PRA2

2022-05-16

Instalamos y cargamos las librerías.

```
# https://cran.r-project.org/web/packages/ggplot2/index.html
if (!require('ggplot2')) install.packages('ggplot2'); library('ggplot2')
# https://cran.r-project.org/web/packages/dplyr/index.html
if (!require('dplyr')) install.packages('dplyr'); library('dplyr')
if (!require('caret')) install.packages('caret'); library('caret')
if (!require('normtest')) install.packages('normtest'); library('normtest')
if (!require('car')) install.packages('car'); library('car')
if (!require('gridExtra')) install.packages('gridExtra'); library('gridExtra')
```

Ejercicio 1

Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

1.1 Carga de los datos

Cargamos el dataset.

```
df <- read.csv('../data/train.csv',stringsAsFactors = FALSE)
```

Confirmamos que el dataset se ha cargado correctamente:

```
head(df)
```

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
## 6           6        0      3
##                                                 Name    Sex Age SibSp Parch
## 1                            Braund, Mr. Owen Harris   male  22     1     0
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1     0
## 3                             Heikkinen, Miss. Laina female  26     0     0
## 4       Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1     0
## 5                           Allen, Mr. William Henry   male  35     0     0
## 6                                   Moran, Mr. James   male  NA     0     0
##             Ticket    Fare Cabin Embarked
## 1        A/5 21171  7.2500              S
```

```
## 2         PC 17599 71.2833   C85        C
## 3 STON/O2. 3101282  7.9250              S
## 4           113803 53.1000  C123        S
## 5           373450  8.0500              S
## 6           330877  8.4583              Q
```

1.2 Descripción del dataset

Descripción del dataset:

```
str(df)
```

```
## 'data.frame':    891 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
##  $ Sex        : chr  "male" "female" "female" "female" ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : chr  "" "C85" "" "C123" ...
##  $ Embarked   : chr  "S" "C" "S" "S" ...
```

Observamos que el conjunto de datos consta de 891 observaciones y 12 variables que describen la tripulación del Titanic.

PassengerId Número identificador del pasajero (Integer).

Survived Variable que especifica si un pasajero ha sobrevivido. 0 = No, 1 = Si (Integer).

Pclass Clase socio económica del pasajero. 1 = Clase Alta, 2 = Clase Media, 3 = Clase baja (Integer).

Name Nombre del pasajero (String).

Sex Sexo del pasajero (String).

Age Edad del pasajero. Para pasajeros de menos de un año, la edad esta fraccionada. Si la edad esta estimada, su formato es xx.5 (Numeric).

SibSp Número de familiares, hermanos o esposos, a abordo (Integer).

Parch Número de familiares, padres o hijos, a bordo (Ingeger).

Ticket Número del billete (String).

Fare Precio del billete (Numeric).

Cabin Número de cabina (String).

Embarked Puerto en el que embarco el pasajero. C = Cherbourg, Q = Queenstown, S = Southampton (String).

Observamos que muchas de estas variables se presentan como Integers o Strings, cuando realmente nos expresan una categoria. Por ello factorizaremos las variables categoricas.

```
df$Survived <- factor(df$Survived)
df$Pclass <- factor(df$Pclass)
df$Sex <- factor(df$Sex)
df$SibSp <- factor(df$SibSp)
df$Parch <- factor(df$Parch)
df$Embarked <- factor(df$Embarked)
```

1.3 Motivación del estudio

Cuando el Titanic fue construido, se considero que este era imposible de hundir. Durante su viaje inaugural, el Titanic se hundió tras chocar contra un iceberg y la escasez de barcos salvavidas convirtió el accidente en una gran tragedia que resultó en la muerte de 1502 personas del conjunto de 2224 pasajeros y tripulación. Claramente la suerte tuvo un papel importante en determinar quien sobrevivió y quien no, pero se puede intuir que ciertos grupos de personas tuvieron más probabilidades de hacerlo que otros.

En este estudio nos centraremos en responder a la pregunta "Que grupos de personas tuvieron más probabilidades de sobrevivir el accidente?". O dicho de otra forma, "¿Cuales fueron los factores de peso que influyeron en que los pasajeros tuvieran mayor probabilidad de sobrevivir?".

Ejercicio 2

Integración y selección de los datos de interés a analizar. Puede ser el resultado de adicionar diferentes datasets o una subselección útil de los datos originales, en base al objetivo que se quiera conseguir.

La principal variable a incluir va a ser **Survived**. Esta será nuestra variable objetivo a estudiar ya que queremos determinar que factores tuvieron más peso al aumentar la probabilidad de sobrevivir de los pasajeros. A continuación querremos seleccionar todas las variables que puedan describir el pasajero, y eliminar las variables irrelevantes. Observando la previa descripción de las variables realizada en el ejercicio 1, rápidamente podemos identificar que las variables **Cabin**, el número de cabina, y **Ticket**, el número del billete, no nos aportan información relevante para describir el pasajero. De mismo modo, las variables **Nombre** y **PassangerId** que sirven para identificar a los pasajeros, tampoco nos aportan información descriptiva sobre ellos.

Procedemos a descartar las variables irrelevantes.

```
irrelevant <- c("Cabin", "Ticket", "Name", "PassengerId")
titanic <- df[ , -which(names(df) %in% irrelevant)]
```

Ejercicio 3

Limpieza de los datos.

3.1 Valores vacios

¿Los datos contienen ceros o elementos vacíos? Gestiona cada uno de estos casos.

Estadísticas de valores vacíos.

```
colSums(is.na(titanic))
```

```
## Survived   Pclass      Sex      Age    SibSp    Parch     Fare Embarked
##        0        0        0      177        0        0        0        0
```

```
colSums(titanic=="")
```

```
## Survived   Pclass      Sex      Age    SibSp    Parch     Fare Embarked
##        0        0        0       NA        0        0        0        2
```

Observamos 177 valores vacíos en la variable **Age** y 2 valores vacíos en la variable **Embarked**.

Si nos centramos en el caso de la variable **Age**, esta contiene 177 valores vacíos. Como eliminar 177 observaciones del conjunto de datos supondría una gran perdida de información ya que este contiene 801 observaciones, optaremos por la estrategia de completar los valores vacíos usando la media de edad del conjunto de datos.

```
titanic$Age[is.na(titanic$Age)] <- mean(titanic$Age, na.rm=T)
```

Debido a que la variable **Embarked** tiene solo 2 valores vacíos, no vemos la necesidad de crear una categoría nueva "Desconocido" para substituir los valores vacíos, creemos que la mejor estrategia en este caso, es substituir estos valores vacíos por la categoría más repetida, evitando así la perdida de información que supondría eliminar las observaciones.

```
# Determinamos la categoría que se repite con mayor frecuencia
table(titanic$Embarked)
```

```
##
##       C   Q   S
##   2 168  77 644
```

Observamos que la categoría 'S' es la que se repite con mayor frecuencia.

Procedemos a sustituir los valores vacíos por 'S'.

```
titanic$Embarked[titanic$Embarked == ""] <- "S"
```

Comprobamos que ya no existen valores vacíos en el conjunto de datos.

