PEC 2 INFERENCIA ESTADÍSTICA 2019-2020 SEMESTRE 1

Marc Bañuls Tornero

3/1/2020

Contents

1.Cuestiones sobre contrastes de hipótesis **2**

A) El dolor postoperatorio es frecuente en los hospitales. Se han desarrollado guías de práctica clínica para el control del dolor postoperatorio (GPCDP) y para mejorar su tratamiento. Con el objetivo de ver si la implementación de la GPCDP tiene impacto en la disminución de la prevalencia del dolor, se elige un grupo de 36 pacientes operados antes de la GPCDP y otro grupo diferente de 30 pacientes operados después de la GPCDP. Indica el contraste de hipótesis y suposiciones para comprobar si ha disminuido el % de dolor tras la implementación de la GPCDP. 2

B) Se desea analizar la asociación entre concentraciones de tres citosinas: interferón-1α (INF1α), interleucina 10 (IL-10) y BLyS con la actividad clínica en el lupus eritematoso sistémico (LES). Para ello se dispone de 142 pacientes con LES y 34 controles a los que se les mide las tres citocinas. Indica el contraste de hipótesis y las suposiciones para comprobar si hay asociación entre cada una de las citocinas (por separado) y el estar afectado por LES. 2

C) La obesidad es un factor de riesgo asociado a la aparición de las diabetes. Se dispone de un grupo de 15 sujetos pre-diabéticos a los que se les pone en dieta apoyada por un tratamiento. Se mide el peso en el momento basal y a los 6 meses del tratamiento. Indica el contraste de hipótesis y las suposiciones para ver si ha habido un cambio en el peso de los pacientes. . . . 3

D) Se estudia un grupo de 33 pacientes afectados de Carcinoma hepatocelular, un grupo de 22 afectados únicamente de cirrosis y un grupo control de 31 donantes de sangre. En cada uno de los sujetos de cada grupo se determinan la actividad celular NK y el número de células CK. (Nota: En las determinaciones efectuadas se observa que la media de actividad celular es de 39 unidades líticas $/10^7$ de linfocitos y la mediana es de 28 y que la media del número de células es de 178 y la mediana de 163). Indica el contraste de hipótesis y las suposiciones para ver si hay diferencias de cada una de las determinaciones (por separado) entre los grupos de pacientes. 3

E) Se dispone de 45 mujeres obesas de las cuales 22 eran premenopáusicas y 23 postmenopáusicas. Se tomaron medidas como el índice de masa corporal, el índice cintura-cadera y cintura-muslo. Además, se clasificó a las mujeres según su obesidad fuera androide o ginoide. Se está interesado en evaluar la relación entre la menopausia y los índices de obesidad. También entre la menopausia y el tipo de obesidad. Indica el contraste de hipótesis y las suposiciones para ver si hay diferencias de los índices de obesidad y el tipo de obesidad (por separado) con el grupo de mujeres según su menopausia. 4

2 Ejercicio práctico **4**

Los datos 4

Muestreando la base de datos 4

Las preguntas 5

Algunas cosas más Ejercicio 3 **18**

a) Siguiendo las instrucciones debajo vamos a generar 1000 veces tres muestras de tamaño 80 bajo la distribución normal con la misma media y varianza. Para cada conjunto de datos comparamos la igualdad de medias entre la muestra 1 y 2, la muestra 1 y 3 y la muestra 2 y 3 con un error tipo I de ($\alpha = 0.05$ cada uno) y guarda si se acepta o no la hipótesis nula. 18

A) Comparaciones múltiples 18

B) Bootstrap. Responder a las preguntas de forma breve y razonada 20

1.Cuestiones sobre contrastes de hipótesis

A continación, se presentan diversas situaciones de diferentes estudios basados en datos reales. Se trata de indicar el tipo de análisis o de pruebas estadísticas a utilizar en cada uno de los casos e indicar, si es necesario, si hay que efectuar algún tipo de prueba adicional para llevar a cabo el análisis. Formula las hipótesis oportunas para cada tipo de prueba.

A) El dolor postoperatorio es frecuente en los hospitales. Se han desarrollado guías de práctica clínica para el control del dolor postoperatorio (GPCDP) y para mejorar su tratamiento. Con el objetivo de ver si la implementación de la GPCDP tiene impacto en la disminución de la prevalencia del dolor, se elige un grupo de 36 pacientes operados antes de la GPCDP y otro grupo diferente de 30 pacientes operados después de la GPCDP. Indica el contraste de hipótesis y suposiciones para comprobar si ha disminuido el % de dolor tras la implementación de la GPCDP.

Nos encontramos con dos variables independientes (con 36 y 30 muestras) a las que se le han medido la misma variable cuantitativa (dolor postoperatorio). Para observar si ha habido disminución de dolor al utilizar las GPCDP. Primero deberíamos comprobar si las poblaciones son paramétricas. Para ello deben de cumplir que sean independientes, con una distribución normal y con una varianza similar (homocedásticas). Primero observamos la homocedasticidad y para ello realizaríamos un test de Levene (test para comparar varianzas entre grupos) tomando como hipótesis nula que no hay diferencias entre las varianzas de ambas poblaciones. Además realizaríamos un test de normalidad para ambas variables independientes teniendo como hipótesis nulas que la variable sigue una distribución normal.
En el caso de que finalmente aceptemos que las poblaciones son paramétricas realizaríamos un contraste de hipótesis en un test-T (test de comparación de medias entre dos grupos paramétricos), donde indicaríamos que la hipótesis nula es que no haya diferencia entre las medias o que el grupo después de la GPCDP tenga mayor % de dolor, y la hipótesis alternativa que la media de dolor del grupo post GPCDP sea menor que la media del grupo pre-GPCDP. De esta manera podemos saber si ha habido una disminución en el dolor postoperatorio debido a utilizar o no las GPCDP.
En el caso de que no aceptemos que las poblaciones son paramétricas realizamos el mismo contraste de hipótesis anterior pero con el test de Mann-Whitney U (test de comparación de medias entre dos grupos no paramétricos).
Si las variables siguen una distribución normal pero sus varianzas no son iguales, utilizaríamos el test de Welch (con este test asumimos que la distribución es normal pero las varianzas no son iguales) utilizando el mismo contraste hipótesis anterior.

B) Se desea analizar la asociación entre concentraciones de tres citosinas: interferón-1α (INF1α), interleucina 10 (IL-10) y BLyS con la actividad clínica en el lupus eritematoso sistémico (LES). Para ello se dispone de 142 pacientes con LES y 34 controles a los que se les mide las tres citocinas. Indica el contraste de hipótesis y las suposiciones para comprobar si hay asociación entre cada una de las citocinas (por separado) y el estar afectado por LES.

Tenemos tres variables cuantitativas (se miden tres tipos de citosinas) recogidas en dos grupos categóricos (LES y Control). En principio este ejercicio se puede resolver de igual manera que en el anterior, pero debemos realizar el mismo proceso para cada tipo de citosinas. Primero observamos si las variables son paramétricas (test de homocedasticidad con el test de Levene y normalidad con el test de Shapiro-Wilk) en los valores del grupo LES y del grupo Control. En caso de que sí sean paramétricas realizamos un test T considerando como hipótesis nula que no hay diferencias entre la media de concentración de una citosina específica en el

grupo control con la media de concentración de la misma citosina del grupo LES, y como alternativa que sí que hay diferencia entre esas medias. Repitiendo esto en las tres citosinas entre el grupo control y el grupo LES sabremos si hay una diferencia en la concentración de citosinas entre los distintos grupos.

Si las variables siguen una distribución normal pero no son homocedásticas, utilizaríamos el test de Welch con el mismo contraste de hipótesis para cada tipo de citosinas.

En el caso de que las muestras no sean paramétricas, realizaríamos el test de Mann-Whitney U para cada tipo de citosinas utilizando el mismo contraste hipótesis.

C) La obesidad es un factor de riesgo asociado a la aparición de las diabetes. Se dispone de un grupo de 15 sujetos pre-diabéticos a los que se les pone en dieta apoyada por un tratamiento. Se mide el peso en el momento basal y a los 6 meses del tratamiento. Indica el contraste de hipótesis y las suposiciones para ver si ha habido un cambio en el peso de los pacientes.

