Titanic Survivors Analysis



*EM PROJECT*

**TITANIC SURVIVORS:**

**A STATISTICAL ANALYSIS**

*Elisabetta Seggioli 11032*

1. **Data Description**

* **Data type :** Titanic passengers
* **Variable description: pclass** : Passenger Class

**survival**: Survival (0 = No ; 1 = Yes)

**name**: Name

**sex**: Sex

**age**: Age

*The sinking of the RMS Titanic (14 April – 15 April 1912) in the north Atlantic Ocean was a real tragedy. It was the largest passenger line in service at the time and it had an estimated 2.224 people on board when it hit an iceberg at 23:40.*

1. **Aim of the project**

The project ultimate aim is to come up with an analytic model to predict survival in Titanic Disaster using the tool presented during the course. More precisely, to determine the key factor (age, class, sex) that had a greater impact on the survival.

**3. First Phase: First Overview of the data – Basic Features**

The first part of the project consists in loading the table with the data containing the personal data and the information about the survival.

**Class Survived** **Name**  **Sex** **Age**

1 1 Allen, Miss. Elisabeth Walton female 29

1 1 Allison, Master. Hudson Trevor male 1

1 0 Allison, Miss. Helen Loraine female 2

1 0 Allison, Mr. Hudson Joshua Creighton male 30

1 0 Allison, Mrs. Hudson J C female 25

*Example of the data contained in the table.*

\*Total number of people on board: 1310 (the crew members are not counted)

Female Male

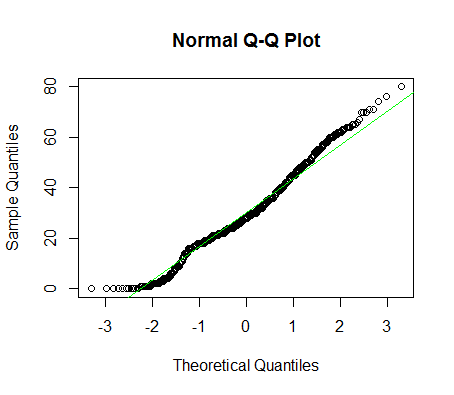
466 843

Before starting with the real analysis, the missing data in the age column have been replaced with the mean value of all the available ages to give uniformity to the data.

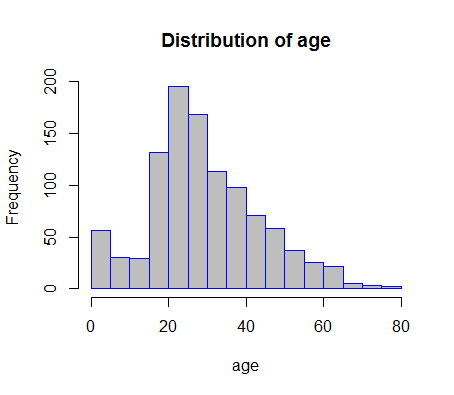
Then, once the data are loaded and stored in a variable called “titanic1”, a basic analysis of the numerical data contained in the age column can be started using the most common statistical tool. In more details, mean, median, standard deviation, variance and distribution are calculated.

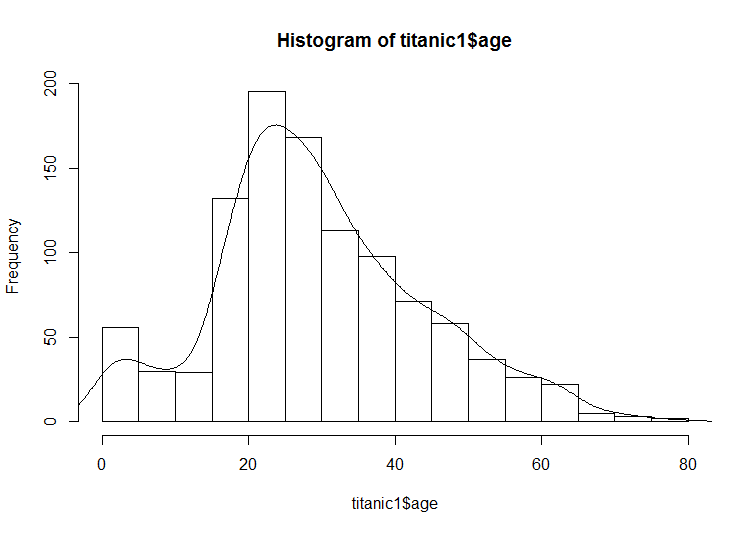
**R RESULTS TABLE AND GRAPHS**

|  |  |
| --- | --- |
| **OPERATIONS** | **RESULTS** |
| mean | 29.8689 |
| median | 28 |
| Standard deviation | 14.42928 |
| variance | 208.2041 |