```
colSums(is.na(titanic))
```

```
## Survived   Pclass      Sex      Age    SibSp    Parch     Fare Embarked
##        0        0        0        0        0        0        0        0
```

```
colSums(titanic=="")
```

```
## Survived   Pclass      Sex      Age    SibSp    Parch     Fare Embarked
##        0        0        0        0        0        0        0        0
```

3.2 Valores extremos

Identifica y gestiona los valores extremos.

Los valores extremos se dan en las variables continuas cuando observamos valores significativamente altos o significativamente bajos, que generan duda sobre su veracidad.

Observemos las estadísticas básicas del conjunto de datos para identificar posibles valores extremos (*outliers*) en las variables continuas **Age**, **SibSp**, **Parch** y **Fare**.

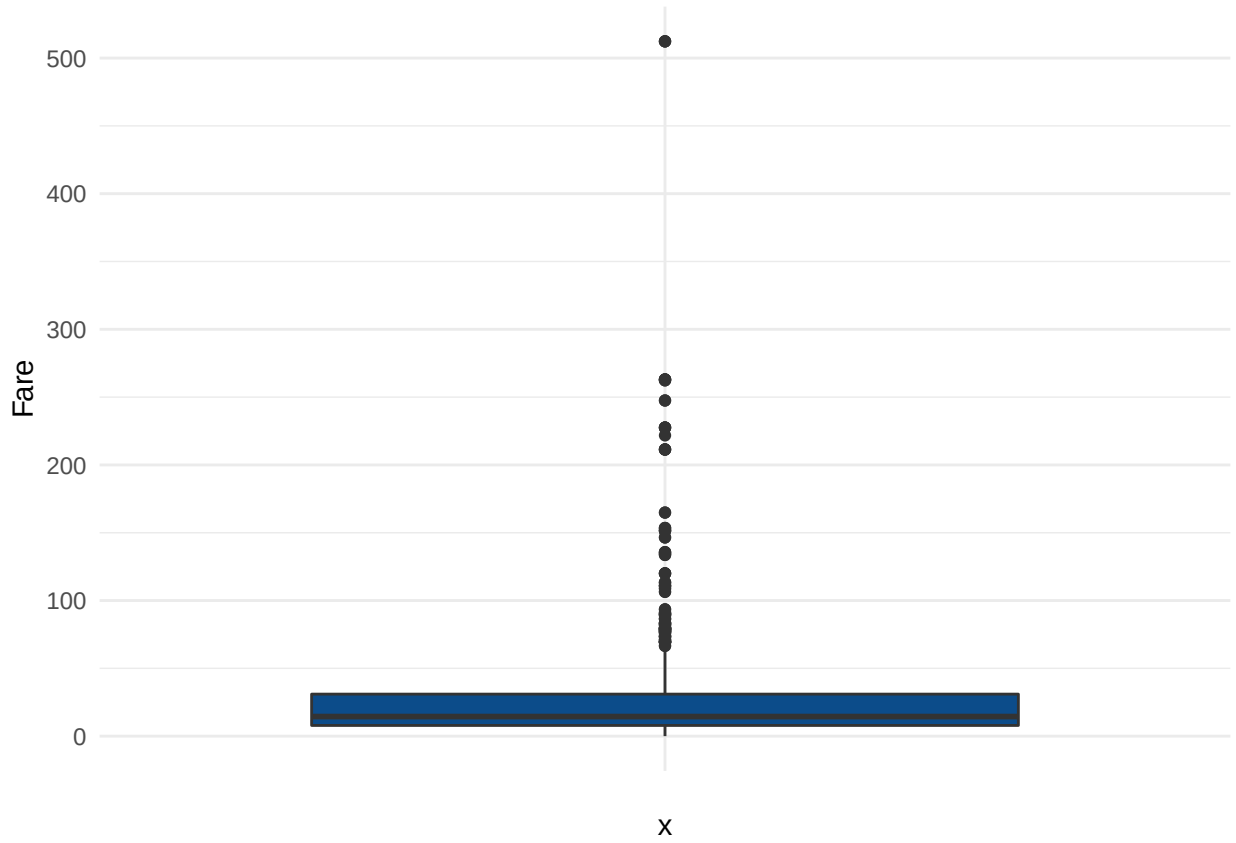
```
summary(titanic)
```

```
##  Survived Pclass      Sex           Age        SibSp   Parch        Fare
##  0:549    1:216   female:314   Min.   : 0.42   0:608   0:678   Min.   :  0.00
##  1:342    2:184   male  :577   1st Qu.:22.00   1:209   1:118   1st Qu.:  7.91
##           3:491                Median :29.70   2: 28   2: 80   Median : 14.45
##                                Mean   :29.70   3: 16   3:  5   Mean   : 32.20
##                                3rd Qu.:35.00   4: 18   4:  4   3rd Qu.: 31.00
##                                Max.   :80.00   5:  5   5:  5   Max.   :512.33
##                                                8:  7   6:  1
##  Embarked
##   :  0
##  C:168
##  Q: 77
##  S:646
##
##
##
```

Observamos que los valores mínimos y máximos de la variable **Age**, se mantienen dentro de un rango lógico. **SibSp** y **Parch**, tienen un valor máximo alto, pero parece consistente con los estándares de familias de la época. La variable que será necesaria analizar con profundidad es **Fare**. Esta variable contiene un valor máximo muy superior al valor medio.

Utilizaremos la representación *boxplot* para identificar los *outliers* de **Fare**.

```
ggplot(titanic) +
  aes(x = "", y = Fare) +
  geom_boxplot(fill = "#0c4c8a") +
  theme_minimal()
```



Con esta representación, podemos identificar claramente *outilers* correspondientes al valor máximo de 512.33. Optaremos por eliminar estos *outliers* ya que la diferencia respecto a las demás observaciones es tan significativa que sospechamos que tales observaciones contienen información no verídica. Los otros puntos que son detectados como *outliers*, menores a 300, corresponden a los precios pagados por los pasajeros de primera clase. A continuación mostramos algunas de estas observaciones para **Fare** mayor a 100.

```
head(titanic[titanic$Fare > 150, ])
```

```
##     Survived Pclass    Sex Age SibSp Parch     Fare Embarked
## 28         0      1   male  19     3     2 263.0000        S
## 89         1      1 female  23     3     2 263.0000        S
## 119        0      1   male  24     0     1 247.5208        C
## 259        1      1 female  35     0     0 512.3292        C
## 269        1      1 female  58     0     1 153.4625        S
## 298        0      1 female   2     1     2 151.5500        S
```

Observamos que efectivamente, esas observaciones corresponden a pasajeros de primera clase.

Eliminamos los *outliers* detectados:

```
titanic <- subset(titanic, titanic$Fare < 500)
```

Ejercicio 4

Análisis de los datos.

4.1 Elección de los datos a analizar

En este análisis pretendemos analizar el peso de las distintas variables del conjunto sobre la variable dependiente **Survived**. En el Ejercicio 2 escogimos el subset del conjunto de datos que consideramos de interés analítico para el estudio. Este consiste de las variables **Survived**, **Pclass**, **Sex**, **Age**, **SibSp**, **Parch**, **Fare** y **Embarked**.

De las variables de interés para el análisis, seleccionaremos ciertas agrupaciones que nos resultan interesantes de analizar en más profundidad. Estas serán la agrupación por genero, la variable **Sex**, y la agrupación por clase socio económica, la variable **Pclass**.

```
# Agrupación por género
titanic.men   <- subset(titanic, Sex == "male")
titanic.women <- subset(titanic, Sex == "female")

