Inferencia estadística 2020-21 semestre 1: PEC2

Lucas Goiriz Beltrán

24 de Diciembre 2020

1 Enunciados teóricos

Indica que tipo de análisis o que pruebas estadísticas utilizarías en cada uno de los apartados y si fuera necesario que algún tipo de prueba adicional harías para llevar a cabo el análisis. Formula la hipótesis a contrastar de acuerdo con las preguntas a responder. Justifica la elección. (La solución puede no ser única pero la escogida debe de estar justificada).

1.1 Se lleva a cabo un estudio para comparar la duración de la estancia hospitalaria de pacientes ingresados con el mismo diagnóstico, en dos hospitales A y B que utilizan dos protocolos de gestión diferentes. En lo dos hospitales A y B se observan los siguientes resultados en estancia para dos grupos de pacientes

- Hospital A — 22, 12, 18, 19, 36, 32, 14, 18, 20, 22
- Hospital B — 26, 32, 33, 38, 24, 27, 19, 29, 30, 26, 24, 16, 28

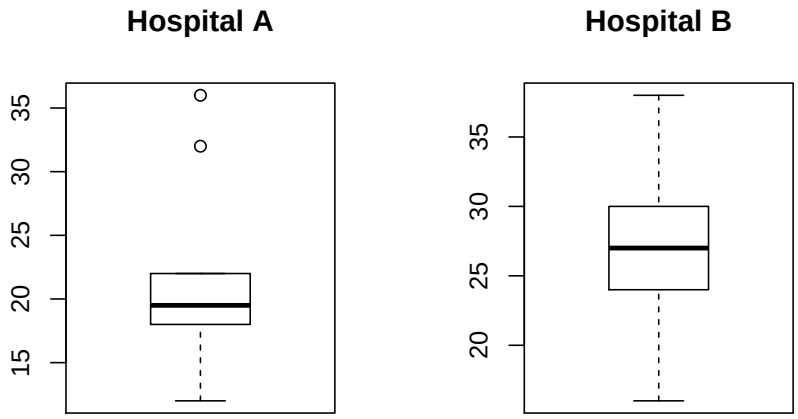
¿Hay diferencias en la duración de la estancia entre los 2 hospitales?

Resolución: La variable a estudiar es de respuesta cuantitativa y supondremos que discreta (la duración de la estancia hospitalaria se mediría en intervalos de 1 día) con dos grupos independientes a ser comparados (los pacientes del hospital A y los pacientes del hospital B).

La hipótesis nula (H_0): la duración de la estancia hospitalaria es la misma en ambos hospitales.
La hipótesis alternativa (H_1): la duración de la estancia hospitalaria no es la misma en ambos hospitales.

El número de muestras en cada grupo es menor que 30. Desconocemos si los datos son normales. Vamos a dibujar primero los boxplot a ver si los datos son simétricos:

```
A <- data.frame(c(22, 12, 18, 19, 36, 32, 14, 18, 20, 22))
names(A) <- c('A')
B <- data.frame(c(26, 32, 33, 38, 24, 27, 19, 29, 30, 26, 24, 16, 28))
names(B) <- c('B')

par(mfrow = c(1, 2))
boxplot(A$A, main = 'Hospital A')
boxplot(B$B, main = 'Hospital B')
```



Podemos ver que mientras los datos del Hospital B tienen aparentemente simetría, los del Hospital A no la tienen. Vamos a hacer una prueba de normalidad, que va a ser el test de Shapiro-Wilks.

```
shapiro.test(A$A)

##
##  Shapiro-Wilk normality test
##
## data:  A$A
## W = 0.88849, p-value = 0.1631

shapiro.test(B$B)

##
##  Shapiro-Wilk normality test
##
## data:  B$B
## W = 0.98134, p-value = 0.9854
```

Para ambos hospitales, el p-valor del test de Shapiro-Wilks es mayor que el nivel de significancia $\alpha = 0.05$, por lo tanto, pese a que en el boxplot hayamos visto asimetría, mediante el criterio del test de Shapiro-Wilks vamos a asumir que los datos son normales. A continuación vamos a comprobar si ambos grupos tienen varianzas iguales, mediante el test de Levene.

```r
library(car)
leveneTest(
    c(A$A, B$B),
    as.factor(c(
        rep(1, length(A$A)),
        rep(2, length(B$B))
        ))
    )
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1  0.2035 0.6565
##       21
```

El p-valor del test de Levene es mayor que el nivel de significancia $\alpha = 0.05$, por lo tanto, mediante el criterio del test de Levene vamos a asumir que los grupos tienen varianzas iguales.

Dadas las suposiciones, tomamos la media como un buen estimador. Esto nos permite considerar una prueba paramétrica de comparación de igualdad de medias (dadas nuestras suposiciones y datos, va a ser un t-test), permitiéndonos a su vez reformular nuestras hipótesis de la siguiente manera:

La hipótesis nula (H_0): $\mu_A - \mu_B = 0$
La hipótesis alternativa (H_1): $\mu_A - \mu_B \neq 0$

Llevamos a cabo entonces el t-test:

```r
t.test(A$A, B$B, mu=0, var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  A$A and B$B
## t = -2.0967, df = 21, p-value = 0.0483
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.50664113  -0.04720503
## sample estimates:
## mean of x mean of y
##  21.30000  27.07692
```

El p-valor del t-test es menor que el nivel de significancia $\alpha = 0.05$, por lo tanto, mediante el criterio del t-test vamos a rechazar la Hipótesis nula, y por ende aceptar la Hipótesis alternativa, es decir, podemos decir que existen diferencias en la duración de la estancia entre ambos hospitales.

