - P(t_{n_1+n_2-2} < t) = 1 - P(t_{203} < t) = 0.285$$

```r
1 - pt(t, 203)
```

```
## [1] 0.2855466
```

5.4 Interpretación

Tanto para un nivel de significación de $\alpha = 0.05$, como para $\alpha = 0.1$, aceptaremos la hipótesis nula, ya que el p-valor $> \alpha$. Es decir, concluiremos que el nivel de estradiol es menor en mujeres que han tenido hijos con una confianza tanto del 95% como del 90% es menor que el nivel de estradiol de las que no han tenido hijos.

6. Estudio longitudinal: ¿Estradiol aumenta con los años?

Los investigadores del estudio encontraron que existía una posible correlación entre la edad de las mujeres y el nivel de estradiol. Por ello, realizaron un estudio longitudinal con una muestra reducida de mujeres. En un grupo de 10 mujeres voluntarias de la muestra original, se midió los niveles de estradiol al cabo de 7 años. El fichero ESTRAD7.csv recoge esta medida. Concretamente, el fichero contiene el identificador de la mujer, el nivel de estradiol original y su nivel de estradiol medido al cabo de 7 años del estudio original. La hipótesis de la investigación es que estradiol aumenta con la edad. ¿Qué dicen los datos en relación a esta hipótesis? ¿Podemos afirmar que aumenta el nivel de estradiol con un nivel de confianza del 97 %?

6.1 Hipótesis nula y alternativa

$$\begin{cases} H_0: & \mu_0 = \mu_1 \\ H_1: & \mu_0 < \mu_1 \end{cases}$$

Siendo μ_0 la media del nivel de estradiol original y μ_1 la media tras 7 años. La hipótesis principal indica que el nivel se mantiene, mientras que la alternativa sostiene que el nivel de estradiol aumenta con la edad.

6.2 Asunción de normalidad

Para determinar el tipo de prueba a aplicar, se comprueba primero si se cumple la asunción de normalidad de los datos. Para ello, podemos examinar si se puede aplicar el teorema del límite central. Además, se puede realizar una visualización gráfica con las curvas Q-Q y aplicar el test de Shapiro-Wilk. En este apartado debéis realizar estas comprobaciones y determinar si se puede asumir normalidad. Justificar vuestra conclusión en base a los resultados obtenidos.

```r
data <- read.csv(paste(current_working_directory,"/ESTRADL7.csv", sep = ""))
X1 <- mean(data$Estrad)
X2 <- mean(data$Estradl7)
```

En este caso, dado que desconocemos la desviación típica de la población, tendremos una distribución t de Student con 6 grados de libertad (t_6). Es una desviación similar a la normal (0, 1): simétrica alrededor de cero. Sin embargo, su desviación típica es ligeramente superior que la normal, ya que los valores que toma