# Agrupación por clase socio económica
titanic.high  <- subset(titanic, Pclass == "1")
titanic.mid   <- subset(titanic, Pclass == "2")
titanic.low   <- subset(titanic, Pclass == "3")
```

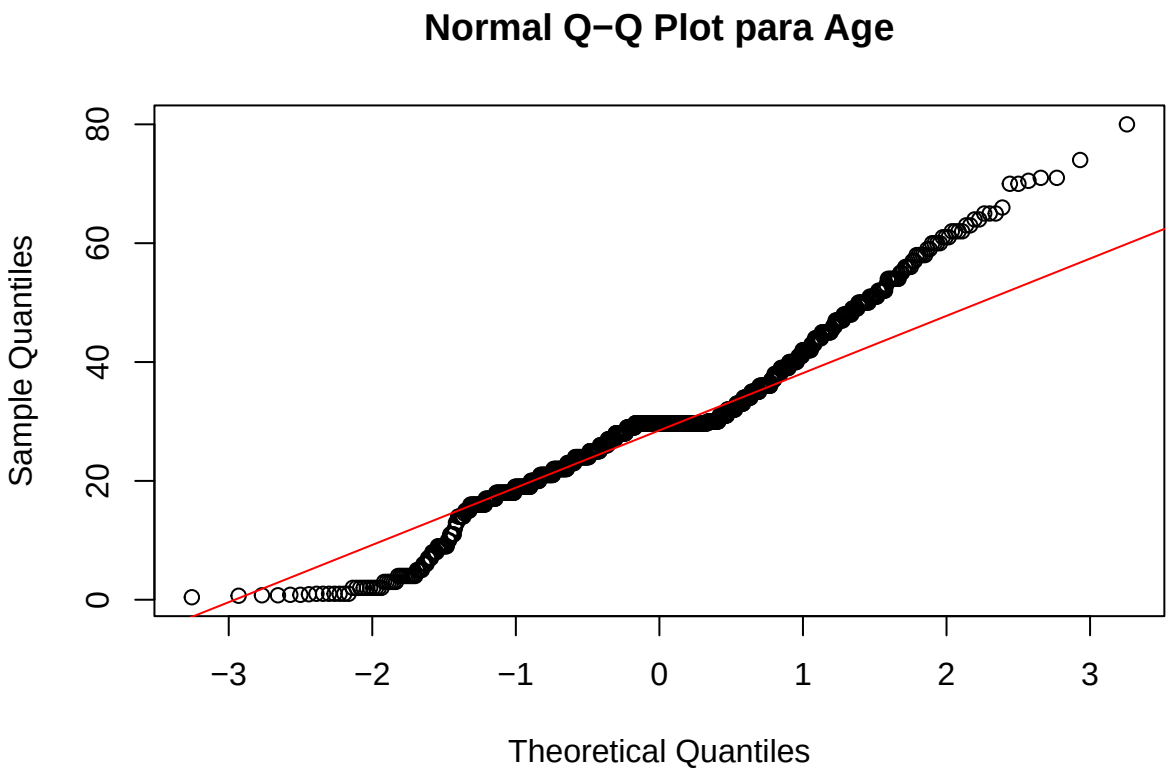
4.2 Normalidad y Homogeneidad

Comprobación de la normalidad y homogeneidad de la varianza.

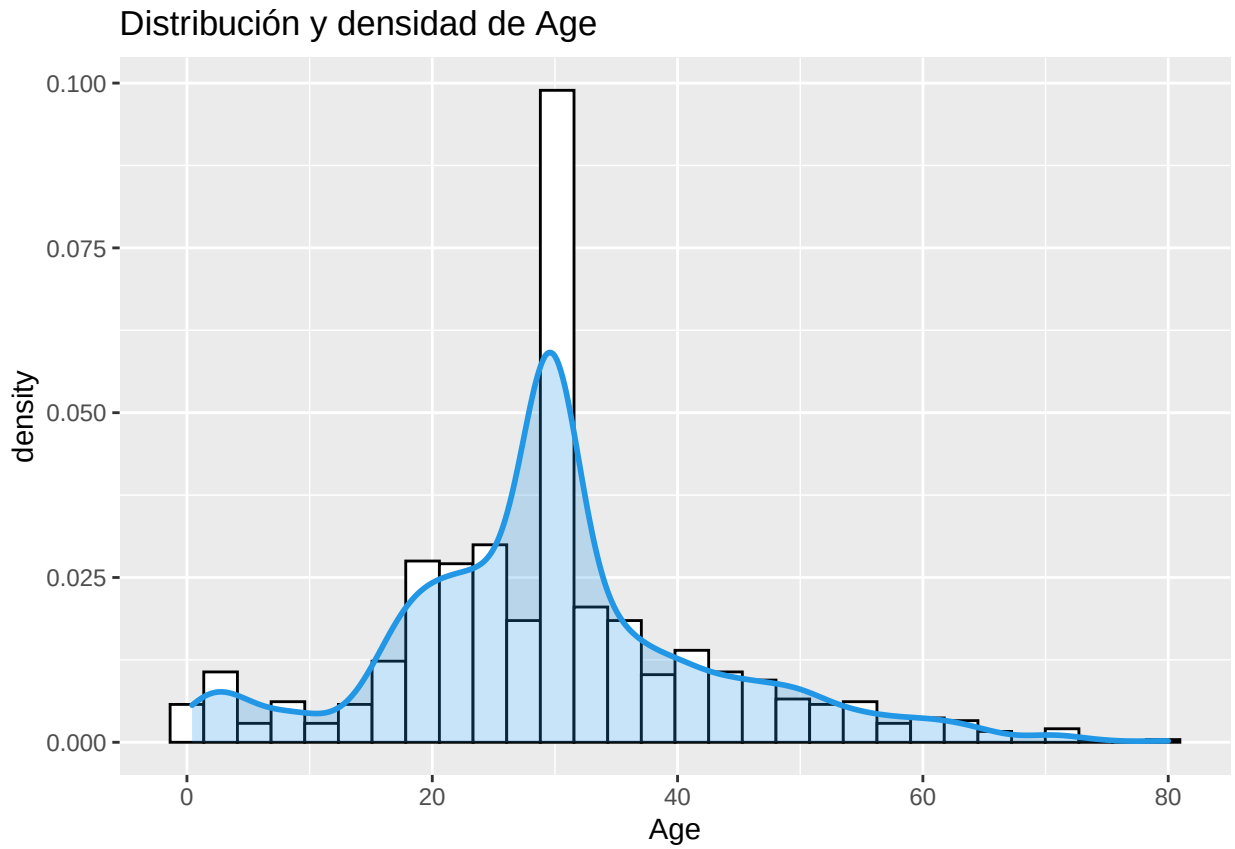
Analizaremos la normalidad y la homogeniedad de la varianza para las variables continuas **Age** y **Fare**.

Age: Análisis visual de normalidad:

```
qqnorm(titanic$Age,main = "Normal Q-Q Plot para Age")
qqline(titanic$Age,col = "red")
```




```
#[1]
ggplot(titanic, aes(x = Age)) +
  geom_histogram(aes(y = ..density..), bins = 30,
                 colour = 1, fill = "white") +
  geom_density(lwd = 1, colour = 4,
               fill = 4, alpha = 0.25) +
  ggtitle("Distribución y densidad de Age")
```



Para determinar que puede haber normalidad, deberíamos poder ver como la mayoría de observaciones se alinea encima de la línea roja (indica el comportamiento esperado de una variable normal) para la representación Quantil-Quantil. Este no es el caso observado para la variable **Age**.

Si nos fijamos en el histograma de la variable con su correspondiente curva de densidad, para que hubiera normalidad, deberíamos observar una curva de densidad simétrica con forma de campana, por lo que podemos determinar que no es una variable normal. Podemos observar de la distribución, que la mayoría de pasajeros eran jóvenes de entre 20 y 30 años.

Análisis por tests de normalidad:

```
#[2]
## Shapiro test
shapiro.test(titanic$Age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  titanic$Age
## W = 0.95851, p-value = 3.678e-15
```

```
## Jarque Bera test
jb.norm.test(titanic$Age)
```

```
##
##  Jarque-Bera test for normality
```

```
##
## data:  titanic$Age
## JB = 61.607, p-value < 2.2e-16
```

```
## Forsini test
frosini.norm.test(titanic$Age)
```

```
##
##  Frosini test for normality
##
## data:  titanic$Age
## B = 1.4197, p-value < 2.2e-16
```

Para los tests de normalidad Shapiro-Wilk, Jarque-Bera y Frosini, la hipótesis nula indica normalidad, mientras que la hipótesis alternativa indica que no hay normalidad. Un valor p superior al nivel de significancia ~0.05, nos indicaría que no podemos descartar la hipótesis nula de normalidad. Para todos los test probados, observamos unos valores p muy inferiores al nivel de significancia, que nos indican que no estamos frente a una variable normal.

Análisis de homogeniedad de la varianza:

```
leveneTest(y = titanic$Age, group = titanic$Survived, center = "median")
```

```
## Levene's Test for Homogeneity of Variance (center = "median")
##        Df F value  Pr(>F)
## group   1  5.7214 0.01697 *
##       886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Para evaluar la homocedasticidad, o la homogeniedad de la varianza, para la variable **Age**, debemos tener en cuenta que, como determinamos previamente, no se trata de una variable normal. Por ello, la elección del test Levene es la más conveniente siendo este menos sensible a que la ausencia de normalidad [3].