**

*Plot of the age distribution using the “qqnorm” and “qqlin” R commands.*

**

**

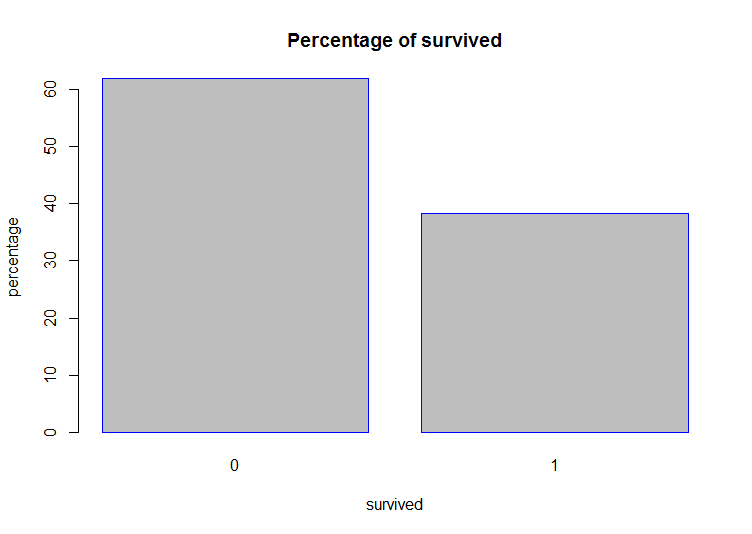
*Histogram using as x-axis the years and as y-axis the number of person having that particular age (age frequency).*

The two graphs shows that the age distribution is an approximately normal data slightly skewed to the right. Moreover, it can be noticed that the higher age frequency is in the interval between 20 and 30; the median and the mean, respectively 29.86 and 28 also confirm this assumption. Furthermore, it is of particular interest to notice that the number of children is higher than the number of seniors on boards.

1. **Second Phase: Checking Correlations**

First, an analysis of the total number of survivors expressed in percentage will be executed, that will be, afterward, useful to check the correlations between survival and the different factors.

|  |  |
| --- | --- |
| 1 | Yes |
| 0 | No |



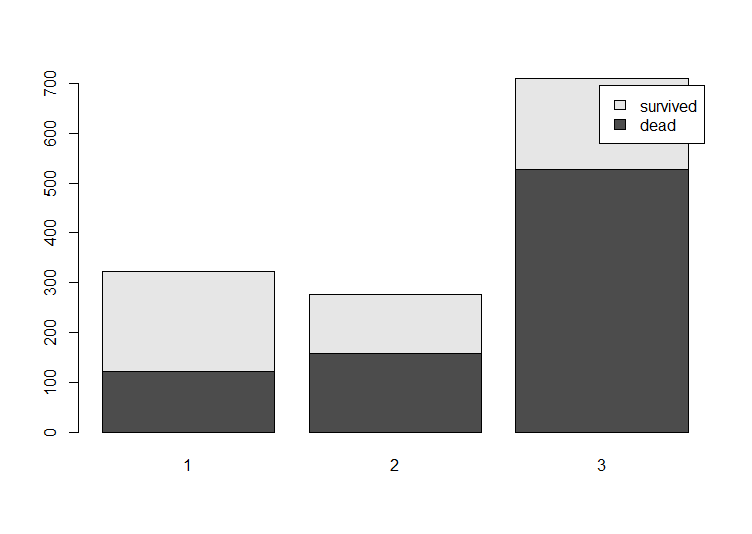
*Number of survived in percentage. The total number of survivors was 500 (38,2 %) against the 809 (61.8%) that have died.*

**4.1 Correlation between class and survival**

After having had an overview of the total number of survivors, the first point of the statistical analysis to check is the correlation between the class and the survival, proving or disproving the hypothesis that the higher number of survived came from the first class.