En este caso nos encontramos con una sola variable independiente dividida en dos partes (peso basal y peso a los 6 meses). Por ello debemos trabajar con la diferencia del valor peso entre los 6 meses y el peso basal y calcular la media poblacional. Al estar hablando de una única variable independiente que cambia con el tiempo, podemos realizar un test-T emparejado (paired T-test, test de comparación de medias entre dos grupos paramétricos con muestras aparejadas) considerando como hipótesis nula que no hay diferencias entre los valores en distintos tiempos (diferencia entre peso en 6 meses y peso basal = 0) y como hipótesis alternativa que sí hay diferencias significativas en el peso.

D) Se estudia un grupo de 33 pacientes afectados de Carcinoma hepatocelular, un grupo de 22 afectados únicamente de cirrosis y un grupo control de 31 donantes de sangre. En cada uno de los sujetos de cada grupo se determinan la actividad celular NK y el número de células CK. (Nota: En las determinaciones efectuadas se observa que la media de actividad celular es de 39 unidades líticas /10^7 de linfocitos y la mediana es de 28 y que la media del número de células es de 178 y la mediana de 163). Indica el contraste de hipótesis y las suposiciones para ver si hay diferencias de cada una de las determinaciones (por separado) entre los grupos de pacientes.

Queremos observar si existen diferencias entre los resultados de la variable cuantitativa "Actividad celular NK" entre los tres grupos de pacientes y si existen diferencias entre los resultados de la variable cuantitativa "Número de células CK" entre los tres grupos de pacientes. Para saber si las dos variables cuantitativas tienen una distribución normal o no se nos muestra la media y mediana de cada variable. Cuando una variable cuantitativa está distribuida normalmente su media y su mediana tienen el mismo valor. En este caso, las dos variables cuantitativas tienen valores distintos de media y mediana, por lo que podemos asumir que estas dos variables no siguen una distribución normal.
Sabiendo esto, podemos realizar un test de Kruskal-Wallis (test de comparación de medias con más de dos grupos no paramétricos) teniendo como hipótesis nula que no hay diferencias significativas en los tres grupos. Como hipótesis alternativa indicaríamos entonces que sí hay diferencias significativas entre los tres grupos. En el caso de que se acepte la hipótesis alternativa, deberíamos realizar un test de Dunn (test de comparacion de medias 2 a 2 en grupos no paramétricos) para encontrar entre qué grupos se encuentran las diferencias significativas. Aquí tendríamos varias hipótesis nulas y hipótesis alternativas, donde las hipótesis nulas indicarían que no hay diferencias significativas entre dos grupos y las hipótesis alternativas indicarían que sí hay diferencias significativas entre dos grupos concretos. Cabe destacar que debemos realizar todas estas hipótesis dos veces, una para la variable cuantitativa "Actividad celular NK" y otra vez para la variable cuantitativa "Número de células CK".

E) Se dispone de 45 mujeres obesas de las cuales 22 eran premenopáusicas y 23 postmenopáusicas. Se tomaron medidas como el índice de masa corporal, el índice cintura-cadera y cintura-muslo. Además, se clasificó a las mujeres según su obesidad fuera androide o ginoide. Se está interesado en evaluar la relación entre la menopausia y los índices de obesidad. También entre la menopausia y el tipo de obesidad. Indica el contraste de hipótesis y las suposiciones para ver si hay diferencias de los índices de obesidad y el tipo de obesidad (por separado) con el grupo de mujeres según su menopausia.

En este apartado medimos tres variables de datos (índice de masa corporal, índice cintura-cadera y índice cintura-muslo) y estas tres variables se encuentran divididas en dos grupos (premenopáusicas y postmenopáusicas). Además tenemos en ambos grupos una subdivisión de los sujetos en otros dos grupos (androide y ginoide). En este apartado debemos realizar dos contrastes de hipótesis, uno para encontrar diferencias en los índices de obesidad y la menopausia y otro contraste para encontrar si hay diferencias en el tipo de obesidad y la menopausia.

Respecto a la relación con los índices de obesidad y menopausia, primeramente deberíamos saber si las muestras son paramétricas o no. Para ello deberíamos realizar un test de normalidad de las variables de los datos para saber si las muestras siguen una distribución normal (test de Shapiro-Wilk) y un test de homocedasticidad (test de Levene) para saber si sus varianzas son similares.

En el caso de que las muestras sean paramétricas, podemos realizar un test-T entre las mujeres premenopáusicas y postmenopáusicas, indicando como hipótesis nula que no hay diferencias significativas en sus medias. Si por el contrario las muestras no son paramétricas, deberíamos usar otro tipo de test. Para dos muestras independientes no paramétricas utilizaríamos el test de Mann-Whitney U con el mismo contraste de hipótesis.

Respecto a la relación con el tipo de obesidad y menopausia estamos tratando con variables categóricas, por lo que utilizaríamos un test Chi-cuadrado donde la hipótesis nula sería que no hay diferencias significativas en la frecuencia de aparición de obesidad androide o ginoide y la menopausia, y la hipótesis alternativa, que sí hay diferencias entre la frecuencia de aparición de estos dos tipos de obesidad y la menopausia. En concreto, podríamos utilizar el test exacto de Fisher para realizar este contraste de hipótesis.

2 Ejercicio práctico

Este ejercicio consta de diversas partes en un intento de simular lo que se lleva a cabo en un estudio real. Se ha simplificado para hacerlo más practicable por lo que no hace falta que os agobiéis si algo no os cuadra del todo. De lo que se trata es que veamos cómo aplicar las distintas técnicas que hemos estudiado, de forma integral, en un problema de análisis de datos.

Los datos

En el fichero MUNS.DAT se encuentran los datos relativos al proyecto MUNS, un estudio sobre las desigualdades en Salud en Unidades de cuidados intensivos. El objetivo del estudio era demostrar las diferencias entre mortalidad en UCI según el nivel socioeconómico de los pacientes. Además de la base de datos en formato texto fijo la base está disponible en SPSS (Muns.sav), en Stata(Muns.dta) y en Excel (Muns.xls).

Muestreando la base de datos

Empezaremos tomando una muestra de la base de datos original de 700 casos. Esto garantiza que estamos trabajando con un tamaño con el que es más manejable y a la vez determina que cada uno trabaje con un "dataset" distinto.

```
library(readxl)
Muns <- read_excel("Muns.xls")
```

Para sacar una muestra de 700 individuos del conjunto de datos activos podéis usar el código siguiente una vez leídos los datos. Primero hay que fijar la semilla para que al repetir el código, seleccione siempre los mismos casos sustituyendo #número por un número. Atención, el código asume que vuestro dataset se llama Muns. Si no es así adaptadlo.

```
set.seed(123)
Muns700 <- Muns[sample(1:nrow(Muns), 700, replace = FALSE),]
```