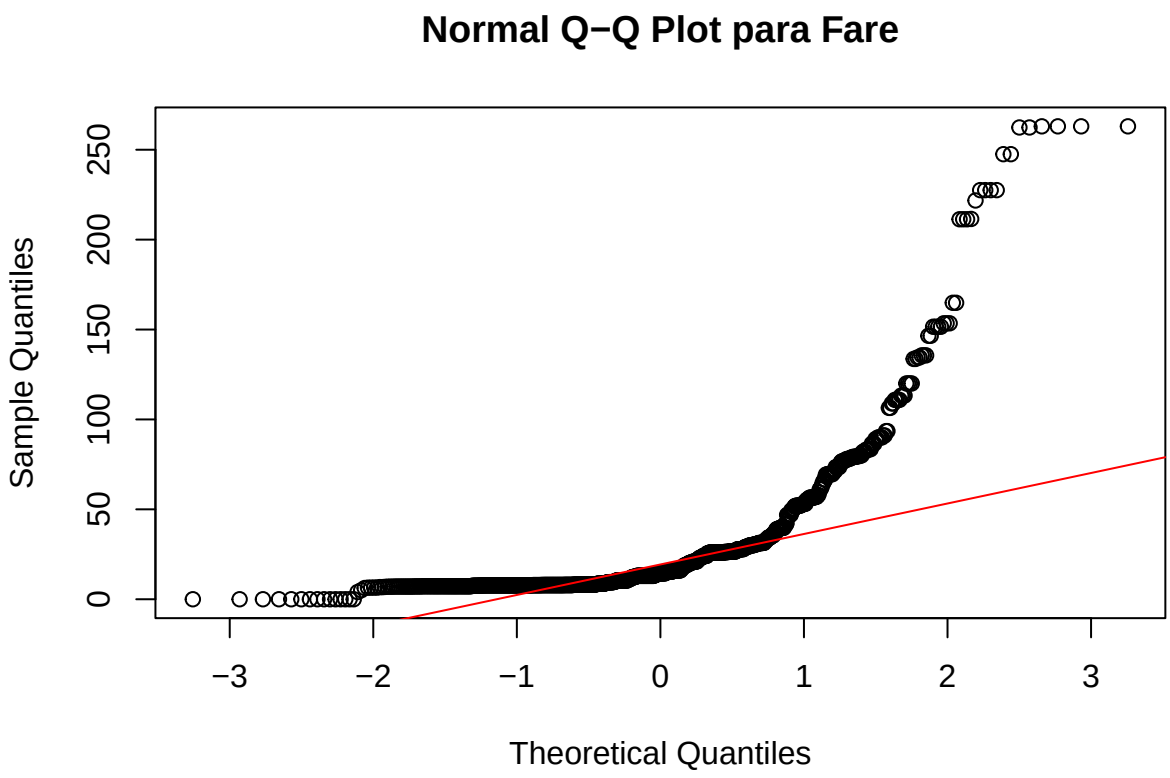
Hipótesis nula: Ambas varianzas son iguales. Hipótesis alternativa: Las varianzas son distintas entre si.

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

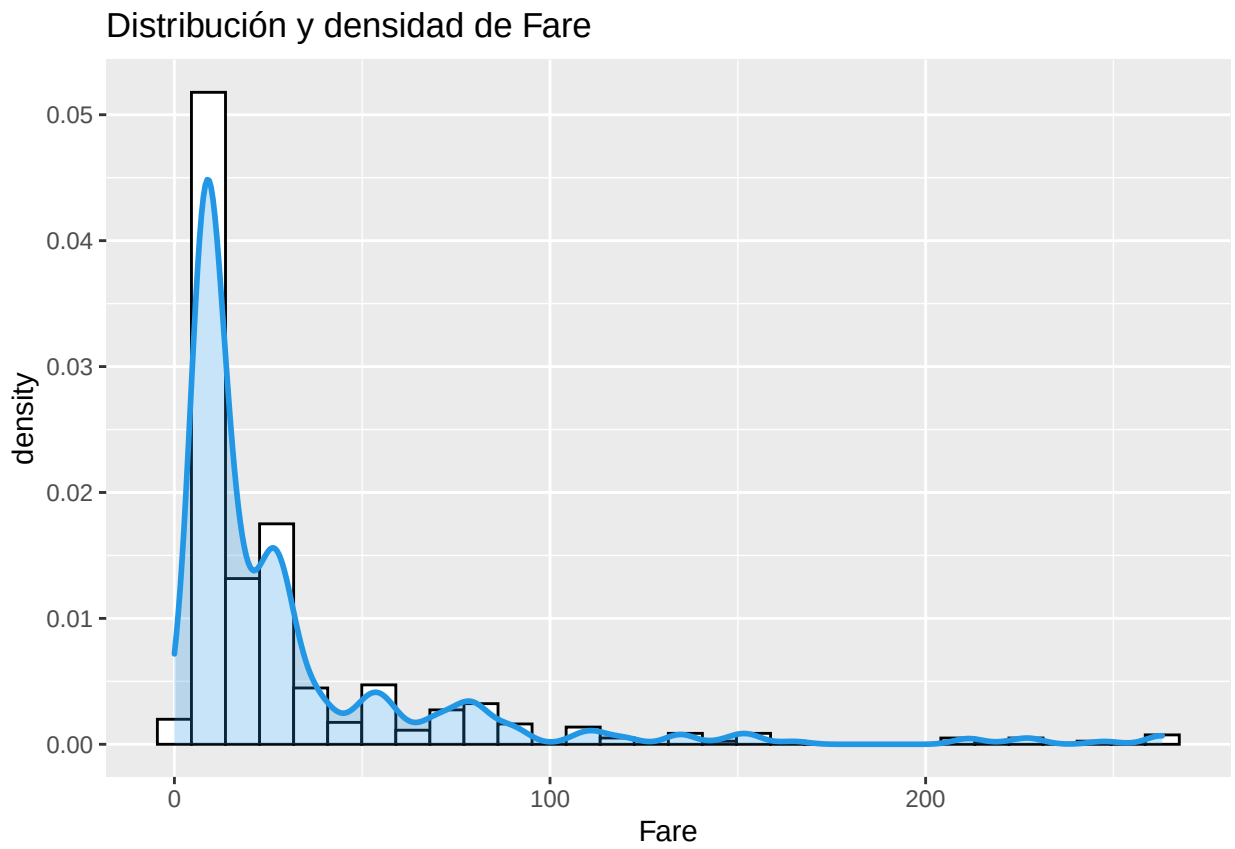
Se aplica el test para la variable **Age** y se comparan las varianzas cuando esta variable es agrupada según **Survived**, la variable dependiente que queremos estudiar. Podemos observar de los resultados un valor p igual a 0.0169. Esto nos indica que debemos descartar la hipótesis nula, es decir, el hecho de que ambas varianzas sean iguales, en el caso de optar por un nivel de significancia de 0.01 o 0.001. Si optamos por un nivel de significancia de 0.05 o 0.1, no podemos descartar la hipótesis nula y por lo tanto las varianzas de ambos grupos deben considerarse iguales.

Fare Análisis visual de normalidad:

```
qqnorm(titanic$Fare,main = "Normal Q-Q Plot para Fare")
qqline(titanic$Fare,col = "red")
```




```
#[1]
ggplot(titanic, aes(x = Fare)) +
  geom_histogram(aes(y = ..density..), bins = 30,
                 colour = 1, fill = "white") +
  geom_density(lwd = 1, colour = 4,
               fill = 4, alpha = 0.25) +
  ggtitle("Distribución y densidad de Fare")
```



Observamos en el gráfico Normal Q-Q, como la mayoría de observaciones no se alinea encima de la línea roja que indica como debería comportarse una variable normal. Este gráfico nos indica que **Fare** no sigue una distribución normal.

Al fijarnos en el histograma de la variable con su curva de densidad, vemos como la mayoría de valores se centran entre 0 y 50, con un pico pronunciado entorno a 10, y decaen casi exponencialmente. La curva de densidad no es simétrica ni tiene forma de campana. Determinamos que no se trata de una variable con distribución normal.

Análisis por tests de normalidad:

```
#[2]
## Shapiro test
shapiro.test(titanic$Fare)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  titanic$Fare
## W = 0.60472, p-value < 2.2e-16
```

```
## Jarque Bera test
jb.norm.test(titanic$Fare)
```

```
##
##  Jarque-Bera test for normality
```

```
##
## data:  titanic$Fare
## JB = 6793.2, p-value < 2.2e-16
```

```
## Forsini test
frosini.norm.test(titanic$Fare)
```

```
##
##  Frosini test for normality
##
## data:  titanic$Fare
## B = 3.9484, p-value < 2.2e-16
```

Para los tests de normalidad Shapiro-Wilk, Jarque-Bera y Frosini, la hipótesis nula indica normalidad, mientras que la hipotesis alternativa indica que no hay normalidad. En todos ellos nos encontramos un valor p muy pequeño, inferior a todos los limites de significancia comunmente usados, por lo que debemos descartar la hipotesis nula de normalidad.