Nota: *si no hubiéramos querido asumir igualdad de varianzas, habríamos llevado a cabo un t-test para varianzas distintas (manteniéndo el contraste de hipótesis de diferencia de medias). Si además no hubiéramos querido asumir que los datos son normales, habríamos empleado el test no paramétrico U-Mann-Whitney para poder contrastar la hipótesis de misma distribución de duración hospitalaria en ambos hospitales.*

1.2 Para determinar si la refracción esférica ocular es la misma en ambos ojos de los individuos, se midió ésta (en dioptrías) en el ojo derecho y en el izquierdo de 15 personas. Se quiere saber si existen diferencias significativas en la cantidad de dioptrías entre ambos ojos.

Resolución: la variable a estudiar es de respuesta cuantitativa contínua (dioptrías oculares). En este escenario disponemos de 2 grupos emparejados (ojo izquierdo y ojo derecho de cada paciente) con 15 replicas cada grupo (ya que son 15 pacientes).

La hipótesis nula (H_0): no hay diferencias significativas en la cantidad de dioptrías de ambos ojos.
La hipótesis alternativa (H_1): hay diferencias significativas en la cantidad de dioptrías de ambos ojos.

Desconocemos si los datos son normales, además el número de muestras de cada grupo es menor que 30. De todas formas podríamos comprobar la simetría mediante un boxplot y podríamos hacer una prueba de normalidad como el test de Shapiro-Wilks. Si el resultado fuese que los datos son normales, tomaríamos la media como buen estimador y reescribiríamos las hipótesis de la siguiente manera:

La hipótesis nula (H_0): la media de las diferencias de dioptrías de los ojos de cada paciente es 0.
La hipótesis alternativa (H_1): la media de las diferencias de dioptrías de los ojos de cada paciente es distinto de 0.

Finalmente procederíamos a hacer un t-test apareado para contrastar la hipótesis de media de diferencias nula.

En el caso de que los datos no fuesen normales, tendríamos que optar por una prueba no paramétrica. Reescribiríamos entonces las hipótesis de la siguiente manera:

La hipótesis nula (H_0): no existen diferencias en la distribución de dioptrías para ambos ojos.

La hipótesis alternativa (H_1): existen diferencias en la distribución de dioptrías para ambos ojos.

Para contrastar la hipótesis de igualdad de distribuciones para ambos ojos procederíamos a hacer un test Signo-rango wilcoxon.

1.3 Algunos trabajos han descrito una relación inversa entre el consumo moderado de alcohol y los niveles de colesterol. Para comprobar esta hipótesis, se aplica un cuestionario sobre consumo de alcohol a un grupo de 520 hombres, trabajadores de un polígono industrial. Se obtienen muestras sanguíneas para determinar sus niveles de colesterol. Los sujetos estudiados son clasificados en tres grupos, de acuerdo con el consumo de alcohol manifestado, a saber, "bajo", "moderado" y "alto". Se desea averiguar si existen diferencias entre los niveles de colesterol de los grupos estudiados, especificando entre qué grupos se producen las diferencias si es que las hay (Nota: El menor grupo supera los 40 sujetos)

Resolución: la variable a estudiar es de respuesta cuantitativa contínua (niveles de colesterol en sangre). Con 3 grupos independientes (personas con consumo bajo, medio y alto de alcohol).

La hipótesis nula (H_0): los niveles de colesterol entre grupos no tienen diferencias significativas.
La hipótesis alternativa (H_1): los niveles de colesterol entre grupos tienen diferencias significativas.

El tamaño de muestra de cada grupo es grande (mayor que 40 para el grupo con menos integrantes), mediante el teorema central del límite, podemos emplear la media de los niveles de colesterol como un buen descriptor. También comprobaremos la simetría de los datos mediante un boxplot además de realizar una prueba de normalidad (usaremos el test de Shapiro-Wilks). A continuación comprobamos la igualdad de varianzas mediante el test de Levene. Si dichas pruebas salen favorables, podemos reformular las hipótesis de la siguiente manera:

La hipótesis nula (H_0): las medias de los niveles de colesterol de cada grupo son iguales.
La hipótesis alternativa (H_1): laz media de los niveles de colesterol de cada grupo son iguales.

Emplearíamos un ANOVA para contrastar la hipótesis de igualdad de medias. El ANOVA nos indicaría si al menos dos de las medias comparadas son significativamente distintas entre sí. Finalmente, si fuese el caso, tendríamos que comparar dos a dos las medias de todos los grupos mediante un t-test para identificar qué grupos tienen medias significativamente distintas, rechazando por lo tanto la hipótesis nula.

Sin embargo si el test de Levene nos saliera desfavorable (por lo tanto no tenemos varianzas iguales) o el test de Shapiro-Wilks nos diera que los datos no siguen la distribución normal, reformularíamos las hipótesis de la siguiente manera:
La hipótesis nula (H_0): las distribuciones de los niveles de colesterol entre grupos no tienen diferencias significativas.
La hipótesis alternativa (H_1): las distribuciones de los niveles de colesterol entre grupos tienen diferencias significativas.

Para contrastar la hipótesis de igualdad de distribuciones emplearíamos el test de Kruskal-Wallis. De nuevo, este test nos indicaría si al menos dos de las distribuciones comparadas son significativamente distintas entre sí. Una vez más, si fuese el caso, tendríamos que comparar dos a dos las distribuciones de todos los grupos mediante un el test U-Mann-Whitney para identificar qué grupos tienen distribuciones significativamente distintas, rechazando por lo tanto la hipótesis nula.

1.4 En el contexto de la pandemia de la COVID-19, numerosos test serológicos han sido propuesto para la detección de anticuerpos Anti-SARS-COV_2. En el laboratorio central del hospital han llegado dos nuevos test de dos empresas biotecnológicas, la A y la B, con las mismas características de sensibilidad y especificidad. El personal del laboratorio central quiere comprobar si ambas pruebas coinciden en los resultados. Para ello quieren efectuar un contraste de hipótesis aplicando los dos test a un grupo de 1200 muestras de suero disponibles. Se desea contrastar si ambos test de antígenos coinciden en el diagnóstico.