Las preguntas

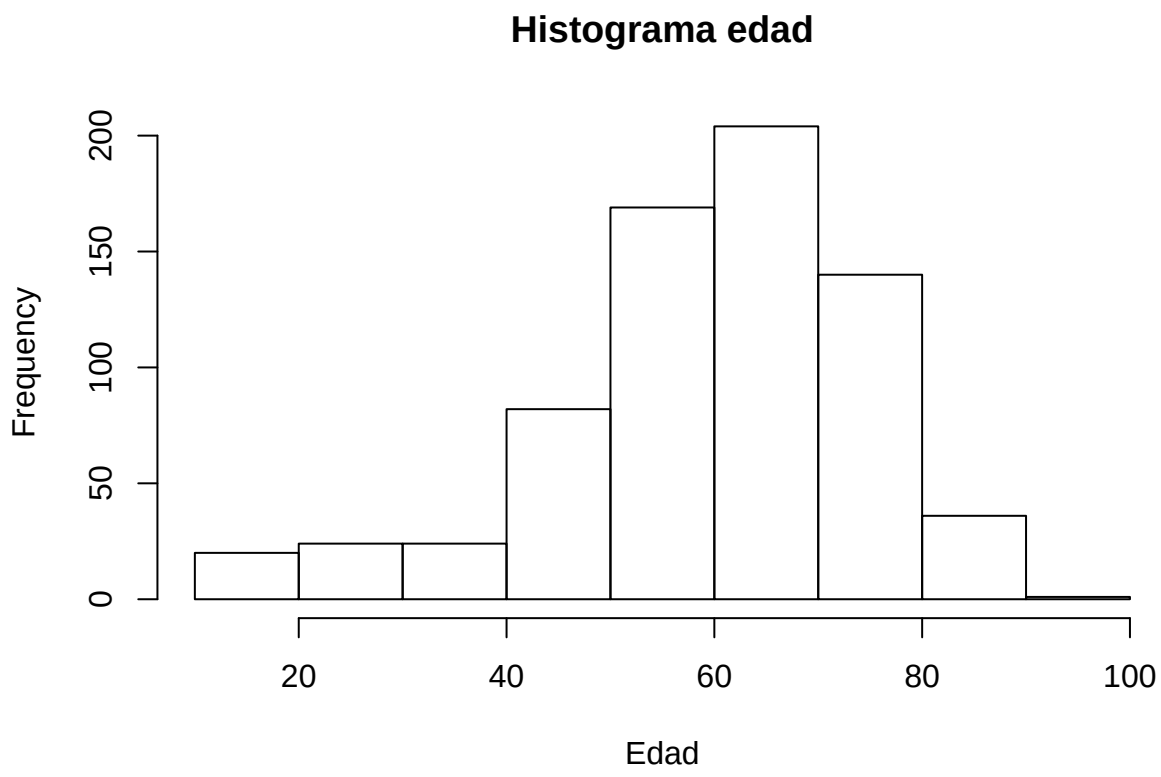
a) En algunas situaciones se requiere que las variables cuantitativas sean normales. Comprueba la normalidad de las variables edad, días de estancia, SAPS y TISS. Muetra los resultados tanto gráfica como analíticamente.

Para observar si existe normalidad en estas variables, podemos realizar un test de normalidad de Shapiro-wilk en cada variable. Además, realizaremos para cada variable un histograma, boxplot y qqnorm para observar gráficamente que las variables tienen una distribución normal.

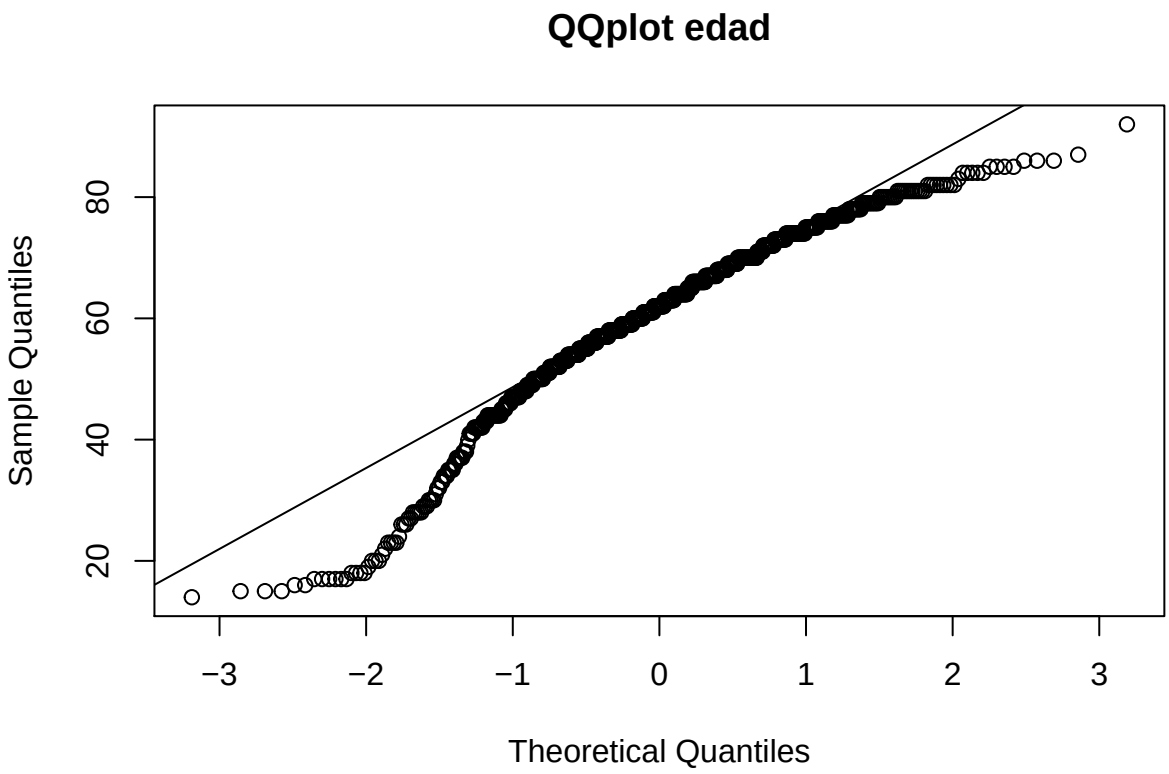
Variable edad:

Realizamos los gráficos para observar la normalidad de la variable:

```
hist(Muns700$edad, main="Histograma edad", xlab = "Edad")
```




```
qqnorm(Muns700$edad, main="QQplot edad")
qqline(Muns700$edad)
```



Realizamos el test de normalidad:

```
shapiro.test(Muns700$edad)
```

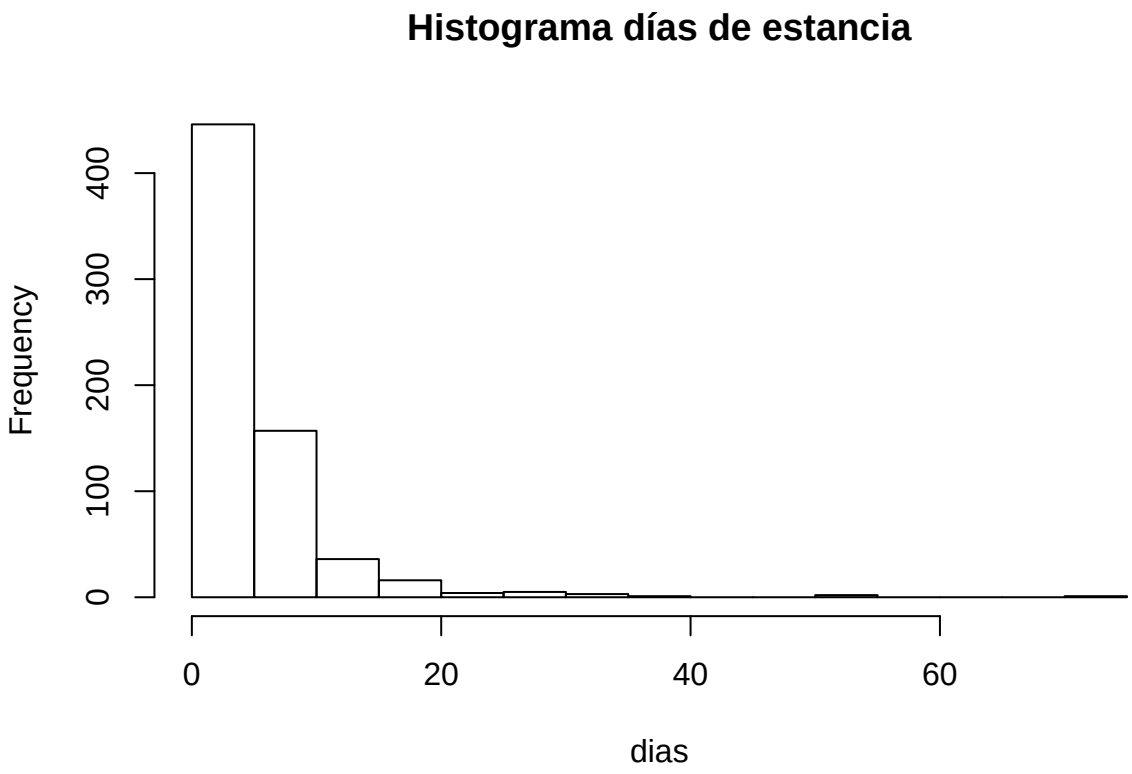
```
##
##  Shapiro-Wilk normality test
##
## data:  Muns700$edad
## W = 0.9469, p-value = 3.771e-15
```

El histograma parece seguir una leve distribución normal con una simetría de los datos, y en el QQplot se observa una tendencia normal de la distribución. Finalmente, estudiando el resultado del test de Shapiro-Wilk, el p-valor sugiere que la variable no tiene (su valor es ampliamente menor a 0.05) una distribución normal. Por lo tanto, basándonos en el histograma y QQplot podemos decir que esta variable cuantitativa sigue una distribución normal.