Análisis de homogeniedad de la varianza:

```
leveneTest(y = titanic$Fare, group = titanic$Survived, center = "median")
```

```
## Levene's Test for Homogeneity of Variance (center = "median")
##        Df F value    Pr(>F)
## group   1  44.703 4.055e-11 ***
##       886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Para evaluar la homocedasticidad, o la homogeniedad de la varianza, para la variable **Fare**, elijemos de nuevo el test Levene siendo este menos sensible a que la ausencia de normalidad [3].

Hipótesis nula: Ambas varianzas son iguales. Hipótesis alternativa: Las varianzas son distintas entre si.

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

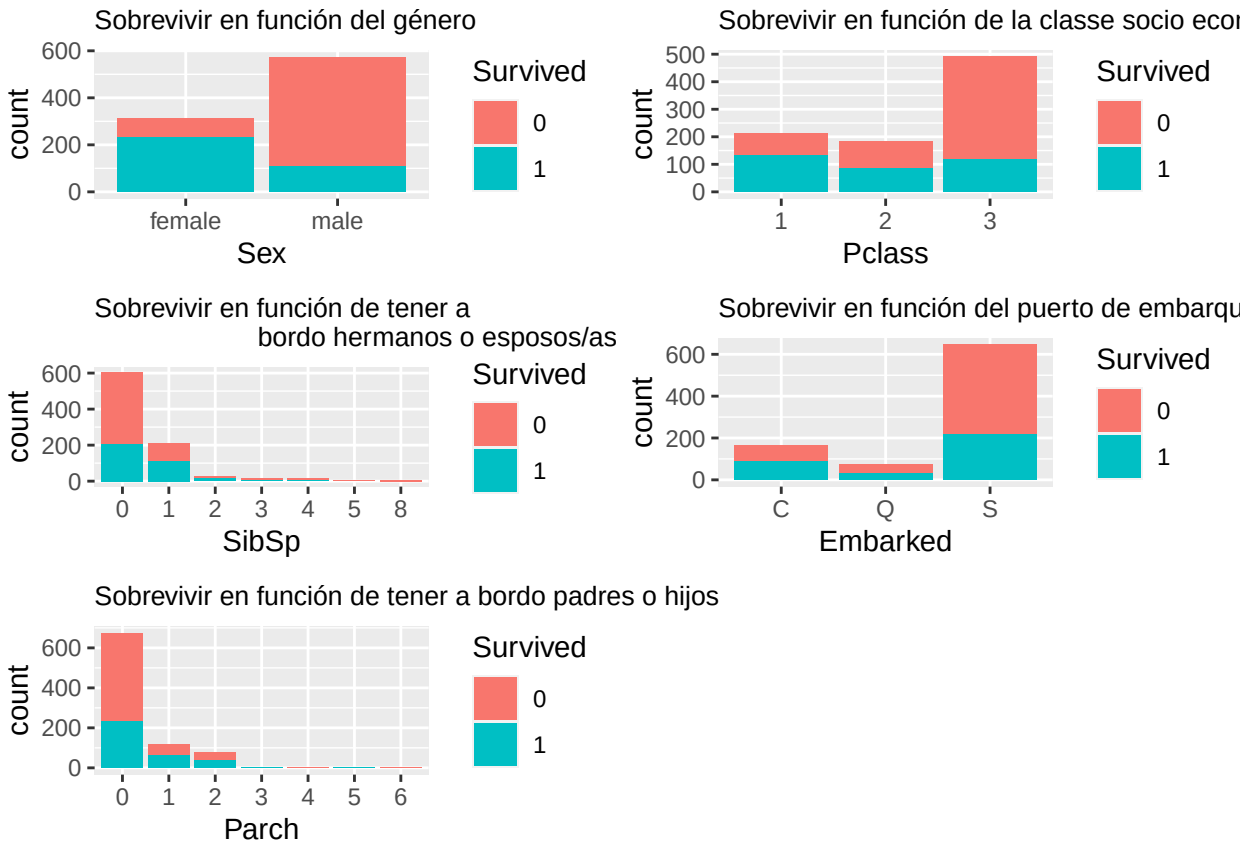
Observamos que al evaluar las varianzas agrupando la variable **Fare** en función de la variable dependiente **Survived**, estas son claramente diferentes entre ellas. Obtenemos un valor p muy pequeño, inferior a todos los límites de significancia comúnmente usados, y por ello debemos descartar las hipótesis nula.

4.3 Análisis

Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

4.3.1 Analisis de la relación entre haber sobrevivido, o no, con las variables categóricas Primeramente analizaremos de forma visual si pueden haber correlaciones entre las variables categóricas del conjunto **Sex**, **Pclass**, **SibSp**, **Parch** y **Embarked** con **Survive**, la variable que nos indica si los pasajeros sobrevivieron al accidente. Para ello utilizaremos un gráfico de frecuencias, que nos permitir al mismo tiempo conocer mejor como se distribuyen estas variables.

```r
# Histogramas
d.Sex <- ggplot(data = titanic, aes(x=Sex, fill=Survived))+geom_bar()+
              ggtitle("Sobrevivir en función del género") +
              theme(plot.title = element_text(size = 10))

d.Pclass <- ggplot(data = titanic, aes(x=Pclass, fill=Survived))+geom_bar()+
              ggtitle("Sobrevivir en función de la classe socio económica") +
              theme(plot.title = element_text(size = 10))

d.SibSp <- ggplot(data = titanic, aes(x=SibSp, fill=Survived))+geom_bar()+
              ggtitle("Sobrevivir en función de tener a
                      bordo hermanos o esposos/as") +
              theme(plot.title = element_text(size = 10))

d.Parch <- ggplot(data = titanic, aes(x=Parch, fill=Survived))+geom_bar()+
              ggtitle("Sobrevivir en función de tener a bordo padres o hijos") +
              theme(plot.title = element_text(size = 10))

d.Embarked <- ggplot(data = titanic, aes(x=Embarked, fill=Survived))+geom_bar()+
                ggtitle("Sobrevivir en función del puerto de embarque") +
                theme(plot.title = element_text(size = 10))

grid.arrange(d.Sex, d.Pclass, d.SibSp, d.Embarked, d.Parch,
             ncol = 2, nrow = 3)
```



#TODO : Comentar el gráfico

#TODO: Introducir bien el Odds ratio y A continuación, se calculará el Odds-Ratio de la variable **Survived** para cada una de las variables categóricas. El OR es un número que permite conocer cómo varía la probabilidad de que la variable dependiente adquiera un valor en función del valor que adquiera la variable independiente. Para calcularlo se crea una función.

Se calcularán los OR de sobrevivir frente a no sobrevivir.

```
# Se muestra la tabla de frecuencia
print(table(titanic$Survived, titanic$Pclass), rownames(c("Not survived", "Survived")))
```

```
##
##       1   2   3
##   0  80  97 372
##   1 133  87 119
```

```
# Se calculan los diferentes Odds-Ratios
a <- as.data.frame(table(titanic$Survived, titanic$Pclass))
odds.high.class <- a[2,]$Freq  / a[1,]$Freq
cat("\nThe odds for the high class is ", odds.high.class, "\n")
```

```
##
## The odds for the high class is  1.6625
```

```
odds.medium.class <- a[4,]$Freq  / a[3,]$Freq
cat("The odds for the medium class is ", odds.medium.class, "\n")
```

```
## The odds for the medium class is  0.8969072
```

```
odds.low.class <- a[6,]$Freq  / a[5,]$Freq
cat("The odds for the low class is ", odds.low.class, "\n\n")
```

```
## The odds for the low class is  0.3198925
```

```
cat("The OR between the high class and the low class was", odds.high.class / odds.low.class, "\n")
```

```
## The OR between the high class and the low class was 5.197059
```

```
cat("The OR between the medium class and the low class was", odds.medium.class / odds.low.class, "\n")
```

```
## The OR between the medium class and the low class was 2.803777
```

```
cat("The OR between the high class and the medium class was", odds.high.class / odds.medium.class, "\n")
```

```
## The OR between the high class and the medium class was 1.853592
```

Analizando los resultados, una persona de clase alta tenía 5,19 veces más probabilidades de sobrevivir que una persona de clase baja.

Una persona de clase media tenía casi el triple de probabilidades de sobrevivir que una persona de clase baja.

Una persona de clase alta tenía casi el doble de probabilidades de sobrevivir que una persona de clase media.

A continuación se calcularán los OR de la variable **Sex**.