Resolución: la variable a estudiar es de respuesta cualitativa dicotómica (toma los valores positivo o negativo). En este escenario disponemos de 2 grupos emparejados (prueba fabricada por la empresa A y prueba fabricada por la empresa B, aplicadas ambas sobre cada mustra de suero) con 1200 replicas cada grupo (ya que son 1200 muestras de suero).

La hipótesis nula (H_0): las proporciones de positivos y negativos para ambos

tests no tienen diferencias significativas
La hipótesis alternativa (H_1): las proporciones de positivos y negativos para ambos tests tienen diferencias significativas.

Construiremos entonces una tabla de proporciones empleando los resultados de los tests para cada muestra de suero, construyendo así una tabla parecida a la siguiente:

		Test A		
		Positivo	**Negativo**	**Total fila**
Test B	**Positivo**	$p_a = \frac{a}{1200}$	$p_b = \frac{b}{1200}$	$p_{a+b} = \frac{a+b}{1200}$
	Negativo	$p_c = \frac{c}{1200}$	$p_d = \frac{d}{1200}$	$p_{c+d} = \frac{c+d}{1200}$
	Total columna	$p_{a+c} = \frac{a+c}{1200}$	$p_{b+d} = \frac{b+d}{1200}$	1

Con esta tabla, podemos reformular las hipótesis planteadas anteriormente. "*H_0:Las proporciones de positivos y negativos para ambos tests no tienen diferencias significativas*", si miramos la tabla, se correspondería a $p_{a+b} = p_{a+c}$ y $p_{c+d} = p_{b+d}$. Mirando la tabla, podemos simplificar estas expresiones, ya que $p_{i+j} = p_i + p_j$, quedándonos con $p_a + p_b = p_a + p_c$ y $p_c + p_d = p_b + p_d$. Es sencillo ver que ambas expresiones se simplifican en una: $p_b = p_c$, es decir que la proporción de muestras que dan positivo con el test B y negativo con el test A es igual a la proporción de muestras que dan negativo con el test B y positivo con el test A. Por lo tanto, reescribiremos nuestras hipótesis de la siguiente forma:

La hipótesis nula (H_0): $p_c = p_b$.
La hipótesis alternativa (H_1): $p_c \neq p_b$.

Para contrastar nuestra hipótesis, llevamos a cabo el test de McNemar. Si este resulta favorable (el p-valor del test es superior al nivel de significancia $\alpha = 0.05$), no tendríamos evidencias suficientes para rechazar la hipótesis nula, por lo tanto la aceptaríamos.

1.5 Se plantea un estudio para examinar los efectos exógenos de la Interlukina-33(IL-33) en las característica biológicas del Carcinoma Hepatocellular(HCC). Se clasifica el nivel de la expresión de la IL-33 en Bajo (H-score<68 en 34 pacientes) y Alto(H-score≥68 en 45 pacientes.). Se desea conocer si por un lado la edad y por otro el volumen del tumor, que es una variable asimétirca, y el tener o no metástasis se asocian con el nivel de IL-33.

Resolución: la variable expresión de IL-33 es una variable cualitativa dicotómica (que toma los valores Alto y Bajo).En el primer caso, se desea conocer

si la edad se asocia con el nivel de expresión. La edad es una variable cuantitativa y vamos a asumir que es discreta (que se mide en intervalos de 1 año). Podemos agrupar la variable edad en 2 grupos independientes: un grupo de edades de las personas con un nivel de expresión de IL-33 Alto y otro grupo de edades de las personas con un nivel de expresión de IL-33 Bajo.

A continuación planteamos las hipótesis a contrastar:
La hipótesis nula (H_0): las edades entre los dos grupos de nivel de expresión de IL-33 no tienen diferencias significativas.
La hipótesis alternativa (H_1): las edades entre los dos grupos de nivel de expresión de IL-33 tienen diferencias significativas.

Ahora deberiamos evaluar si la distribución de la edad en nuestros grupos sigue la distribución normal. Primero comprobaríamos si las distribuciones son simétricas mediante un boxplot, a continuación haríamos un el test de Shapiro-Wilk para ver si las distribuciones son normales y un test de Levene para comprobar si las varianzas son iguales entre ambos grupos. Si se cumpliesen todas las condiciones (simetría, normalidad e igualdad de varianzas) tomaríamos la media de las edades de cada grupo como un buen descriptor de la edad, lo que nos permitiría reescribir las hipótesis de la siguiente manera:

La hipótesis nula (H_0): las medias de las edades entre los dos grupos de nivel de expresión de IL-33 no tienen diferencias significativas.
La hipótesis alternativa (H_1): las medias de las edades entre los dos grupos de nivel de expresión de IL-33 tienen diferencias significativas.

Para contrastar estas hipótesis emplearíamos un t-test de igualdad de medias.

Si no se hubiesen dado las condiciones de simetría, normalidad e igualdad de varianzas, reescribiríamos las hipótesis de la siguiente manera:

La hipótesis nula (H_0): la distribución de las edades en los dos grupos de nivel de expresión de IL-33 no tienen diferencias significativas.
La hipótesis alternativa (H_1): la distribución de las edades en los dos grupos de nivel de expresión de IL-33 tienen diferencias significativas.

Para contraster estas hipótesis emplearíamos el test no paramétrico U-Mann-Whitney.

Para el segundo caso, se desea conocer si el volumen del tumor se asocia con el nivel de expresión. El volumen del tumor es una variable cuantitativa y vamos a asumir que es contínua (imagino que se medirá en mm^3) además de que se nos indica explícitamente que esta variable es asimétrica (por lo tanto de inicio incumple el criterio de normalidad).Nuevamente podemos agrupar la

variable volumen del tumor en 2 grupos independientes: un grupo de volúmenes de tumor de las personas con un nivel de expresión de IL-33 Alto y otro grupo de volúmenes de tumor de las personas con un nivel de expresión de IL-33 Bajo.