**PASSENGERS FATE BY TRAVELLING CLASS GRAPH**

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1st | 2nd | 3rd |
| NO | 123 | 158 | 528 |
| YES | 200 | 119 | 181 |

**

*Bar plot showing the proportion of the survivors for each class.*

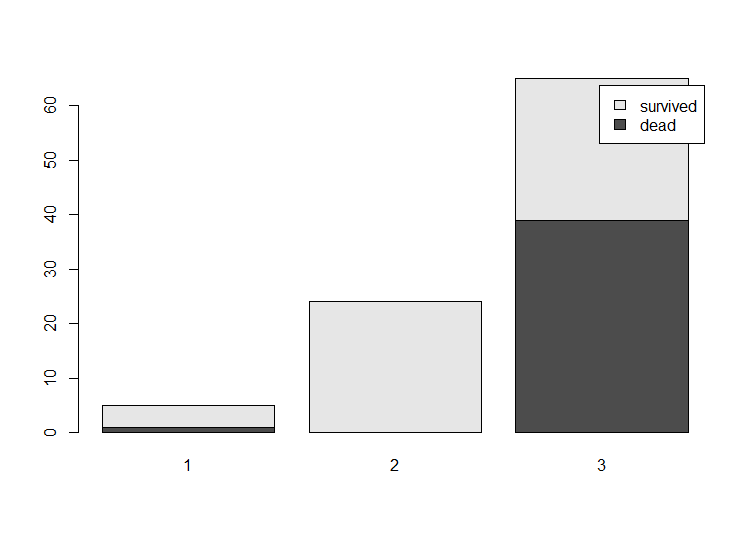
As the previous hypothesis said, this graph confirms the alternative hypothesis that the class has a weight in the survival rate. More precisely, the higher the class, the lower the number of survived. In reality, the difference between the first and the second class is not so net if compared with the third class that register the higher number of victims. The two factors can be considered inversely proportional. An explanation can be found in the distance of the different class from the part of the Titanic, in which the lifeboats were positioned. More precisely, the first class was not only the most advantaged for the social status, but also for the position, indeed it was the nearest class to the lifeboats.

* 1. **Correlation between age and survival**

Once the correlation between class and survival has been checked, the next step will be controlling the correlation, this time, between age and survival. To do this, since age is a continuous variable, it is needed to change it into a so-called “categorical” one, establishing age’s ranges. For example a first range from 0 – 12, a second from 12 – 30 and greater than 30. This is necessary to check the hypothesis that the passengers below 12 years, the children, had greater chance of surviving than adults. In addition, it can useful in this case, to use a one-sided test to examine only the part of the data that have less than 10 years.

**ANALYSIS TAKING INTO ACCOUNT ONLY CHILDREN**

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1st | 2nd | 3rd |
| NO | 1 | 0 | 39 |
| YES | 4 | 24 | 26 |