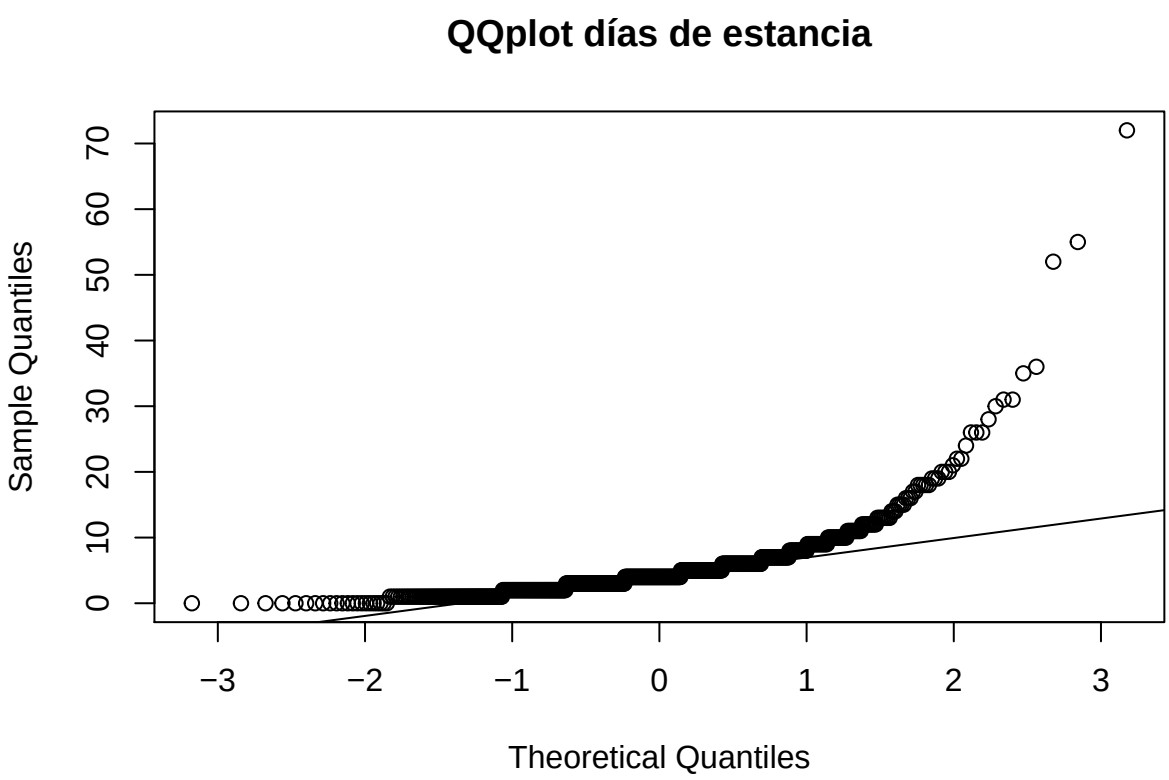
Variable días de estancia

Realizamos los gráficos para observar la normalidad de la variable:

```
hist(Muns700$dias, main="Histograma días de estancia", xlab = "dias")
```




```
qqnorm(Muns700$dias, main="QQplot días de estancia")
qqline(Muns700$dias)
```



Realizamos el test de normalidad:

```
shapiro.test(Muns700$dias)
```

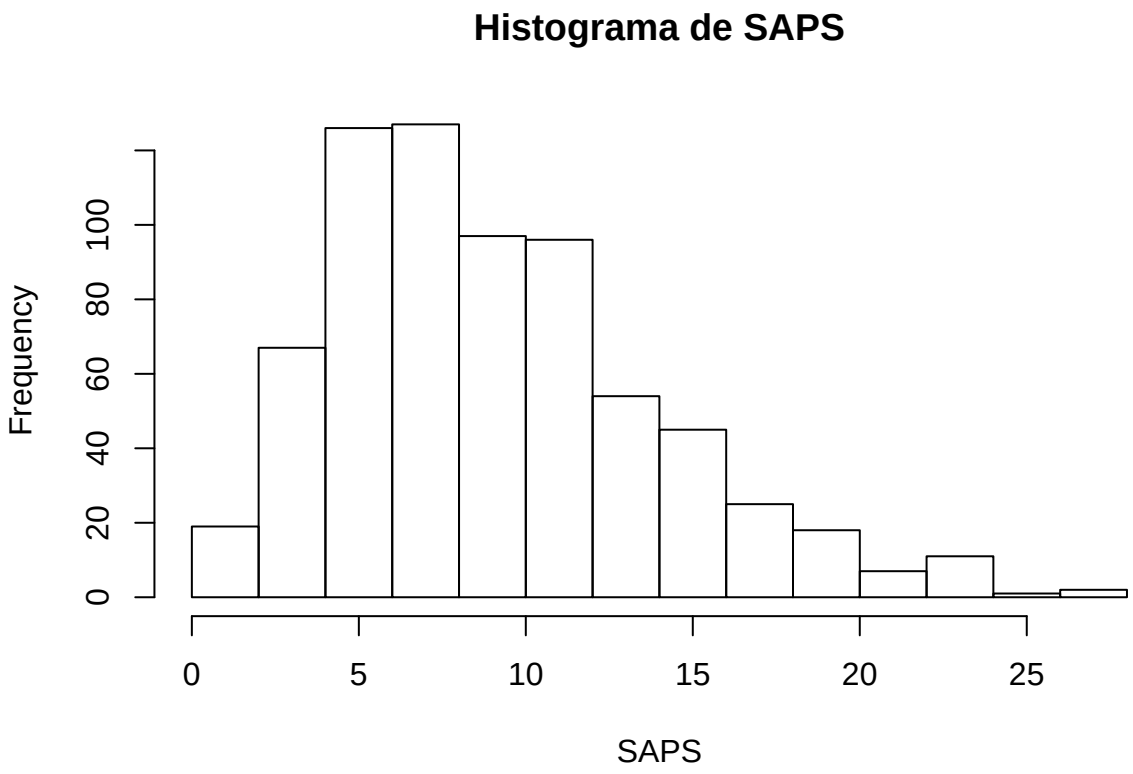
```
##
##  Shapiro-Wilk normality test
##
## data:  Muns700$dias
## W = 0.62015, p-value < 2.2e-16
```

El histograma y el QQplot indican de manera clara que la variable edad no sigue una distribución normal, cosa que se confirma con el test de shapiro-Wilk con un p-valor mucho menor a 0.05.