Al no tener criterio de normalidad, hemos de emplear un test no paramétrico. La formulación de las hipótesis es la siguiente:

La hipótesis nula (H_0): la distribución de los volúmenes de los tumores en los dos grupos de nivel de expresión de IL-33 no tienen diferencias significativas.
La hipótesis alternativa (H_1): la distribución de los volúmenes de los tumores en los dos grupos de nivel de expresión de IL-33 tienen diferencias significativas.

Nuevamente, para contrastar estas hipótesis emplearíamos el test de U-Mann-Whitney.

Para el tercer caso, se desea conocer si el hecho de tener metástasis o no tenerla se asocia con el nivel de expresión. Tener o no tener metástasis es una variable cualitativa dicotómica (toma los valores "Sí" y "No"). Como además desconocemos las proporción de individuos con metástasis (por lo tanto desconocemos si se cumplen las condiciones de $np > 5$ y $n(1-p) > 5$ donde n es número de individuos y p es proporción de los grupos), hemos de optar por un test no paramétrico. Para ello, hemos de construir primero una tabla de proporciones similar a la siguiente:

		Expresión IL-33 Alta		
		Sí metástasis	**No metástasis**	**Total fila**
Expresión IL-33 Baja	**Sí metástasis**	p_a	p_b	p_{a+b}
	No metástasis	p_c	p_d	p_{c+d}
	Total columna	p_{a+c}	p_{b+d}	1

Con la ayuda de esta tabla calculamos las proporciones de interés (p_b y p_c) y formularemos nuestras hipótesis:

La hipótesis nula (H_0): La proporción de individuos con expresión baja de IL-33 y sin metástasis no tiene diferencias significativas con la proporción de individuos con expresión alta de IL-33 y con metástasis, es decir que $p_c = p_b$.
La hipótesis alternativa (H_1): La proporción de individuos con expresión baja de IL-33 y sin metástasis tiene diferencias significativas con la proporción de individuos con expresión alta de IL-33 y con metástasis, es decir que $p_c \neq p_b$.

Para contrastar nuestras hipótesis emplearemos el test no paramétrico signo-rango Wilcoxon.

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 con la realidad. 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.

Se dispone de 1092 pacientes mayores de 65 años que fueron ingresados en un hospital por tres patologías(tipocas): síndrome coronario agudo (SCA), accidente cerebrovascular (AVC) y Neumonía(Neumo). Los sujetos fueron asignados a dos grupos: Los casos y controles en función de un tratamiento preventivo recibido (caso). Se midieron algunas variables como el sexo , la edad, antecedentes de cardiopatía y diabetes. Además se disponía de la tensión arterial sistólica(TAS) y el índice de Barthel que es una escala de autonomía del paciente. El fichero lo tenéis disponible en Stata (`data_admision.dta`) o en Excel (`data_admission.xlsx`).

```
# Selecciona 900 casos al azar del fichero para generar la base datos del trabajo
set.seed(101010)
library(foreign)
datos <- as.data.frame(read.dta("data_admission.dta"))
datos900 <- datos[sample(1:nrow(datos), 900, replace=FALSE),]
```

En las siguientes preguntas además de escoger las hipótesis adecuadas, justifica el uso de la prueba o pruebas utilizadas e interpreta los resultados.

2.1 ¿Existe asociación entre el tener una cardiopatía previa y el tipo de enfermedad por la que ingresó el paciente? ¿Y con tener diabetes y el tipo de enfermedad por la que ingresó el paciente?

Empecemos por el par `cardiopatía` y `tipocaso`. La variable `cardiopatía` es una variable cualitativa dicotómica (con valores: "sí" y "no"). Por otra parte la variable `tipocaso` es también cualitativa pero tiene 3 posibles valores ("SCA", "ACV" y "Neumo").

La hipótesis nula (H_0): el hecho de haber tenido previamente una cardiopatía y el tipo de patología del paciente son independientes, es decir, el % de personas que previamente han sufrido una cardiopatía no varía entre los distintos tipos de patología de ingreso.
La hipótesis alternativa (H_1): el hecho de haber tenido previamente una cardiopatía y el tipo de patología del paciente no son independientes, es decir, el % de personas que previamente han sufrido una cardiopatía varía entre los dis-

tintos tipos de patología de ingreso.

Como las variables problema son cualitativas y el número de poblaciones es mayor que 2, nos decantamos por un test no paramétrico: test χ^2. Sin embargo, para este test necesitamos la tabla de conteos y además que no más del 20% de las celdas tengan un valor inferior a 5.

Mediante `R` construiremos la tabla de conteos.

```
cols <- c("SCA", "ACV", "Neumo")
rows <- c("si", "no")

cont_table <- data.frame(
                         SCA = double(),
                         ACV = double(),
                         Neumo = double()
                         )

for (col in c(1, 2, 3)) {
    for (row in c(1, 2)) {
        cont_table[row, col] <-
            nrow(
                 datos900[
                          (datos900$cardiopatia == rows[row]) &
                          (datos900$tipocas == cols[col]),
                          ]
                 )
    }
}
row.names(cont_table) <- paste(rows, 'cardiopatia')
cat("La tabla de conteos es la siguiente:\n")

## La tabla de conteos es la siguiente:

cont_table

##                 SCA ACV Neumo
## si cardiopatia 135  81   115
## no cardiopatia 156 185   228
```

En la tabla podemos ver que todos los conteos son mayores que 5, por lo tanto procedemos sin miedo a hacer la prueba χ^2 de 2 grados de libertad.

```
chisq.test(cont_table)

##
```

```
##  Pearson's Chi-squared test
##
## data:  cont_table
## X-squared = 17.705, df = 2, p-value = 0.000143
```