```r
# Se muestra la tabla de frecuencia
print(table(titanic$Survived, titanic$Sex), rownames(c("Not survived", "Survived")))
```

```
##
##     female male
##   0     81  468
##   1    232  107
```

```r
# Se calculan los diferentes Odds-Ratios
a <- as.data.frame(table(titanic$Survived, titanic$Sex))
odds.female <- a[2,]$Freq  / a[1,]$Freq
cat("\nThe odds for the female is ", odds.female, "\n")
```

```
##
## The odds for the female is  2.864198
```

```r
odds.male <- a[4,]$Freq  / a[3,]$Freq
cat("The odds for the male is ", odds.male, "\n")
```

```
## The odds for the male is  0.2286325
```

```r
cat("The OR between the female and the male was", odds.female / odds.male, "\n")
```

```
## The OR between the female and the male was 12.52752
```

Una mujer tenía 12,5 veces más probabilidades de sobrevivir que un hombre.

A continuación, para confirmar si hay relación, utilizaremos el test Chi-cuadrado de Pearson [4] relación entre las variables categóricas **Sex**, **Pclass**, **SibSp**, **Parch** y **Embarked** con la variable **Survive** que nos indica si los pasajeros sobrevivieron o no al accidente.

Para el test Chi-cuadrado, las hipótesis son las siguientes:

Hipótesis nula: No existe relación entre ambas variables. Hipótesis alternativa: Existe relación entre las variables.

Un valor p inferior a los valores comunes de significancia ~0.05, nos indica que podemos descartar la hipótesis nula, por lo tanto, existe relación, de lo contrario nos indicaría que no hay relación entre ambas variables.

Sex:

```r
with(titanic, chisq.test(Survived, Sex))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  Survived and Sex
## X-squared = 262.28, df = 1, p-value < 2.2e-16
```

Hay relación.

Pclass:

```
with(titanic, chisq.test(Survived, Pclass))
```

```
##
##  Pearson's Chi-squared test
##
## data:  Survived and Pclass
## X-squared = 100.03, df = 2, p-value < 2.2e-16
```

Hay relación.

SibSp:

```
with(titanic, chisq.test(Survived, SibSp))
```

```
## Warning in chisq.test(Survived, SibSp): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  Survived and SibSp
## X-squared = 37.981, df = 6, p-value = 1.133e-06
```

TODO: Este error proviene de: https://stats.stackexchange.com/questions/81483/warning-in-r-chi-squared-approximation-may-be-incorrect

Warning causado por categorias con muy pocas observaciones.

Parch:

```
with(titanic, chisq.test(Survived, Parch))
```

```
## Warning in chisq.test(Survived, Parch): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  Survived and Parch
## X-squared = 27.665, df = 6, p-value = 0.0001087
```

Embarked:

```
with(titanic, chisq.test(Survived, Embarked))
```

```
##
##  Pearson's Chi-squared test
##
## data:  Survived and Embarked
## X-squared = 23.755, df = 2, p-value = 6.944e-06
```

Hay relación.

4.3.2 Analisis de la relación entre haber sobrevivido, o no, con las variables numéricas Para estudiar si existe relación entre las variables numéricas **Age** y **Fare** con la variable **Survived**, se crea un modelo de regresión logística para cada una de las variables.

Age:

```
model.age <- glm(Survived ~ Age, family = binomial(link=logit), data=titanic)
summary (model.age)
```

```
##
## Call:
## glm(formula = Survived ~ Age, family = binomial(link = logit),
##     data = titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1127  -0.9827  -0.9383   1.3599   1.6501
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.142468   0.172296  -0.827   0.4083
## Age         -0.011535   0.005403  -2.135   0.0328 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1180.9  on 887  degrees of freedom
## Residual deviance: 1176.3  on 886  degrees of freedom
## AIC: 1180.3
##
## Number of Fisher Scoring iterations: 4
```

#TODO: Reescrivir -> Existe relción ya que el p-valor es menor que el nivel de significancia.

Fare:

```
model.fare <- glm(Survived ~ Fare, family = binomial(link=logit), data=titanic)
summary (model.fare)
```

```
##
## Call:
## glm(formula = Survived ~ Fare, family = binomial(link = logit),
##     data = titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4899  -0.8885  -0.8531   1.3458   1.5941
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.941130   0.095192  -9.887  < 2e-16 ***
## Fare         0.015189   0.002236   6.794 1.09e-11 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1180.9  on 887  degrees of freedom
## Residual deviance: 1117.6  on 886  degrees of freedom
## AIC: 1121.6
##
## Number of Fisher Scoring iterations: 4
```

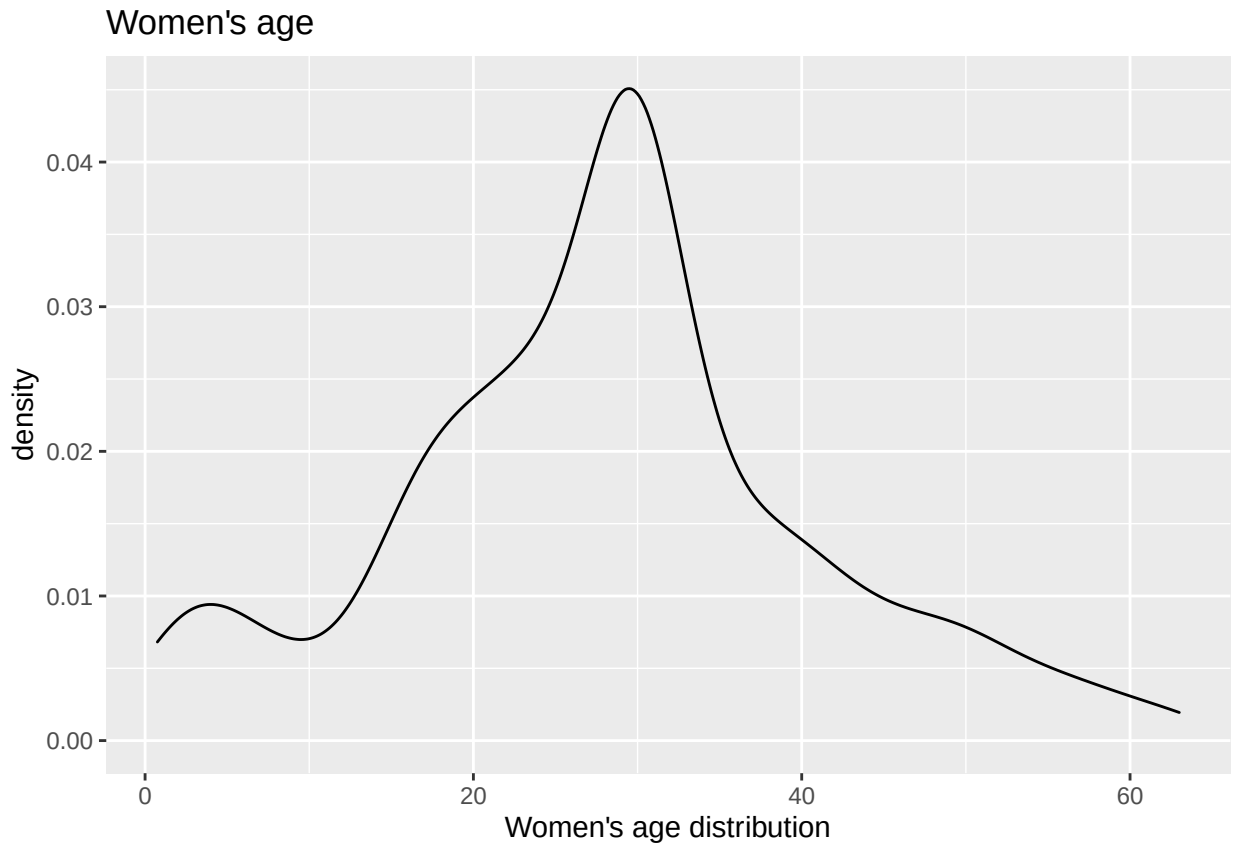
#TODO: comentar

4.3.3 ¿La media de edad de las mujeres es igual o inferior a la de los hombres?

Nos preguntamos si la media de edad de hombres es mayor que la de las mujeres.

Análisis visual #TODO: Agrupar ambas densidades men/wmn en un mismo plot