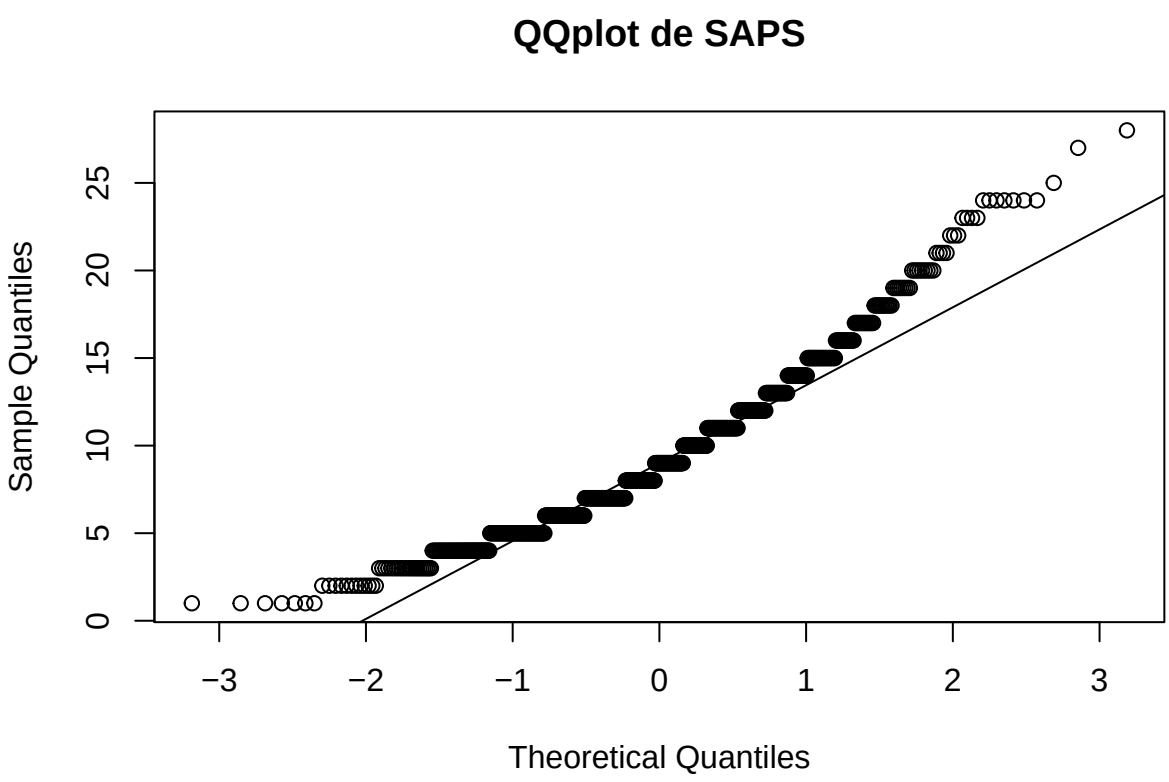
Variable SAPS

Realizamos los gráficos para observar la normalidad de la variable:

```
hist(Muns700$saps, main="Histograma de SAPS", xlab = "SAPS")
```




```
qqnorm(Muns700$saps, main="QQplot de SAPS")
qqline(Muns700$saps)
```



Realizamos el test de normalidad:

```
shapiro.test(Muns700$saps)
```

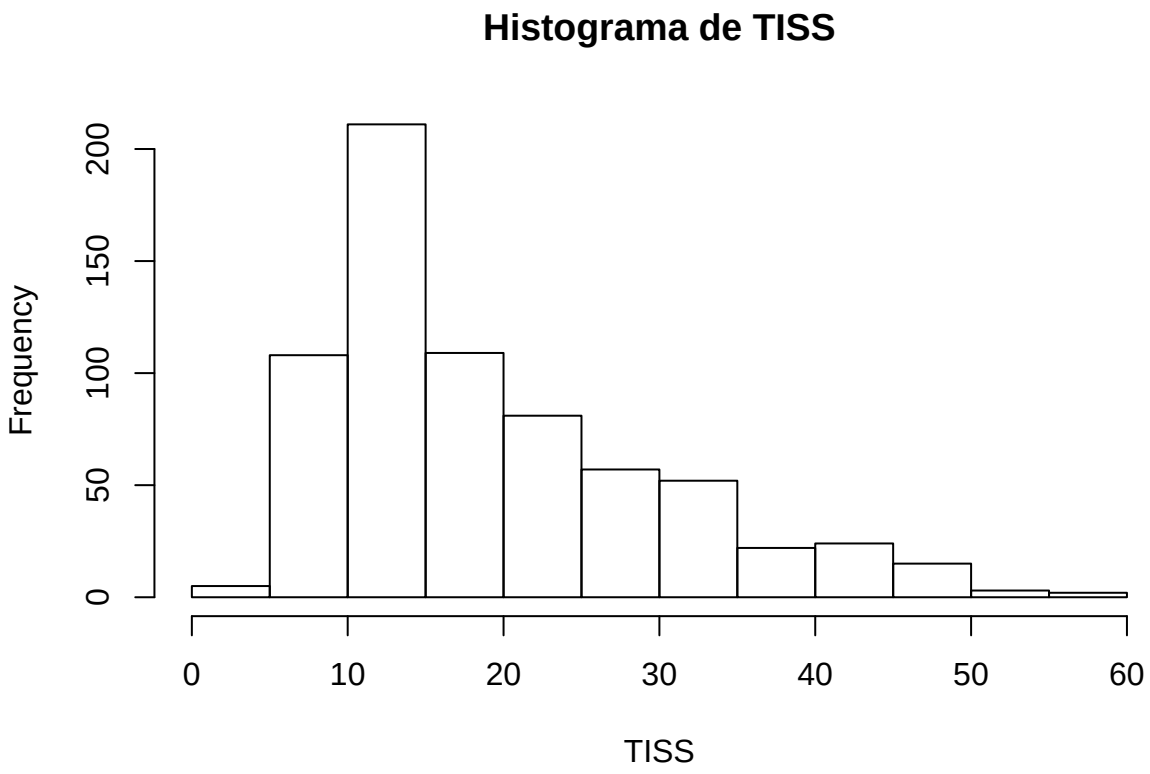
```
##
##  Shapiro-Wilk normality test
##
## data:  Muns700$saps
## W = 0.94756, p-value = 5.525e-15
```

El histograma indica que la variable sigue una distribución normal, ya que se observa simetría, hipótesis que se mantiene al observar el QQplot. En el test de normalidad de Shapiro-Wilk sin embargo observamos que la variable no parece seguir una distribución normal (su p valor es menor a 0.05). Por lo tanto, teniendo en cuenta la simetría de los valores y la linearidad del QQ-plot podemos decir que la variable "SAPS" sigue una distribución normal.

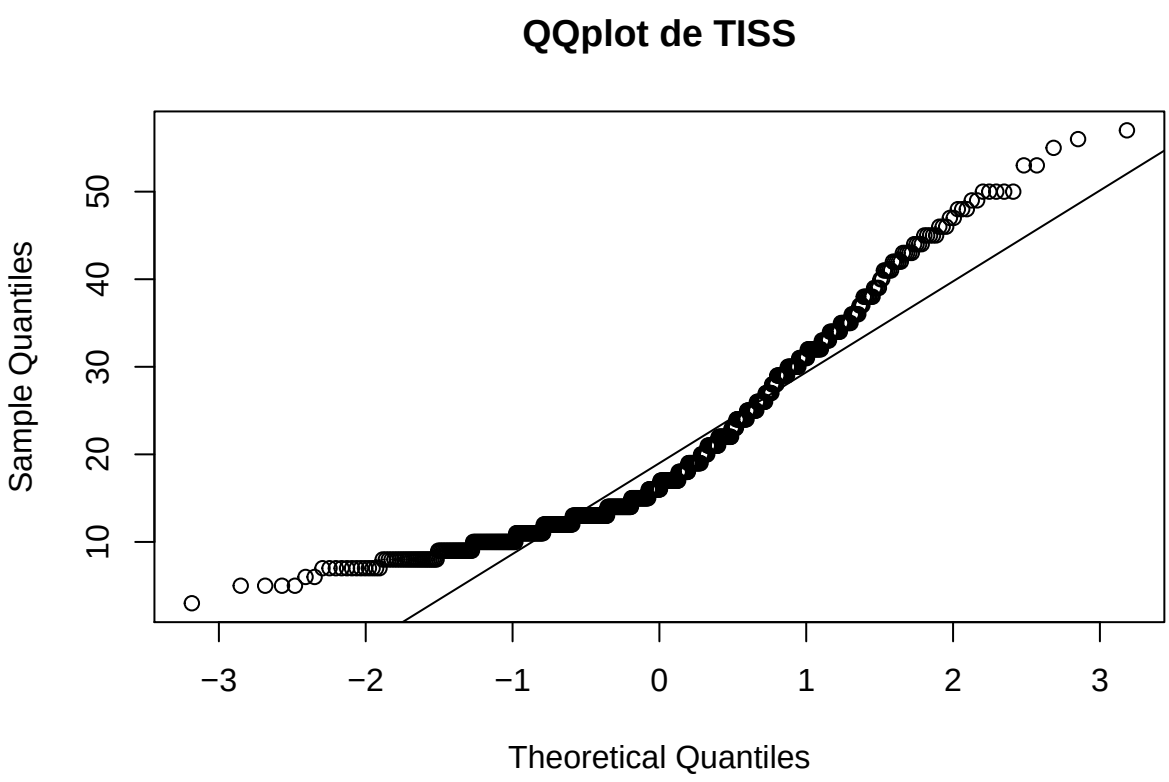
Variable TISS

Realizamos los gráficos para observar la normalidad de la variable:

```
hist(Muns700$tiss, main="Histograma de TISS", xlab = "TISS")
```




```
qqnorm(Muns700$tiss, main="QQplot de TISS")
qqline(Muns700$tiss)
```