El p-valor de la prueba χ^2 es inferior a $\alpha = 0.05$, por lo tanto rechazamos la hipótesis nula y aceptamos la hipótesis alternativa, por ende el hecho de haber tenido previamente una cardiopatía influye en el tipo de patología del paciente ingresado. Para estudiar qué grupos en concreto tienen diferencias significativas, vamos a realizar tests χ^2 (con corrección del nivel de significania α) para cada par de columnas de nuestra tabla de conteos.

```
iters <- list()
iters[[1]] <- c(1,2)
iters[[2]] <- c(1,3)
iters[[3]] <- c(2,3)

for (i in iters){
    t <- cont_table[, i]
    n <- names(t)
    cat("Pareja de columnas: ", n[1], "_", n[2])
    print(chisq.test(t))
    cat(rep("#", 63), "\n", sep="")
}
```

```
## Pareja de columnas:  SCA _ ACV
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  t
## X-squared = 14.21, df = 1, p-value = 0.0001635
##
## ###############################################################
## Pareja de columnas:  SCA _ Neumo
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  t
## X-squared = 10.376, df = 1, p-value = 0.001276
##
## ###############################################################
## Pareja de columnas:  ACV _ Neumo
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  t
## X-squared = 0.5164, df = 1, p-value = 0.4724
##
## ###############################################################
```

Podemos ver que el p-valor es inferior a $\alpha = 0.05$ para las parejas formadas con la variable `SCA` mientras que la pareja de variables `ACV` y `Neumo` el p-valor es superior a $\alpha = 0.05$. Esto quiere decir que haber tenido o no cardiopatías previas no influye para los ingresados por neumonía y por accidente cardiovascular, sin embargo sí que influye para aquellos incresados por síndrome coronario agudo.

Para el par `tipocaso` y `diabetes` tenemos una situación similar. La variable `diabetes` es, como la variable `cardiopatía`, cualitativa dicotómica (tiene dos valores: "sí" y "no").

La hipótesis nula (H_0): el hecho de tener diabetes y el tipo de patología del paciente son independientes, es decir, el % de personas que tienen diabetes no varía entre los distintos tipos de patología de ingreso.
La hipótesis alternativa (H_1): el hecho de tener diabetes y el tipo de patología del paciente no son independientes, es decir, el % de personas que que tienen diabetes varía entre los distintos tipos de patología de ingreso.

De nuevo, como las variables problema son cualitativas y el número de poblaciones es mayor que 2, nos decantamos por un test no paramétrico: test χ^2. Sin embargo, para este test necesitamos la tabla de conteos y además que no más del 20% de las celdas tengan un valor inferior a 5. Mediante `R` construiremos la tabla de conteos.

```
cols = c("SCA", "ACV", "Neumo")
rows = c("no", "si")

cont_table <- data.frame(
                          SCA = double(),
                          ACV = double(),
                          Neumo = double()
                          )

for (col in c(1, 2, 3)) {
    for (row in c(1, 2)) {
        cont_table[row, col] <-
            nrow(
                 datos900[
                          (datos900$diabetes == rows[row]) &
                          (datos900$tipocas == cols[col]),
                          ]
                 )
    }
}
row.names(cont_table) <- paste(rows, 'diabetes')
cat("La tabla de conteos es la siguiente:\n")
```

```
## La tabla de conteos es la siguiente:

cont_table

##             SCA ACV Neumo
## no diabetes 209 197   261
## si diabetes  82  69    82
```

En la tabla podemos ver que todos los conteos son mayores que 5, por lo tanto procedemos sin miedo a hacer la prueba χ^2 de 2 grados de libertad.

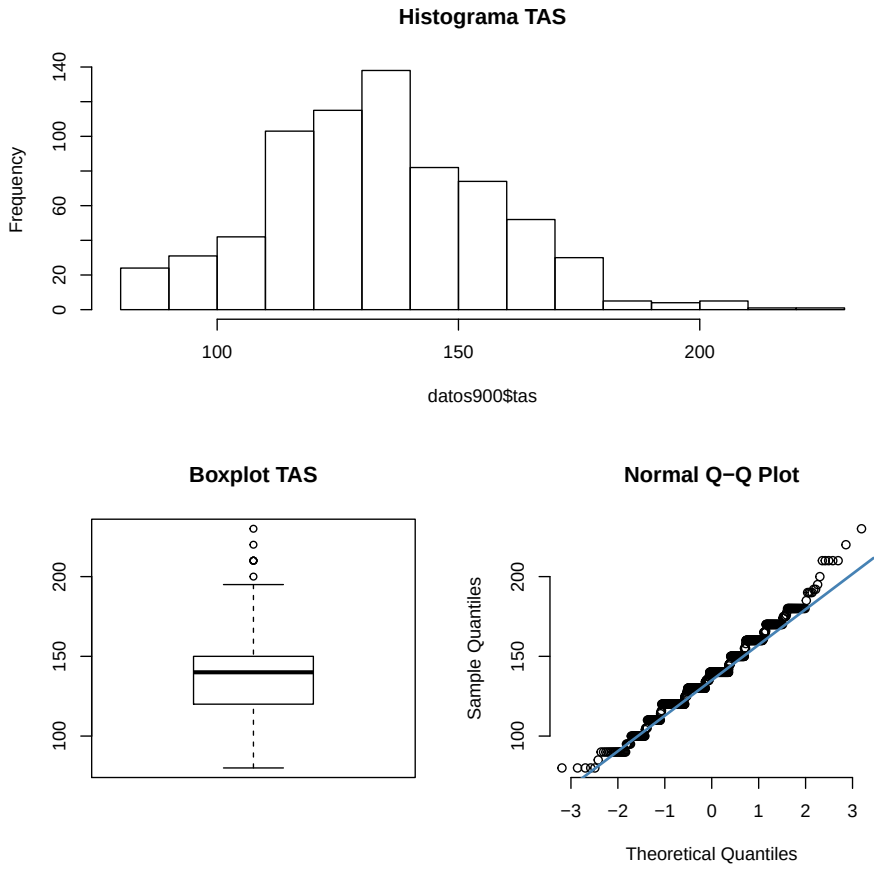
```
chisq.test(cont_table)

##
##  Pearson's Chi-squared test
##
## data:  cont_table
## X-squared = 1.498, df = 2, p-value = 0.4728
```

El p-valor de la prueba χ^2 no es inferior a $\alpha = 0.05$, es decir, no tenemos evidencias suficientes para rechazar la hipótesis nula, así que aceptaremos la hipótesis nula y por lo tanto el hecho de tener diabetes no influye en el tipo de patología del paciente ingresado.