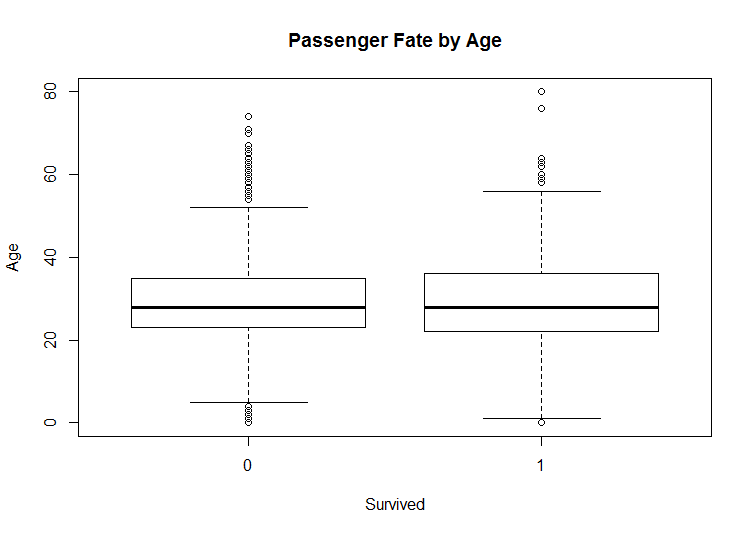
**

*Plot of the survival rate locking only to the children on board of titanic.*

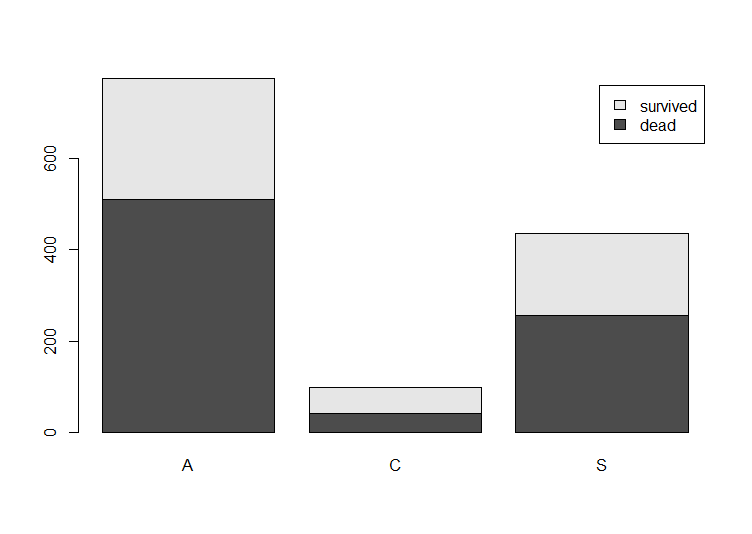
It can be easily observe that the majority of the children have survived, this still considering the difference between classes. Moreover, worthy to be noticed is that, although the great survivor’s number of the overall titanic passengers is registered in the first class. For what concerns children, the class with the higher number of survivors is the second class in which 100 % of the children have survived. Finally, regarding the third class, comparing the previous bar chart and the last one, the number of children survived is higher than the number of third class passengers.

To better examine the age survival, it is useful, as already mentioned, to divide the age by categories in which C stand for children, A for adults and S for seniors. For the age a box plot and a bar plot will be created using the r commands.

|  |  |  |  |
| --- | --- | --- | --- |
|  | A | C | S |
| NO | 511 | 42 | 256 |
| YES | 264 | 57 | 179 |



*Box plot of the age survival divided for categories.*

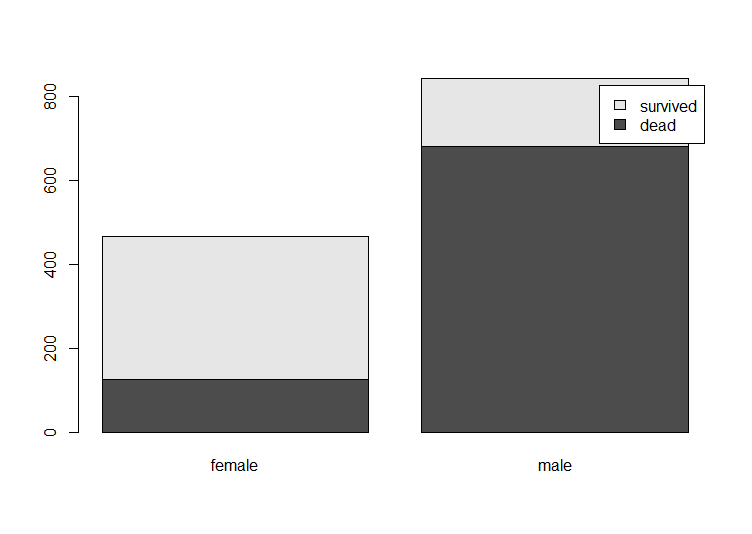
**

*Bar plot of the ages divided by categories.*

The bar plot and box plot confirm that the greater proportion between dead and survived can be found in the children bar where 57 % have survived the tragedy against the 34 % in the adults’ bar and 41 % in the seniors bar.

**4.1 Correlation between sex and survival**

Last point to take into consideration is, verify the correlation between the sex and the rate of survival, in this case, proving or disproving the null hypothesis that the sex had no impact on the survival or the alternative hypothesis that the sex was a decisive factor.