Realizamos el test de normalidad:

```
shapiro.test(Muns700$tiss)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Muns700$tiss
## W = 0.89468, p-value < 2.2e-16
```

En el histograma podríamos pensar que los datos siguen una distribución normal aunque existe poca simetría (los valores víran hacia la parte derecha del histograma) que pierde significancia en el QQplot, y se descarta la normalidad de la variable TISS al observar el p-valor claramente menor a 0.05.

b) ¿Existen diferencias entre SAPS y la mortalidad? ¿Existen diferencias en el número de días de estancia en Uci y la mortalidad? Indica los contrastes de hipótesis a realizar y comenta los resultados de su aplicación.

Diferencias entre SAPS y mortalidad

Antes de todo, podemos modificar la variable "mortdos" que ahora es numérica para presentarla como factor con los niveles equivalentes a los grupos que tiene esta variable. De esta manera los resultados serán más entendibles:

```
Muns700$mortdos <- as.factor(Muns700$mortdos)
levels(Muns700$mortdos) = c("Vivo", "Fallecido", "Desconocido")
```

Para analizar si hay diferencias entre SAPS y la mortalidad debemos saber si las muestras son paramétricas o no. En el anterior apartado ya hemos visto que la variable "saps" sigue una distribución normal, por lo que podemos asumir que trabajamos con variables paramétricas. En este caso queremos observar las diferencias

entre dos variables, la variable cuantitativa "saps" y la variable cualitativa con tres niveles "mortdos". Al asumir normalidad utilizaremos entonces el test ANOVA para ver si hay diferencias entre los tres grupos de la variable "mortdos" y la variable "saps". Para ello, observamos mediante el test de Levene si las variables son homocedásticas:

```
library(car)
```

```
## Warning: package 'car' was built under R version 3.6.2
```

```
## Loading required package: carData
```

```
leveneTest(Muns700$saps ~ Muns700$mortdos)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value  Pr(>F)
## group   2   3.039 0.04852 *
##       692
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Observamos que por centésimas se puede aceptar la hipótesis alternativa en la que se indicaría que las varianzas no son iguales. Igualmente, al ser un valor que puede llegar a interpretarse tanto como aceptación como rechazo de la hipótesis nula, consideraremos que estas varianzas son iguales entre los grupos en la variable "saps". De esta manera, podemos realizar el test ANOVA, el cual suele ser más recomendado usar en estos casos.
Realizamos el test ANOVA teniendo como hipótesis nula que no hay diferencias significativas entre las medias de todos los grupos y como hipótesis alternativa que existen diferencias significativas entre las medias de dos grupos (aunque no se indique qué dos grupos):

```
anova_saps_mortdos <- aov(Muns700$saps ~ Muns700$mortdos)
summary(anova_saps_mortdos)
```

```
##                  Df Sum Sq Mean Sq F value Pr(>F)
## Muns700$mortdos   2   3568  1783.8    96.1 <2e-16 ***
## Residuals       692  12846    18.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 5 observations deleted due to missingness
```

Como el p valor del test ANOVA es menor a 0.05 podemos confirmar que sí hay diferencias entre la variable "saps" y alguno de los tres grupos de la variable "mortdos". Para saber en qué pareja de grupos de la variable "mortdos" existen diferencias, utilizaremos el test T apareado con la corrección de bonferroni:

```
pairwise.t.test(Muns700$saps, Muns700$mortdos, p.adj = "bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  Muns700$saps and Muns700$mortdos
##
##             Vivo   Fallecido
## Fallecido   <2e-16 -
## Desconocido 0.83   <2e-16
##
## P value adjustment method: bonferroni
```

Como el p valor entre el grupo fallecido y el grupo desconocido y el p valor entre el grupo vivo y el grupo fallecido son menores a 0.05, podemos decir que hay diferencias entre estos grupos y sus valores de saps. En cambio, entre el grupo vivo y el grupo desconocido al ser el p valor mayor a 0.05 (en concreto, 0.83), no se

observan diferencias significativas entre estos dos grupos. Por lo tanto, concluimos que existen diferencias significativas entre SAPS y la mortalidad entre los grupos fallecido y desconocido y entre los grupos de vivos y fallecidos.

Diferencias entre días de estancia y mortalidad

Para buscar diferencias entre la variable "dias" y la variable "mortdos" al no seguir la variable "dias" una distribución normal, podemos realizar los mismos tests previamente hechos con la variable "saps". Por lo tanto, empezamos realizando el test Kruskal-Wallis para ver si hay diferencias entre los tres grupos de la variable "mortdos" y la variable "dias":

```
kruskal.test(Muns700$dias ~ Muns700$mortdos)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Muns700$dias by Muns700$mortdos
## Kruskal-Wallis chi-squared = 15.055, df = 2, p-value = 0.0005381
```

Al ser el p valor menor a 0.05, aceptamos la hipótesis alternativa, indicando que sí hay diferencias significativas entre estas variables. Para saber entre que par de grupos existen estas diferencias significativas, realizamos el test de Dunn:

```
library(dunn.test)
dunn.test(Muns700$dias, Muns700$mortdos, kw = TRUE, method = "bonferroni")
```

```
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 15.0551, df = 2, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   Desconoc   Fallecid
## ---------+----------------------
## Fallecid |   2.076213
##          |     0.0568
##          |
##     Vivo |  -0.763352  -3.880074
##          |     0.6679    0.0002*
##
## alpha = 0.05
## Reject Ho if p <= alpha/2
```

En los resultados observamos que el grupo que tiene diferencias significativas en el número de días es entre el grupo de vivos y el grupo de fallecidos. Los otros pares de grupos al tener un p-valor mayor a 0.05 indica que no tienen diferencias significativas en el número de días (entre el grupo de vivos y desconocidos y entre el grupo de fallecidos y desconocidos). Por lo tanto concluímos que sólo existen diferencias significativas entre el número de días de estancia en Uci y la mortalidad entre los grupos de vivos y fallecidos.

c) Es cierto que existe una relación entre el nivel de estudios y la mortalidad? Indica los contrastes a realizar y comenta los resultados de su aplicación.

Para encontrar la relación entre la variable categórica con 4 niveles "educacio" y la variable categórica con 3 niveles "mortdos" podemos realizar un test de correlación, ya que estamos buscando la relación entre dos

variables.
Para poder observar mejor estos valores podemos convertir la variable "educacio" en un factor (de la misma manera que hemos hecho con "mortdos" en el apartado b) para ver más fácilmente los resultados:

```
Muns700$educacio <- as.factor(Muns700$educacio)
levels(Muns700$educacio) = c("Ninguno", "1 a 7", "8 a 12", "12", "Desconocido")
```

Ahora que tenemos las dos variables con los grupos correctamente nombrados, podemos realizar una tabla de contingencia para hacernos una idea de la distribución de los datos.