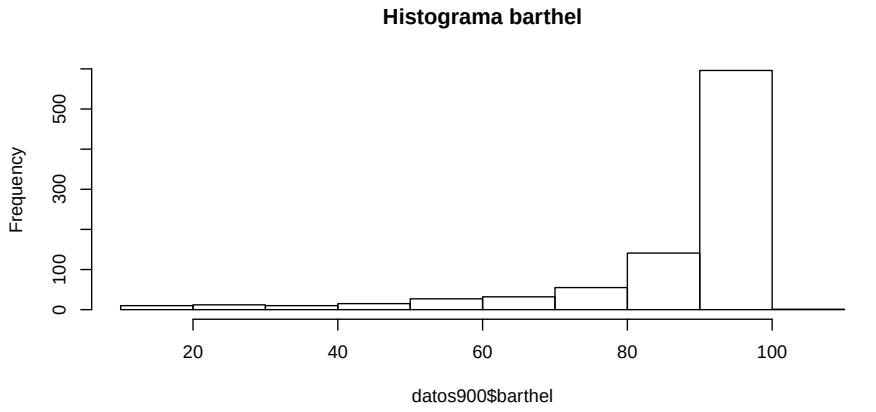
2.2 Comprueba la normalidad de las variables tensión arterial(TAS) y score de Barthel.

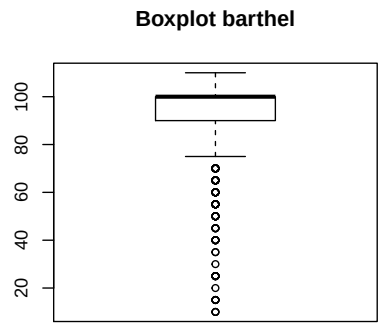
Ambas variables son cuantitativas. La tensión arterial es cuantitativa contínua mientras que el score de Barthel es una variable discreta. Para estudiar la normalidad vamos a dibujar los boxplots de dichas variables, un histograma y un Q-Q plot (para ver las desviaciones con respecto a una normal teórica).

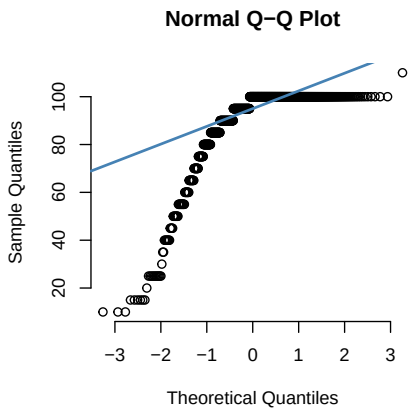
```
layout(matrix(c(1,1,2,3), 2, 2, byrow = TRUE))
hist(datos900$tas, main='Histograma TAS')
boxplot(datos900$tas, main = 'Boxplot TAS')
qqnorm(datos900$tas, frame = FALSE)
qqline(datos900$tas, col = "steelblue", lwd = 2)
```




```
layout(matrix(c(1,1,2,3), 2, 2, byrow = TRUE))
hist(datos900$barthel, main='Histograma barthel')
boxplot(datos900$barthel, main = 'Boxplot barthel')
qqnorm(datos900$barthel, frame = FALSE)
qqline(datos900$barthel, col = "steelblue", lwd = 2)
```







A la vista de los gráficos, nuestras variables no parecen seguir la distribución normal. A continuación realizaremos una test de normalidad (Shapiro-Wilks) para confirmar.

```
cat("Test de normalidad para TAS\n")

## Test de normalidad para TAS

shapiro.test(datos900$tas)

##
##  Shapiro-Wilk normality test
##
## data:  datos900$tas
## W = 0.98261, p-value = 1.925e-07

cat("Test de normalidad para barthel\n")
```

```
## Test de normalidad para barthel

shapiro.test(datos900$barthel)

##
##  Shapiro-Wilk normality test
##
## data:  datos900$barthel
## W = 0.63314, p-value < 2.2e-16
```

Dado que los p-valores para ambas son menores que $\alpha = 0.05$, podemos decir que las variables `tas` y `barthel` no siguen la distribución normal.

2.3 ¿Existe asociación entre la TAS y el tipo de enfermedad por el que ingreso el paciente? ¿Y entre el score de Barthel y el tipo de Caso?

En el apartado anterior hemos visto que la variable `tas` no sigue una distribución normal, por lo tanto para contrastar la asociación de esta variable con la variable `tipocas` (que es una variable cualitativa) emplearemos tests no paramétricos.

La hipótesis nula (H_0): la tensión arterial sistólica y el tipo de patología del paciente son independientes, es decir, la variable `tas` se distribuye igual en los distintos grupos de `tipocas`.
La hipótesis alternativa (H_1): la tensión arterial sistólica y el tipo de patología del paciente no son independientes, es decir, la variable `tas` se distribuye de forma distinta en los distintos grupos de `tipocas`.