```
ggplot(data=titanic.women, mapping= aes(x=Age)) + geom_density() + scale_x_continuous("Women's age dist
```

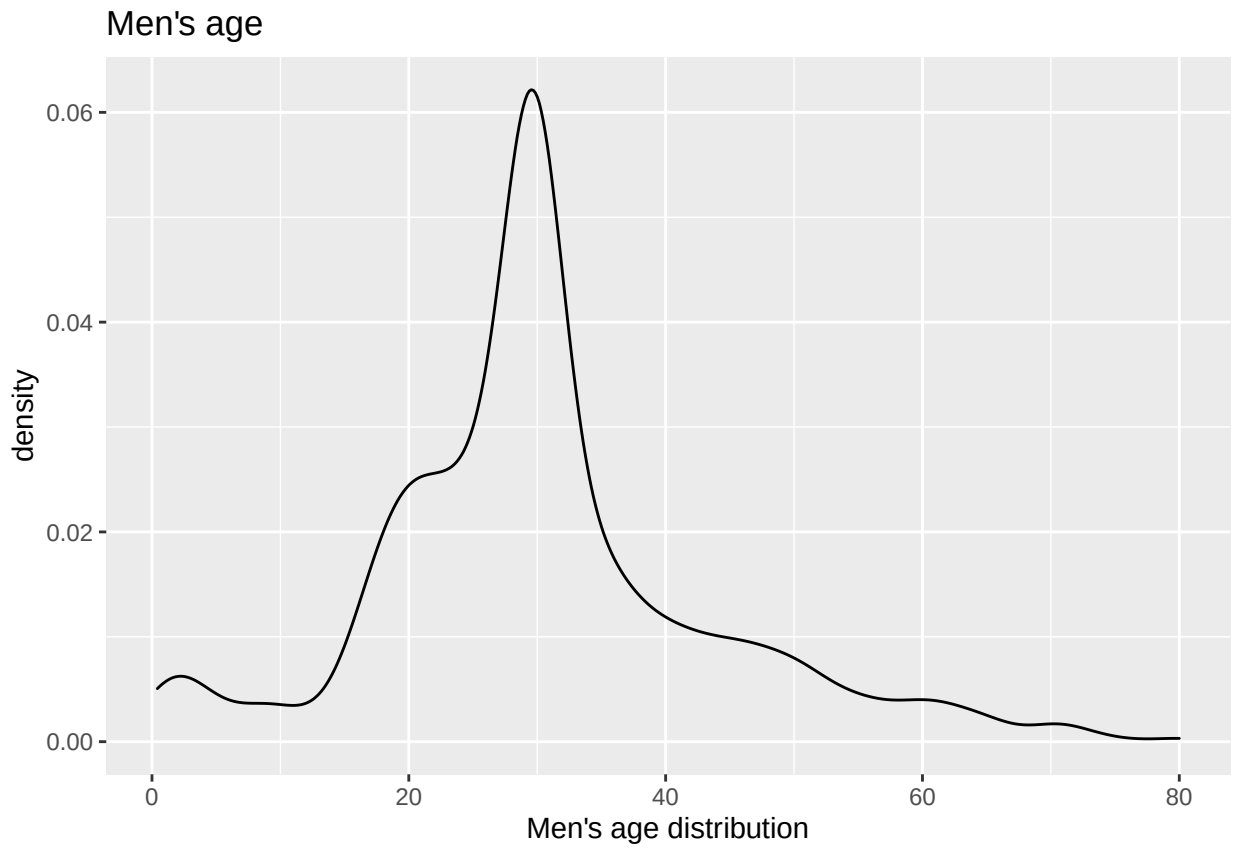



```
shapiro.test(titanic.women$Age)
```

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  titanic.women$Age
## W = 0.97511, p-value = 3.009e-05
```

```
ggplot(data=titanic.men, mapping= aes(x=Age)) + geom_density() + scale_x_continuous("Men's age distribu
```



#TODO: Comentar. ????Aunque las muestras no presentan distribuciones normales, aplicando el Teorema del Límite Central se puede concluir que las medias muestrales siguen una distribución normal

El test de contraste que se aplicará en este caso es el de un contrate unilateral de dos muestras independientes sobre la media con varianzas desconocidas. Para saber si las varianzas son diferentes, se debe realizar un test de homoscedasticidad.

```
var.test(titanic.women$Age, titanic.men$Age)
```

```
##
##  F test to compare two variances
##
## data:  titanic.women$Age and titanic.men$Age
## F = 0.97918, num df = 312, denom df = 574, p-value = 0.8404
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.807806 1.193718
## sample estimates:
## ratio of variances
##           0.979184
```

Como el p-valor es tan alto, no se puede rechazar la hipótesis nula de que las varianzas son diferentes, por lo tanto el test seguirá una distribución *t-student* con $n_1 + n_2 - 2$ grados de libertad.

#TODO: Plantear que hacemos un contraste de hipotesis.

Pregunta: La media de edad de las mujeres es igual o inferior a la de los hombres?

Hipótesis nula: $H_0 : \mu_W = \mu_M$

Hipótesis alternativa: $H_1 : \mu_W < \mu_M$

????Es un test paramétrico ya que los datos siguen una distribución normal.

```
t.test(titanic.women$Age, titanic.men$Age, alternative = "l", var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  titanic.women$Age and titanic.men$Age
## t = -2.5152, df = 886, p-value = 0.006036
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.7920278
## sample estimates:
## mean of x mean of y
##  28.19506  30.48845
```

Como el p-valor es menor que el nivel de significancia, se rechaza la hipótesis nula y se acepta que la media de edad de las mujeres fue menor que la de los hombres.

4.4 Modelo de regresión logística

A continuación, se irán añadiendo variables independientes al modelo de regresión logística para tratar de obtener el modelo que mejor prediga el dataset test.

En primer lugar, se divide el dataset en un dataset de train y otro de test para evaluar los modelos con una proporción 3 a 1.

```
set.seed(23)

trainIndex=createDataPartition(titanic$Survived, p=0.75)$Resample1

titanic_train=titanic[trainIndex, ]
titanic_test= titanic[-trainIndex, ]

cat("The train set has", nrow(titanic_train), "rows\n\n")
```

```
## The train set has 667 rows
```

```
cat("The train set has", nrow(titanic_test), "rows")
```

```
## The train set has 221 rows
```

Para evaluar los modelos, se calcula su precisión a partir del dataset de test.

En primer lugar, se calcula la precisión con las variables por separado.

```
for (var in colnames(titanic_train)){
  if (var != "Survived"){
    formula <- as.formula(sprintf("Survived ~ %s", var))
    model <- glm(formula, family = binomial(link=logit), data=titanic_train)

    titanic_test$pred <- predict(model, titanic_test, type = "response")
    titanic_test$pred_d <- ifelse(titanic_test$pred < 0.5, 0, 1)
    titanic_test$is_ok <- ifelse(titanic_test$Survived == titanic_test$pred_d,
                                 1, 0)
    accuracy <- sum(titanic_test$is_ok)/nrow(titanic_test)
    cat("The accuracy of the model with ", var, " is ", accuracy, "\n")
  }
}
```

```
## The accuracy of the model with  Pclass  is  0.7104072
## The accuracy of the model with  Sex  is  0.7692308
## The accuracy of the model with  Age  is  0.6199095
## The accuracy of the model with  SibSp  is  0.6742081
## The accuracy of the model with  Parch  is  0.5972851
## The accuracy of the model with  Fare  is  0.6968326
## The accuracy of the model with  Embarked  is  0.6334842
```

A continuación, se van añadiendo las variables con mejores resultados al modelo.

```
for (var in colnames(titanic_train)){
  if (var != "Sex" && var != "Survived"){
    formula <- as.formula(sprintf("Survived ~ Sex * %s", var))
    model <- glm(formula, family = binomial(link=logit), data=titanic_train)

    titanic_test$pred <- predict(model, titanic_test, type = "response")
    titanic_test$pred_d <- ifelse(titanic_test$pred < 0.5, 0, 1)
    titanic_test$is_ok <- ifelse(titanic_test$Survived == titanic_test$pred_d,
                                 1, 0)
    accuracy <- sum(titanic_test$is_ok)/nrow(titanic_test)
    cat("The accuracy of the model with Sex and ", var, " is ", accuracy, "\n")
  }
}
```