```
table(Muns700$educacio, Muns700$mortdos)
```

```
##
##               Vivo Fallecido Desconocido
##   Ninguno       96        45          20
##   1 a 7        223        68          47
##   8 a 12       103        19          13
##   12            28         7           5
##   Desconocido   10        15           1
```

Con esta tabla en principio parece que los pacientes vivos tienen principalmente estudios de 1 a 7 años y en menor medida de 8 a 12 años. Por lo que respecta a los pacientes fallecidos parece que también la mayoría de pacientes tienen de 1 a 7 años de estudios seguido por ningún tipo de estudios.

Para analizar si hay una relación entre el nivel de estudios y mortalidad, primero debemos saber si estamos tratando con muestras paramétricas o no paramétricas. Para ello, analizaremos la normalidad del nivel de estudios con un test de shapiro-Wilk. Marcamos los valores como numéricos para poder realizar el test de normalidad utilizando sus niveles, mencionados en el enunciado de la PEC.

```
shapiro.test(as.numeric(Muns700$educacio))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  as.numeric(Muns700$educacio)
## W = 0.84014, p-value < 2.2e-16
```

Como la variable "educacio" (nivel de estudios) no se comporta como una normal (p-valor menor a 0.05 en el test de normalidad), deberemos usar tests que traten muestras no paramétricas.

Para estudios de relación con muestras no paramétricas realizamos el test de correlación de Spearman, donde la hipótesis nula será que no hay correlación entre las variables y la hipótesis alternativa que sí hay correlación entre las variables. Marcamos los valores como numéricos para poder realizar la correlación mediante sus niveles, mencionados en el enunciado de la PEC.

```
cor.test(as.numeric(Muns700$educacio), as.numeric(Muns700$mortdos), method = "spearman")
```

```
## Warning in cor.test.default(as.numeric(Muns700$educacio),
## as.numeric(Muns700$mortdos), : Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  as.numeric(Muns700$educacio) and as.numeric(Muns700$mortdos)
## S = 60700372, p-value = 0.1022
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##         rho
## -0.06181626
```

El p-valor superior a 0.05 indica que aceptamos la hipótesis nula, por lo que afirmamos que no existe una correlación entre el nivel de estudios y la mortalidad.

d) Una de las causas de la relación entre la gravedad y la mortalidad es que la gravedad es diferente según el nivel de estudios. Comprueba si esta afirmación es cierta, es decir, si hay diferencias entre las variables de gravedad (TISS y SAPS por separado) y el nivel de estudios. Indica los contrastes de hipótesis a realizar y comenta los resultados.

Relación entre TISS y nivel de estudios

Hemos visto anteriormente que la variable TISS no tiene una distribución normal, por lo que deberemos usar un test no paramétrico. Como debemos comparar entre la variable "tiss" y una variable cualitativa con 5 grupos, podemos buscar la diferencia entre las medias de TISS y cada grupo de la variable estudios. Por lo tanto, realizamos un test no paramétrico para analizar la diferencia de medias, por lo que utilizaremos el test de Kruskal-Wallis. Tendremos como hipótesis nula que no hay diferencias entre distintos grupos de estudios y la variable TISS y la hipótesis alternativa que sí hay diferencias entre grupos.
Realizamos el test entre la variable "tiss" y la variable "educacio":

```
kruskal.test(Muns700$tiss ~ Muns700$educacio)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Muns700$tiss by Muns700$educacio
## Kruskal-Wallis chi-squared = 21.029, df = 4, p-value = 0.0003125
```

Como el p-valor del test es menor a 0.05, podemos concluir que hay diferencias signficativas entre los grupos de la variable "educacio" en la variable "tiss", es decir, que algún par de grupos tiene diferencias significativas entre ellos. Para saber que par de grupos tienen diferencias significativas entre sí, realizaremos el test de Dunn:

```
dunn.test(Muns700$tiss, Muns700$educacio, kw = TRUE, method = "bonferroni")
```

```
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 21.0287, df = 4, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |      1 a 7         12     8 a 12   Desconoc
## ---------+-------------------------------------------
##       12 |  -1.114787
##          |     1.0000
##          |
##   8 a 12 |   0.115738   1.102425
##          |     1.0000     1.0000
##          |
## Desconoc |  -4.382156  -2.842221  -4.233307
##          |    0.0001*    0.0224*    0.0001*
##          |
##  Ninguno |  -1.214349   0.396066  -1.101715   3.686348
##          |     1.0000     1.0000     1.0000    0.0011*
##
## alpha = 0.05
```

```
## Reject Ho if p <= alpha/2
```

Con los resultados obtenidos en el test de Dunn podemos observar que hay dos pares de grupos con diferencias significativas entre ellos, ya que son los únicos que tienen un p valor menor a 0.05 (aceptando la hipótesis alternativa del test). En concreto, estos dos pares de grupos son entre desconocido nivel de estudios y estudios de 8 a 12 años y entre ningún nivel de estudios y el nivel de estudios desconocido. Como ambos pares de grupos con diferencias significativas tienen en cuenta el grupo desconocido, podemos suponer que estos resultados no son significativos, ya que la variable desconocidos no aporta en principio ningún valor a la base de datos (no nos indica si el paciente tiene o no estudios). Por lo tanto podríamos decir que, descontando el grupo de pacientes con estudios desconocidos, no existen diferencias significativas entre la variable "tiss" y el nivel de estudios de los pacientes.

Relación entre SAPS y nivel de estudios

Como consideramos que la variable cuantitativa SAPS sigue una distribución normal, podemos usar tests paramétricos para buscar si hay diferencias entre la variable SAPs y los distintos niveles de estudios de la variable "educacio". Para buscar estas diferencias, el contraste de hipótesis que podemos realizar es como hipótesis nula que no hay diferencias significativas entre la variable SAPS y los distintos niveles de estudios, y la hipótesis alternativa que sí que hay diferencias significativas entre estos. Para ello podemos realizar un test ANOVA:

```r
anova_saps_estudios <- aov(Muns700$saps ~ Muns700$educacio)
summary(anova_saps_estudios)
```

```
##                    Df Sum Sq Mean Sq F value   Pr(>F)
## Muns700$educacio    4    973  243.33   10.87 1.49e-08 ***
## Residuals         690  15440   22.38
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 5 observations deleted due to missingness
```

Con un p-valor menor a 0.05 también podemos confirmar que hay diferencias significativas entre la variable "saps" y los distintos grupos de nivel de estudios. Para saber en que par de grupos existen diferencias significativas teniendo en cuenta que usamos tests paramétricos, podemos utilizar el test t apareado con corrección de bonferroni:

```r
pairwise.t.test(Muns700$saps, Muns700$educacio, p.adj = "bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  Muns700$saps and Muns700$educacio
##
##             Ninguno 1 a 7   8 a 12  12
## 1 a 7       0.01113 -       -       -
## 8 a 12      0.00011 0.45250 -       -
## 12          0.25987 1.00000 1.00000 -
## Desconocido 0.01720 2.1e-05 5.5e-07 0.00030
##
## P value adjustment method: bonferroni
```

Observando los p valores entre los distintos grupos de nivel de estudio según la variable SAPS, podemos decir que hay diferencias significativas entre no tener estudios y entre 1 a 7 años, 8 a 12 y estudios desconocidos. También hay diferencias significativas entre niveles de estudios desconocidos y los otros cuatro niveles de estudios. Como hemos comentado en el subapartado anterior (relación entre tiss y nivel de estudios), el grupo de nivel de estudios desconocidos marca diferencias significativas con todos los grupos pero este grupo no da mucha información al respecto. Por lo tanto podríamos descartar las diferencias significativas con este grupo.

Entonces, podemos concluir que existen diferencias significativas entre ningún tipo de estudios y estudios de 1 a 7 años y entre ningún tipo de estudios y estudios de 8 a 12 años.

Algunas cosas más Ejercicio 3

a) Siguiendo las instrucciones debajo vamos a generar 1000 veces tres muestras de tamaño 80 bajo la distribución normal con la misma media y varianza. Para cada conjunto de datos comparamos la igualdad de medias entre la muestra 1 y 2, la muestra 1 y 3 y la muestra 2 y 3 con un error tipo I de ($\alpha = 0.05$ cada uno) y guarda si se acepta o no la hipótesis nula.