Como la variable `tipocas` tiene 3 niveles, emplearemos el test de Kruskal-Wallis.

```
kruskal.test(datos900$tas, datos900$tipocas)

##
##  Kruskal-Wallis rank sum test
##
## data:  datos900$tas and datos900$tipocas
## Kruskal-Wallis chi-squared = 16.183, df = 2, p-value = 0.0003061
```

Como el p-valor del test es menor a $\alpha = 0.05$, rechazamos la hipótesis nula y por tanto aceptamos la hipótesis alternativa, es decir que la variable `tas` se distribuye de forma distinta en los distintos grupos de `tipocas`. Para ver exactamente qué grupos tienen distribuciones distintas, haremos comparaciones dos a dos mediante el test de Dunn y empleando el ajuste de Benjamini-Hochberg.

```
library(dunn.test)
dunn.test(datos900$tas, datos900$tipocas, method='bh')
```

```
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 16.183, df = 2, p-value = 0
##
##
##                             Comparison of x by group
##                               (Benjamini-Hochberg)
## Col Mean-|
## Row Mean |        ACV      Neumo
## ---------+----------------------
##    Neumo |   3.090524
##          |    0.0015*
##          |
##      SCA |   3.720938   0.263347
##          |    0.0003*     0.3961
##
## alpha = 0.05
## Reject Ho if p <= alpha/2
```

Podemos ver que el grupo que muestra diferencias significativas en la presión arterial sistólica es el compuesto por los pacientes ingresados por accidente cardiovascular ya que en el test de Dunn, todos los emparejamientos con este grupo muestran p-valores inferiores al nivel de significancia $\alpha/2 = 0.025$. Es equivalente decir que los grupos compuests por pacientes ingresados por neumonía y por síndrome coronario agudo no tienen diferencias significativas en la presión arterial sistólica.

Con respecto a la variable `barthel`, en el apartado anterior hemos visto también que la variable `barthel` no sigue una distribución normal, por lo tanto para contrastar la asociación de esta variable con la variable `tipocas` emplearemos tests no paramétricos.

La hipótesis nula (H_0): el Barthel score y el tipo de patología del paciente son independientes, es decir, la variable `barthel` se distribuye igual en los distintos grupos de `tipocas`.
La hipótesis alternativa (H_1): el Barthel score y el tipo de patología del paciente no son independientes, es decir, la variable `barthel` se distribuye de forma distinta en los distintos grupos de `tipocas`.

De nuevo emplearemos el test de Kruskal-Wallis ya que la variable `tipocas` tiene 3 niveles.

```r
kruskal.test(datos900$barthel, datos900$tipocas)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  datos900$barthel and datos900$tipocas
## Kruskal-Wallis chi-squared = 12.072, df = 2, p-value = 0.002391
```

De la misma manera que en el caso para la variable `tas`, el p-valor del test es menor a $\alpha = 0.05$, rechazamos la hipótesis nula y por tanto aceptamos que la variable `barthel` se distribuye de forma distinta en los distintos grupos de de `tipocas`. Para ver exactamente qué grupos tienen distribuciones distintas, haremos otra vez comparaciones dos a dos mediante el test de Dunn y empleando el ajuste de Benjamini-Hochberg.

```r
library(dunn.test)
dunn.test(datos900$barthel, datos900$tipocas, method='bh')
```

```
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 12.0725, df = 2, p-value = 0
##
##
##                             Comparison of x by group
##                               (Benjamini-Hochberg)
## Col Mean-|
## Row Mean |        ACV      Neumo
## ---------+----------------------
##    Neumo |  -0.876391
##          |     0.1904
##          |
##      SCA |  -3.300398  -2.612095
##          |    0.0014*    0.0067*
##
## alpha = 0.05
## Reject Ho if p <= alpha/2
```

Podemos ver que el grupo que muestra diferencias significativas en el Barthel score es el compuesto por los pacientes ingresados por síndrome coronario agudo ya que en el test de Dunn, todos los emparejamientos con este grupo muestran p-valores inferiores al nivel de significancia $\alpha/2 = 0.025$. Es equivalente decir que los grupos compuestos por pacientes ingresados por neumonía y por accidente cardiovascular no tienen diferencias significativas en la presión arterial sistólica.

2.4 ¿Existe asociación entre la TAS y el ser caso o control? ¿Y entre el Barthel y el ser caso o control?

Este apartado es similar al anterior. Recordemos que la variable `tas` no sigue una distribución normal, por lo tanto, de la misma manera que en el apartado anterior, para contrastar la asociación de esta variable con la variable `caso` (que es una variable cualitativa dicotómica) emplearemos el test no paramétrico U-Mann-Whitney.

La hipótesis nula (H_0): la tensión arterial sistólica y ser caso o control es independiente, es decir, la variable `tas` se distribuye igual en los distintos grupos de `caso`.
La hipótesis alternativa (H_1): la tensión arterial sistólica y ser caso o control es independiente, es decir, la variable `tas` se distribuye de forma distinta en los distintos grupos de `caso`.

Aplicamos el test U-Mann-Whitney:

```
wilcox.test(tas~caso, data=datos900)

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  tas by caso
## W = 44077, p-value = 0.003123
## alternative hypothesis: true location shift is not equal to 0
```

El p-valor del test es menor a $\alpha = 0.05$, rechazamos la hipótesis nula y por tanto aceptamos que la variable `tas` se distribuye de forma distinta en los distintos grupos de `caso`.

Para las variables `barthel` y `caso`, procedemos de forma idéntica (empleando el test no paramétrico U-Mann-Whitney.) ya que `barthel` es una variable cuantitativa discreta cuya distribución no sigue a la distribución normal, y la variable `caso` es cualitativa dicotómica.

La hipótesis nula (H_0): el barthel score y ser caso o control es independiente, es decir, la variable `barthel` se distribuye igual en los distintos grupos de `caso`.
La hipótesis alternativa (H_1): el barthel score y ser caso o control es independiente, es decir, la variable `barthel` se distribuye de forma distinta en los distintos grupos de `caso`.

Aplicamos el test U-Mann-Whitney:

```
wilcox.test(barthel~caso, data=datos900)

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  barthel by caso
## W = 92518, p-value = 0.5966
## alternative hypothesis: true location shift is not equal to 0
```

El p-valor del test es mayor a $\alpha = 0.05$ por lo tanto no tenemos suficientes evidencias para rechazar la hipótesis nula y por ello aceptamos que la variable `tas` se distribuye igual en los distintos grupos de `caso`.