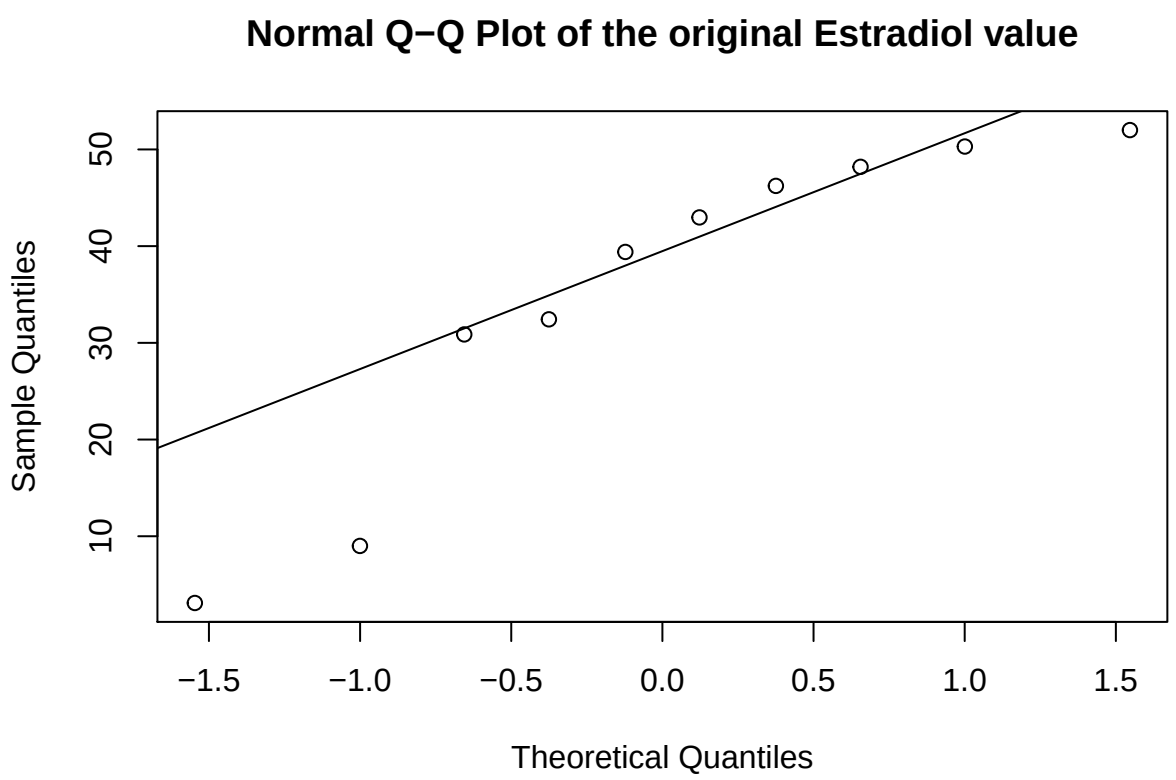
esta variable están un poco más dispersos (cuanto mayor es el número de grados de libertad, más se acerca a la distribución normal).

En este caso, podemos suponer que la media del nivel de estradiol en mujeres sigue una distribución normal, aunque desconocemos la desviación típica.

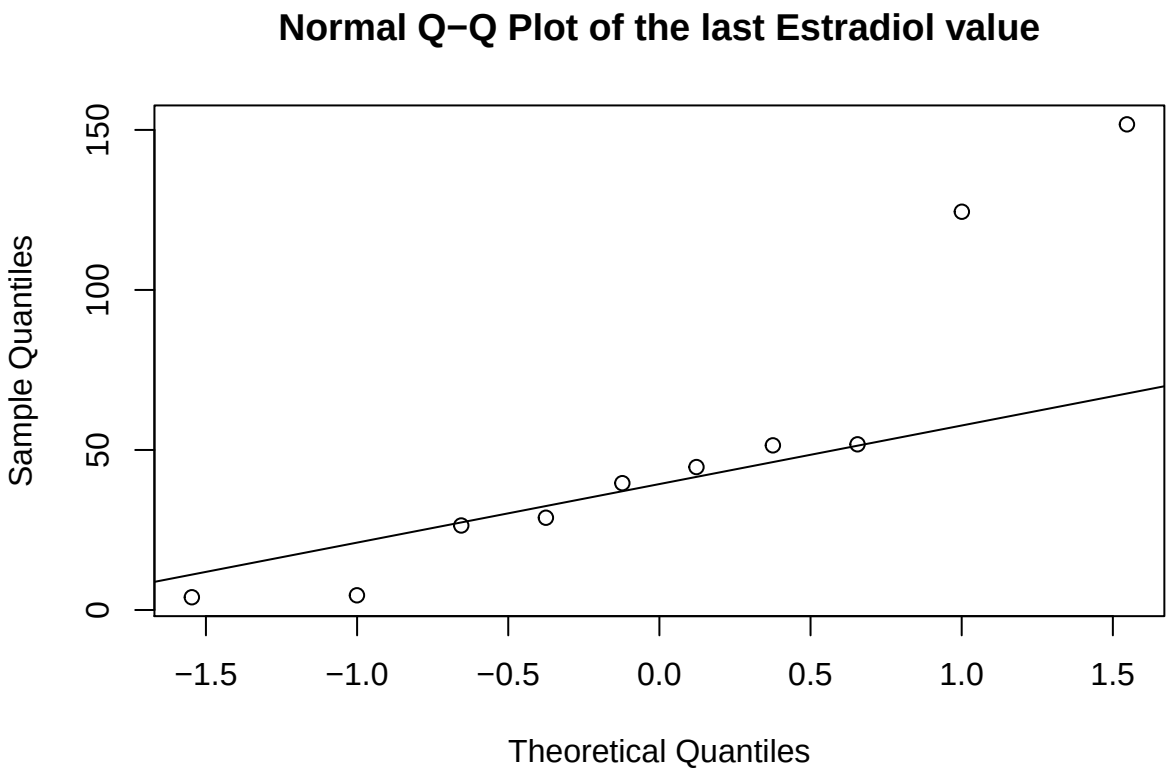
- **Teorema del límite central**: Dicho teorema enuncia que las muestras de tañano suficientemente grande, siendo n $>$ 30, podemos afirmar que se trata de una distribución normal. Dado que no es el caso, debemos comprobar que la distribución de la variable $\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}$ es una normal estándar.