Sex y el resto de variables

```
## The accuracy of the model with Sex and  Pclass  is  0.7692308
## The accuracy of the model with Sex and  Age  is  0.7692308
## The accuracy of the model with Sex and  SibSp  is  0.7782805
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
```

```
## The accuracy of the model with Sex and  Parch  is  0.7737557
## The accuracy of the model with Sex and  Fare  is  0.7692308
## The accuracy of the model with Sex and  Embarked  is  0.7692308
```

Las variables que mejores resultados muestran son **Parch** y **SibSp**. Como la variable **Parch** muestra un warning de correlación, se decide incluir la variable **SibSp** en el modelo.

```
for (var in colnames(titanic_train)){
  if (var != "Sex" && var != "Survived" && var != "SibSp" && var != "Parch"){
    formula <- as.formula(sprintf("Survived ~ Sex * SibSp + %s", var))
    model <- glm(formula, family = binomial(link=logit), data=titanic_train)

    titanic_test$pred <- predict(model, titanic_test, type = "response")
    titanic_test$pred_d <- ifelse(titanic_test$pred < 0.5, 0, 1)
    titanic_test$is_ok <- ifelse(titanic_test$Survived == titanic_test$pred_d,
                                 1, 0)
    accuracy <- sum(titanic_test$is_ok)/nrow(titanic_test)
    cat("The accuracy of the model with Sex, SibSp and ", var, " is ", accuracy, "\n")
  }
}
```

Sex, SibSp y el resto de variables

```
## The accuracy of the model with Sex, SibSp and  Pclass  is  0.7828054
## The accuracy of the model with Sex, SibSp and  Age  is  0.7782805
## The accuracy of the model with Sex, SibSp and  Fare  is  0.7782805
## The accuracy of the model with Sex, SibSp and  Embarked  is  0.7782805
```

Se hace el estudio sin las interacciones, ya que la gran cantidad de términos de correlación harían que los modelos generados no fuesen fiables.

Se decide añadir ahora la variable **Pclass**.

```
for (var in colnames(titanic_train)){
  if (var != "Sex" && var != "Survived" && var != "SibSp" && var != "Parch"
      && var != "Pclass"){
    formula <- as.formula(sprintf("Survived ~ Sex * SibSp + Pclass * %s", var))
    model <- glm(formula, family = binomial(link=logit), data=titanic_train)

    titanic_test$pred <- predict(model, titanic_test, type = "response")
    titanic_test$pred_d <- ifelse(titanic_test$pred < 0.5, 0, 1)
    titanic_test$is_ok <- ifelse(titanic_test$Survived == titanic_test$pred_d,
                                 1, 0)
    accuracy <- sum(titanic_test$is_ok)/nrow(titanic_test)
    cat("The accuracy of the model with Sex, SibSp, Pclass and ", var, " is ",
        accuracy, "\n")
  }
}
```

Sex, SibSp, Pclass y el resto de variables.

```
## The accuracy of the model with Sex, SibSp, Pclass and  Age  is  0.8099548
## The accuracy of the model with Sex, SibSp, Pclass and  Fare  is  0.7828054
## The accuracy of the model with Sex, SibSp, Pclass and  Embarked  is  0.7828054
```

La variable **Age** muestra los mejores resultados.

```
for (var in colnames(titanic_train)){
  if (var != "Sex" && var != "Survived" && var != "SibSp" && var != "Parch"
      && var != "Pclass" && var != "Age"){
    formula <- as.formula(sprintf("Survived ~ Sex * SibSp + Pclass * Age * %s",
                                  var))
    model <- glm(formula, family = binomial(link=logit), data=titanic_train)

    titanic_test$pred <- predict(model, titanic_test, type = "response")
    titanic_test$pred_d <- ifelse(titanic_test$pred < 0.5, 0, 1)
    titanic_test$is_ok <- ifelse(titanic_test$Survived == titanic_test$pred_d,
                                 1, 0)
    accuracy <- sum(titanic_test$is_ok)/nrow(titanic_test)
    cat("The accuracy of the model with Sex, SibSp, Pclass, Age and ", var,
        " is ", accuracy, "\n")
  }
}
```

Sex, SibSp, Pclass, Age y el resto de variables.

```
## The accuracy of the model with Sex, SibSp, Pclass, Age and  Fare  is  0.7963801
## The accuracy of the model with Sex, SibSp, Pclass, Age and  Embarked  is  0.7918552
```

La precisión de los modelos es menor. Se decide que el modelo final sea el del apartado anterior.

```
model.final <- glm(Survived ~ Sex * SibSp + Pclass * Age, family = binomial(link=logit), data=titanic_t
summary(model.final)
```

```
##
## Call:
## glm(formula = Survived ~ Sex * SibSp + Pclass * Age, family = binomial(link = logit),
##     data = titanic_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5545  -0.6092  -0.4380   0.6159   2.4114
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     3.535e+00  6.130e-01   5.767 8.08e-09 ***
## Sexmale        -2.830e+00  2.721e-01 -10.401  < 2e-16 ***
## SibSp1         -2.487e-01  3.548e-01  -0.701  0.48347
## SibSp2         -5.191e-01  7.559e-01  -0.687  0.49228
```

```
## SibSp3         -1.836e+00  8.142e-01  -2.255  0.02412 *
## SibSp4         -6.150e-01  1.287e+00  -0.478  0.63279
## SibSp5         -1.762e+01  2.400e+03  -0.007  0.99414
## SibSp8         -1.718e+01  1.697e+03  -0.010  0.99192
## Pclass2        -5.431e-01  8.032e-01  -0.676  0.49893
## Pclass3        -1.959e+00  7.288e-01  -2.689  0.00717 **
## Age            -3.133e-02  1.388e-02  -2.258  0.02396 *
## Sexmale:SibSp1  3.045e-01  4.909e-01   0.620  0.53504
## Sexmale:SibSp2 -1.111e-01  1.325e+00  -0.084  0.93317
## Sexmale:SibSp3 -1.366e+01  1.006e+03  -0.014  0.98916
## Sexmale:SibSp4  7.214e-02  1.655e+00   0.044  0.96523
## Sexmale:SibSp5  2.540e+00  2.935e+03   0.001  0.99931
## Sexmale:SibSp8  2.830e+00  2.190e+03   0.001  0.99897
## Pclass2:Age    -1.871e-02  2.275e-02  -0.823  0.41076
## Pclass3:Age    -1.204e-03  2.114e-02  -0.057  0.95458
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 887.35  on 666  degrees of freedom
## Residual deviance: 589.10  on 648  degrees of freedom
## AIC: 627.1
##
## Number of Fisher Scoring iterations: 15
```

```
titanic_test$pred <- predict(model.final, titanic_test, type = "response")
titanic_test$pred_d <- ifelse(titanic_test$pred < 0.5, 0, 1)
titanic_test$is_ok <- ifelse(titanic_test$Survived == titanic_test$pred_d, 1, 0)
accuracy.final <- sum(titanic_test$is_ok)/nrow(titanic_test)
cat("The accuracy of the final model is ", accuracy.final)
```

```
## The accuracy of the final model is  0.8099548
```

Aunque haya varios términos no significativos, la precisión del modelo final es la más alta de entre todos los que se han estudiado.

Ejercicio 5

Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

#TODO: - Como afectan las variables a la probabilidad de sobrevivir o no. - Comentar resultados relevantes del estudio para crear un modelo de regresión.

Referencias

[1] Histogram with density in ggplot2 [fecha de consulta: 30/05/22] Disponible en: https://r-charts.com/distribution/histogram-density-ggplot2/

[2] Sigüeñas Gonzales, Manuel. Pruebas de Normalidad [fecha de publicación: 28/10/2015] [fecha de consulta: 30/05/22] Disponible en: https://rpubs.com/MSiguenas/122473

[3] Amat Rodrigo, Joaquín. Análisis de la homogeneidad de varianza (homocedasticidad)[fecha de publicación: 01/01/2016] [fecha de consulta: 30/05/22] Disponible en: https://rpubs.com/Joaquin_AR/218466

[4] The Chi-Square Test for Independence [fecha de consulta: 30/05/22] Disponible en: https://soc.utah.edu/sociology3112/chi-square.php