Selecciona 40 casos al azar de la base de datos

```
# Selecciona 40 casos al azar
set.seed(202020)
datos40 <- datos[sample(1:nrow(datos), 40, replace=FALSE),]
```

2.5 ¿Existe asociación entre la TAS y el ser caso o control? ¿y entre el Barthel y el ser caso o control?

Nuevamente planteamos las hipótesis del apartado anterior para cada caso:

2.5.1 Tensión arterial sistólica

La hipótesis nula (H_0): la tensión arterial sistólica y ser caso o control es independiente, es decir, la variable `tas` se distribuye igual en los distintos grupos de `caso`.
La hipótesis alternativa (H_1): la tensión arterial sistólica y ser caso o control es independiente, es decir, la variable `tas` se distribuye de forma distinta en los distintos grupos de `caso`.

Aplicamos el test U-Mann-Whitney:

```
wilcox.test(tas~caso, data=datos40)

## Warning in wilcox.test.default(x = c(120L, 145L, 130L, 120L, 100L,
140L, :  cannot compute exact p-value with ties

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  tas by caso
## W = 55, p-value = 0.03225
## alternative hypothesis: true location shift is not equal to 0
```

Tal y como ha ocurrido en el apartado anterior, el p-valor del test es menor a $\alpha = 0.05$, rechazamos la hipótesis nula y por tanto aceptamos que la variable `tas` se distribuye de forma distinta en los distintos grupos de `caso`. ***Nota**: como nuestros datos tienen un tamaño menor a 50 y ocurren "empates" (ties) en el ranking que hace el test, el p-valor que se nos muestra es un p-valor aproximado a una normal.*

2.5.2 Barthel score

La hipótesis nula (H_0): el barthel score y ser caso o control es independiente, es decir, la variable `barthel` se distribuye igual en los distintos grupos de `caso`.
La hipótesis alternativa (H_1): el barthel score y ser caso o control es independiente, es decir, la variable `barthel` se distribuye de forma distinta en los distintos grupos de `caso`.
`caso`.

Aplicamos el test U-Mann-Whitney:

```
wilcox.test(barthel~caso, data=datos40)

## Warning in wilcox.test.default(x = c(35L, 100L, 50L, 100L, 100L,
95L, 75L, :  cannot compute exact p-value with ties

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  barthel by caso
## W = 198, p-value = 0.4914
## alternative hypothesis: true location shift is not equal to 0
```

Nuevamente, el p-valor del test no es menor a $\alpha = 0.05$, por lo tanto no tenemos suficientes evidencias para rechazar la hipótesis nula y por ello aceptamos que la variable `tas` se distribuye igual en los distintos grupos de `caso`.

3 Ejercicios Adicionales

3.1 Indica para que te podrían servir las técnicas de bootstrap en los análisis estadísticos que has realizado. Calcula el intervalo utilizando Bootstrap de la mediana de la tensión arterial

Las técnicas de bootstrap consisten en llevar a cabo remuestreos no de la distribución teórica de la población si no de la empírica (a partir de los datos de una muestra grande que tengamos de dicha población). Estas técnicas permiten aproximar la varianza, construir intervalos de confianza... En mi opinión,

la técnica bootstrap más útil para los análisis estadísticos de este informe es el test de permutaciones para comparaciones de grupos. Imaginemos que queremos contrastar una hipótesis nula de igualdad de medias de grupos (como hemos hecho anteriormente), pero desconocemos la distribución nula. El procedimiento a seguir es computar el estadístico (la diferencia de medias en este ejemplo concreto), permutar (sin sustitución) las etiquetas (en inglés, labels) de los grupos y recomputar el estadístico. Repitiendo esto muchas veces (10000 por ejemplo) y dibujando un histograma de conteos de los estadísticos, obtendríamos una aproximación a la distribución nula.

Con esta distribución podemos calcular el p-valor de la siguiente manera: contamos el número de veces que obtuvimos un estadístico igual o más alto que el calculado inicialmente y lo dividimos entre el total de estadísticos computadas. De esta forma podremos rechazar o aceptar la hipótesis nula.

Para el cómputo del intervalo de confianza bootstrap al 95% de la mediana de la presión arterial seguimos este procedimiento:

```
# Como hay algunos NaNs, vamos a eliminarlos para que no nos dé problemas.
TAS_no_NA = na.omit(datos900$tas)

# Vamos a hacer un Boostrap de 10000 remuestreos.
B <- 10000
remuestreos <- matrix(
                      sample(
                             TAS_no_NA,
                             length(TAS_no_NA)*B,
                             replace = TRUE
                             ),
                      B,
                      length(TAS_no_NA)
                      )

# Calculamos la mediana de cada una de los 10000 remuestreos
medianas <- apply(remuestreos, 1, median)

# Computamos los percentiles 2.5 y 97.5
intervalo <- quantile(medianas, c(0.025, 0.975))

cat(
    sprintf(
        "El intervalo de confianza bootstrap al 95%% para la tensión arterial es [%.d, %.d].
        intervalo[[1]],
        intervalo[[2]]
        )
    )

## El intervalo de confianza bootstrap al 95% para la tensión arterial es [135, 140].
```

3.2 Indica porqué hay que utilizar técnicas de comparacion múltiple como la corrección de Bonferroni

Existen escenarios en el que el investigador desea comparar más de 2 grupos de datos para hallar diferencias entre estos. Habitualmente se emplea el ANOVA, el cual indica la existencia de diferencias entre las medianas de los grupos objetivo de estudio, sin embargo no nos indica qué grupos en concreto difieren o cómo difieren dos grupos. Es en este escenario donde las técnicas de comparación múltiple son necesarias. No obstante estas técnicas requieren especial cuidado, ya que a mayor número comparaciones, más aumenta el error tipo I. La corrección de Bonferroni intenta compensar este aumento del error tipo I y consiste en cambiar el p-valor de p a p/n donde n es el número de grupos de datos a comparar.