```
set.seed(123)
m <- 20
s <- 3
comp1.2 <- NULL
comp2.3 <- NULL
comp1.3 <- NULL
for(i in 1:1000) {
  m1 <- rnorm(80, m, s)
  m2 <- rnorm(80, m, s)
  m3 <- rnorm(80, m, s)
  IC1.2 <- c(t.test(m1,m2)$conf.int[1],t.test(m1,m2)$conf.int[2])
  comp1.2[i] <- (IC1.2[1]<0 & IC1.2[2]>0)
  IC2.3 <- c(t.test(m2,m3)$conf.int[1],t.test(m2,m3)$conf.int[2])
  comp2.3[i] <- (IC2.3[1]<0 & IC2.3[2]>0)
  IC1.3 <- c(t.test(m1,m3)$conf.int[1],t.test(m1,m3)$conf.int[2])
  comp1.3[i] <- (IC1.3[1]<0 & IC1.3[2]>0)
}
```

```
IC1.2
```

```
## [1] -1.0900595  0.8662538
```

A) Comparaciones múltiples

a1) ¿Qué porcentaje de veces no rechazaríamos cada una de las tres hipótesis. Comprobad el resultado en vuestra muestra

En cada comparación tenemos 1000 valores que pueden ser TRUE si se acepta la hipótesis nula y FALSE si se acepta la hipótesis alternativa (se rechaza la hipótesis nula). Para encontrar el porcentaje de veces que no rechazamos la hipótesis nula, podemos hacer una proporción de la tabla TRUE y FALSE para cada comparación:

```
prop.table(table(comp1.2))
```

```
## comp1.2
## FALSE  TRUE
## 0.046 0.954
```

El 95.4% de las veces que se compara la muestra 1 con la 2 la hipótesis nula no se rechaza.

```
prop.table(table(comp2.3))
```

```
## comp2.3
## FALSE  TRUE
## 0.046 0.954
```

Entre la muestra 2 y 3 también se acepta la hipótesis nula un 95.4% de las veces.

```
prop.table(table(comp1.3))
```

```
## comp1.3
## FALSE  TRUE
## 0.035 0.965
```

Entre la muestra 1 y 3 no se rechaza la hipótesis nula un 96.5% de las veces.

a2) ¿Qué % de las veces no rechazamos ninguna de las tres hipótesis de que las tres medias son iguales?

Al estar hablando de 3 grupos (muestras) paramétricos (independientes, que siguen distribución normal y tienen misma varianza), podemos realizar un test ANOVA para observar en cada una de las 1000 muestras si se acepta la hipótesis nula (sus medias son iguales). Para ello realizamos otro loop con la misma seed: Como tenemos las probabilidades de las comparaciones con los tests T de las tres hipótesis nulas, multiplicando las probabilidades de estas comparaciones obtendremos la probabilidad de no rechazar la hipótesis nula en las tres comparaciones. Por lo tanto, con una multiplicación podemos resolver el apartado:

```
tres_hipotesis <- 0.954 * 0.954 * 0.965
tres_hipotesis
```

```
## [1] 0.8782619
```

Obtenemos entonces que no rechazamos ninguna de las tres hipótesis de que las tres medias sean iguales un 87.8% de las veces.

a3) Explicar qué es la corrección de Bonferroni y cuál sería en este caso. Aplícalo a las muestras que habías obtenido anteriormente y vuelve a calcular el % de no rechazo de las tres hipótesis nulas: las tres medias son iguales. Comentar los resultados.

La corrección de Bonferroni es un método estadístico que permite corregir la acumulación del Error de Tipo I en tests estadísticos en los que se realiza una comparación de múltiples variables o grupos. En este caso se utilizaría la corrección de Bonferroni para corregir el resultado de la ANOVA realizada para comparar las medias entre tres grupos. Al utilizar tres grupos acumulamos la probabilidad de aceptar la hipótesis alternativa: α=1-P(aceptar cada una de las tres hipótesis) = 1 - 0.95 * 0.95 * 0.95 = 0.14, habiéndose acumulado consecuentemente la probabilidad de aceptar la hipótesis alternativa inicial de 0.05. Para calcular el % de veces que no se rechazan las tres hipótesis nulas con el nuevo valor de α, debemos realizar de nuevo los test T con el nuevo nivel de confianza de 0.14:

```
set.seed(123)
comp1.2_corrected <- NULL
comp2.3_corrected <- NULL
comp1.3_corrected <- NULL
for(i in 1:1000) {
  m1 <- rnorm(80, m, s)
  m2 <- rnorm(80, m, s)
  m3 <- rnorm(80, m, s)
  IC1.2 <- c(t.test(m1,m2, conf.level = 0.14)$conf.int[1],t.test(m1,m2)$conf.int[2])
  comp1.2_corrected[i] <- (IC1.2[1]<0 & IC1.2[2]>0)
  IC2.3 <- c(t.test(m2,m3, conf.level = 0.14)$conf.int[1],t.test(m2,m3)$conf.int[2])
  comp2.3_corrected[i] <- (IC2.3[1]<0 & IC2.3[2]>0)
  IC1.3 <- c(t.test(m1,m3, conf.level = 0.14)$conf.int[1],t.test(m1,m3)$conf.int[2])
  comp1.3_corrected[i] <- (IC1.3[1]<0 & IC1.3[2]>0)
}
```

Con el nuevo nivel de confianza calculamos los porcentajes entre las distintas comparaciones de nuevo:

```r
prop.table(table(comp1.2_corrected))
```

```
## comp1.2_corrected
## FALSE  TRUE
## 0.471 0.529
```

El 52.9% de las veces que se compara la muestra 1 con la 2 la hipótesis nula no se rechaza.

```r
prop.table(table(comp2.3_corrected))
```

```
## comp2.3_corrected
## FALSE  TRUE
## 0.454 0.546
```

Entre la muestra 2 y 3 también se acepta la hipótesis nula un 54.6% de las veces.

```r
prop.table(table(comp1.3_corrected))
```

```
## comp1.3_corrected
## FALSE  TRUE
##  0.46  0.54
```

Entre la muestra 1 y 3 no se rechaza la hipótesis nula un 54% de las veces.
Por lo tanto, realizamos ahora la multiplicación con los nuevos porcentajes:

```r
tres_hipotesis_corrected <- 0.529 * 0.546 * 0.540
tres_hipotesis_corrected
```

```
## [1] 0.1559704
```

Por lo tanto podemos concluir que el porcentaje de no rechazo de las tres hipótesis nulas corregidas con Bonferroni (es decir, con nivel de confianza = 0.14) es el 15.6% de las veces.

B) Bootstrap. Responder a las preguntas de forma breve y razonada

b1) ¿La estimación bootstrap sirve para estimar la distribución de un estadístico o bien para estimar la distribución teórica de la población?

La estimación Bootstrap sirve principalmente para estimar de manera eficiente y con el menor sesgo posible la distribución de una muestra. Con este método no podemos estimar la distribución teórica de la población porque estamos haciendo simulaciones de las muestras obtenidas, por lo que obtendremos la distribución estimada de las muestras, no la distribución teórica de la población.

b2) Indica brevemente como procederías para obtener la estimación bootstrap de la mediana de la variable TISS. Calcula el estimador bootstrap de la mediana y su intervalo de confianza al 95%. (Nota: En uno de los capítulos recomendados indica como realizarlo).

Para realizar el bootstrap de la variable TISS, haría un sample de x observaciones de la variable guardando la mediana de cada sample. Antes de esto será mejor eliminar los valores perdidos (NA) para evitar posteriores errores. Después repetiría el sample repetidas veces (10000 por ejemplo) con substitución de la observación en cada elección del sample, para simular una elevada cantidad de medianas:

```r
set.seed(123)
muestras <- na.omit(Muns700$tiss)
n <- length(muestras)
repeticiones <- 10000
resamples <- matrix(sample(muestras, n*repeticiones, replace = TRUE),repeticiones,n)
medianas <- apply(resamples,1, median)
```

Ahora calculando el percentil 2.5 y el percentil 97.5 obtenemos el intervalo de confianza de las medianas al 95%:

```
quantile(medianas,c(0.025,0.975))
```

```
##  2.5% 97.5%
##    15    17
```

Nos encontramos entonces que el intervalo de medianas que se encuentra dentro del intervalo de confianza de 95% es entre la mediana de valor 15 y la mediana de valor 17.