- **Curvas Q-Q**:

```
qqnorm(y = data$Estrad, main = "Normal Q-Q Plot of the original Estradiol value")
qqline(y = data$Estrad)
```




```
qqnorm(y = data$Estradl7, main = "Normal Q-Q Plot of the last Estradiol value")
qqline(y = data$Estradl7)
```



- **Shapiro-Wilk**: La inspección visual no siempre es del todo fiable, ya que en este caso, dada la escasa muestra, la diferencia entre los puntos en el gráfico es más notable. Realizaremos un test en R lamado Shapiro-Wilk para comprobar si se trata de una distribución normal. Este test será positivo si el p-valor resultante es mayor que 0.05 convencionalmente, aunque en nuestro caso $\alpha = 0.03$.

```
shapiro.test(data$Estrad)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$Estrad
## W = 0.84957, p-value = 0.05741
```

```
shapiro.test(data$Estradl7)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$Estradl7
## W = 0.83177, p-value = 0.03515
```

Este test nos indica que ambas variables tienen una distribución normal, ya que p-value $> \alpha$.

6.3 Método

Independientemente de la conclusión obtenida en el apartado anterior, aplicar un test no paramétri- co a la muestra. Existen dos tipos de tests no paramétricos para el contraste de dos muestras: 1) el test de suma

de rangos (también conocido como test U de Mann-Whitney) y 2) el test de rangos y signos de Wilcoxon. Decidid cuál es el contraste que debe aplicarse en este caso. Debéis justificar vuestra elección.

A la hora de escoger entre los tests no-paramétricos *"Suma de rangos (o test U)"* y *"test de rangos y signos del Wilcoxon"* partimos de la base de que el primero establece la comparación entre medias de poblaciones independientes, mientras que el segundo se trata de un test para comparar medias entre muestras dependientes, como es nuestro caso. Por ello, aplicaremos el **test de rangos y signos de Wilcoxon**.

6.4 Cálculo e Interpretación

Aplicad el contraste e interpretar el resultado. Podéis usar funciones R (no desarrolléis los cálculos en este apartado).

```
wilcox.test(data$Estrad, data$Estradl7, paired = TRUE, alternative = "l", conf.level = 0.97)
```

```
##
##  Wilcoxon signed rank test
##
## data:  data$Estrad and data$Estradl7
## V = 25, p-value = 0.4229
## alternative hypothesis: true location shift is less than 0
```

El resultado indica que p-value = 0.4229, el cual es mayor que el nivel de significación ($\alpha = 0.03$), por tanto aceptamos la hipótesis alternativa H_1, que indica que el nivel de estradiol en mujeres es mayor con el paso del tiempo.

6.5 Expicar el test escogido

Explicar brevemente cómo se calcula el test que habéis escogido en el apartado anterior. La explicación no debe ser en base a un código, sino en vuestras propias palabras. Tratad de ser claros y concisos en la explicación

El **test de rangos y signos de Wilcoxon** es una alternativa al test de t-Student para muestras dependientes cuando la muestra es pequeña y no puede asumirse normalidad.