*Bar plot showing the correlation between sex and survival*

As guessed the higher rate of survival can be register in the female bar, where the survival is double respecting the male column, which has the lower number of survived.

1. **CONCLUSIONS**

From this analysis, it can be inferred that the least likely to survive to the sink of the titanic were men, while children and seniors had greater chance. Moreover, the class played also an important role , allowing the first class to place first in the surviving list. Furthermore, women were advantages respect men as can be observed by the previous bar chart.

1. **R CODE SOURCE**

titanic1 <- read.csv("C:/Users/Elisabetta/Desktop/EM/titanic3.csv",header= TRUE) #load the csv table containing data

# 1. FIRST OVERVIEW OF THE DATA - Basic Features

# AGE

titanic1$age[is.na(titanic1$age)] <- median(titanic1$age, na.rm=T) #mean value of all the available ages to replace NA

age= titanic1$age

qqnorm(age)

qqline(age,col="green") #testing the normality

mean(titanic1$age,na.rm=TRUE) #calculate the mean

median(titanic1$age,na.rm=TRUE) #calculate the median

sd(titanic1$age,na.rm=TRUE) #calculate the standard deviation

var(titanic1$age,na.rm=TRUE) #calculate the variance

utils::str(hist(titanic1$age, main="Distribution of age",

xlab="age", col = "gray", border="blue"))#age distribution

myhist <- hist(titanic1$age, main="Distribution of age",

xlab="age", col = "gray", border="blue")

multiplier <- myhist$counts / myhist$density

mydensity <- density(titanic1$age,na.rm=TRUE)

mydensity$y <- mydensity$y \* multiplier[1]

plot(myhist)

lines(mydensity) # plot age distribution

#SEX

summary(titanic1$sex) #count the number of female and male

#SURVIVAL

table(titanic1$survived)

table(titanic1$survived)-> survived

prop.table(survived)\*100 ->psurvived #calculate percentage of survived

round(psurvived, digits=1)->psurvived

barplot(psurvived, main="Percentage of survived",

xlab="survived", ylab="percentage",col = "gray", border="blue") #plotting the percentage of survived

#CLASS

table(titanic1$pclass) #number of passengers for each class

table(titanic1$pclass) -> pclass

prop.table(pclass)\*100 -> class #percentage of passengers for each class

round(class, digits=1) ->class

class

#SUMMARY OF sex, age and class

table (titanic1[,c("survived","sex","pclass")])

# 2. CHECK CORRELATION BETWEEN CLASS AND SURVIVAL

mosaicplot(titanic1$pclass ~ titanic1$survived,

main="Passenger Fate by Traveling Class", shade=FALSE,

color=TRUE, xlab="Pclass", ylab="Survived")

table(titanic1[,c("survived","pclass")]) #comparison between class and survival using table

barplot(table(titanic1[,c("survived","pclass")]),legend.text =c("dead","survived")) #plotting the previous table using barchar

# 3. CHECK CORRELATION BETWEEN AGE AND SURVIVAL

boxplot(titanic1$age ~ titanic1$survived,

main="Passenger Fate by Age",

xlab="Survived", ylab="Age")

# 3.1 ANALYSIS TAKING INTO ACCOUNT ONLY CHILDREN

titanic1.children <- titanic1[which (titanic1$age < 13),] #select all the children

table(titanic1.children[,c("survived","pclass")]) #number of children survived

barplot(table(titanic1.children[,c("survived","pclass")]),legend.text =c("dead","survived"))

# 3.2 DIVIDE DATA INTO CATEGORIES

# C = "children", A= "adults" and S = "senior"

test1 <- titanic1

test1$age[test1$age<=13] <- "C"

test1$age[test1$age>13 & test1$age<=30 & test1$age!="C"] <- "A"

test1$age[test1$age>30 & test1$age!="C" & test1$age!="A"] <- "S"

table(test1[,c("survived","age")])

barplot(table(test1[,c("survived","age")]),legend.text =c("dead","survived"))

# 4. CHECK CORRELATION BETWEEN SEX AND SURVIVAL

table(titanic1[,c("survived","sex")]) #comparison between and survival sex using table

barplot(table(titanic1[,c("survived","sex")]),legend.text =c("dead","survived"))