El test establece como hipótesis principal que la muestra es simétrica alrededor de 0, y como hipótesis alternativa que no es una muestra de distribución normal. A continuación calcula un valor V, que representa la suma de los rangos positivos de nuestra diferencia de muestras. Es decir, se calcula la diferencia de medias de cada mujer y sus valores absolutos se ordenan de menor a mayor, asignando un índice a cada valor. La suma de los indices de la diferencia de medias que fueron positivas será V y la suma de los indices negativos será V'. En nustro caso V = 25 y V' = 30. Podemos comprobar que dichos resultados son correctos a través de la siguiente fórmula.

$$V + V' = \frac{n(n+1)}{2}$$

Si V + V' > 20, podemos asumir que tiene una distribución normal, y procedemos a calcular el p-valor. [1]

7. Conclusiones

Para finalizar, se presenta a continuación un resumen de los conceptos aprendidos en esta práctica, como los pásos a seguir para realizar un contraste de hipótesis de dos muestras.

Para realizar un contraste de hipótesis, debemos establecer una hipótesis nula (H_0) y otra alternativa (H_1), una de las cuales aceptaremos y otra rechazaremos tras el estudio. Estas hipótesis determinarán de si se

[1] https://www.statisticssolutions.com/how-to-conduct-the-wilcox-sign-test/

trata de un test bilateral o unilateral. Si hay dos areas de rechazo, se trata de un test bilateral, mientras que si solo hay un valor crítico, que determina un arera de rechazo, se tratará de una hipótesis unilateral. En segundo lugar, estableceremos el nivel de significación, que se trata del error máximo que podemos asumir. Es decir, si el enunciado indica que debemos establecer una hipótesis con una confianza del 99%, el error máximo que podemos asumir es $\alpha = 0.01$. A continuación, para estudiar si aceptamos la hipótesis nula o no, elaboramos el estadístico de contrase. Si no disponemos de la varianza, el estadístico de contraste seguirá una distribución t-Student. Dicha distribución aproximará la varianza (σ) a un valor S. Asumiendo que la muestra tiene una distribución normal, realizaremos los cálculos pertinentes para obtener un p-valor, el cual determinará si aceptamos la hipótesis nula (p-valor $\geq \alpha$) o la rechazamos (aceptamos la alternativa si p-valor $< \alpha$). En el caso de los contrastes bilaterales, también podemos realizar el contraste de hipótesis con intervalos de confianza, es decir, dado un cierto valor de significación, determinamos los valores entre los que se encuentra el rango de aceptación de H_0, así si el valor que queremos contrastar se encuentra en ese rango, aceptamos H_0. Si en vez de contrastar con un valor, queremos hacerlo sobre la media, únicamente podremos hacerlo buscando el p-valor, sin calcular el valor crítico. Al igual que en el contraste de un valor, si no disponemos de la varianza, seguiremos una distribución t-Student. Podremos seguir los mismos pasos aunque no sepamos si se trata de muestras normales, siempre que n > 30. Si disponemos de valores aparejados, es decir, observaciones de variables diferentes que pertenecen a los mismos individuos, podremos obtener la diferencia para estudiar una única muestra.

En esta práctica también hemos realizado contrastes de hipótesis sobre diferencia de medias poblacionales, debemos determinar en primer lugar si se trata de una distribución normal o no, como anteriormente. En el caso en el que nuestra muestra no tuviese una distribución normal, aplicamos el teorema del limite central, con el cual aproximaremos la muestra a una distribución normal para muestras grandes (n > 30). Diferenciaremos los casos en los que las medias son dependientes o independientes, es decir, si las observaciones corresponden a los mismos individuos o no. También diferenciaremos los casos en los que conocemos la varianza (σ) o no, ya que si no la conocemos la aproximaremos a un valor S. Además, como anteriormente, podemos tomar una decisión, para hipótesis bilaterales, además de basándonos en el p-valor, haciendo uso del intervalo de confianza.

Las pruebas previamente resumidas son paramétricas, ya que conocemos que se ajustan a una distribución normal o lo hemos asumido, si bien es cierto que también hemos aprendido a aplicar algunos test no paramétricos, que se usan cuando la distrubución de la muestra no puede ser asumida. Este es, por ejemplo, el **test de rangos y signos de Wilcoxon** utilizado el el punto 6.