MA5701 R02 Summer 2020

Statistical Summary
of
Titanic Dataset

Date: 08/16/2020

Student names: Prateek Goyal

Submitted to Dr. Ray Molzon in partial fulfillment for the course MA5701 Statistical Methods

Michigan Tech

Contents

Table of Figures

Introduction

If you have ever been on a large boat or watched a movie that involved a shipwreck, then you might have heard the saying “women and children first”. This saying refers to who will be saved in the event of a life-threatening situation. A life-threatening situation could mean any dangerous situation such as a house fire or abandoning ship when resources such as lifeboats are limited. But how did this rule of thumb come to be? The phrase "women and children first" was initially associated with the sinking of HMS Birkenhead in 1852 but was subsequently also closely associated with the Titanic.

The Titanic is arguably one of the most famous shipwrecks of all time. In 1912, a large passenger ship struck an iceberg. The ship did not have enough lifeboats for each passenger to be saved. This shortage of lifeboats is what proposed the question: who exactly got a spot on a lifeboat? Was it a complete free for all, or was there some sort of order in deciding who received a spot? This report will analyze which passengers survived the shipwreck based on class, age, and sex. We are interested in finding the relationships between these variables in order to get a better idea of how they chose who was allowed on the lifeboats. The ‘Sex’ variable will be compared to ‘Age’ to determine if “women and children” were in fact the first to receive a spot on the lifeboat. The ‘Passenger class’ can also suggest what parts of the boat are affected when considering the number of people survived. Along with all these observations, some of the target interests that will be highlighted is which group of passengers was most likely to be saved, as well as how important socioeconomic class was in determining who was allowed on a lifeboat. Our initial hypothesis is that younger children and women will in fact have a larger survival rate compared to the remaining population.

Methods

For our experiment, the **sampling unit** would be every passenger traveled on the Titanic ship. The **sampling scheme** is going to be random for our hypothesis.

As we are planning to conduct 3 hypothesis testing based on Gender, Pclass, and Age and we have chosen the below-mentioned sample scheme for our hypothesis.

For the 1st Gender-based hypothesis, we are planning to follow the Stratified Random sampling scheme as a purely random sample have a little probability of having disproportion of an equal number of gender category and may introduce little bias. For this hypothesis, we will be considering 100 samples from the entire population with 50 females and 50 males each to avoid any disproportion-based bias.

For the 2nd Pclass-based hypothesis, we will follow the same approach as used in the 1st hypothesis.

For the 3rd Age-based hypothesis, we can use a random sampling scheme here as we do not have any Age-based category to deal with and the random scheme will work fine here for 100 samples from the population.

We will be having overall 12 variables in our dataset (mentioned below) with corresponding information about the measured unit/ or assigning/observing levels to each sampling unit.

PassengerID (Categorical): This is going to the unique ID to distinguish all the Passengers from one another.

Survived (Categorical): This is the dependent variable for our experiment. This variable contains the following 2 levels.

- 0 -> Not Survived
- 1 -> Survived

Pclass (Categorical): This variable is going to be one of the independent variables for our experiment. This variable contains the following 3 levels.

- 1 -> 1st Class
- 2 -> 2nd Class
- 3 -> 3rd Class

Name (Categorical): This variable represents the Name of the Passengers. As this variable won't help much for our experiment, hence this will be discarded along with other few variables from our experiment.

Sex (Categorical): This variable is going to be another independent variable for our experiment. This variable contains the following 2 levels.

- Male -> Gender category 1
- Female -> Gender category 2

Age (Numeric): This variable is going to be the final independent variable for our experiment. It is a continuous variable and measured in the 'Years' scale/unit.

SibSp (Numeric): This is one of the variables among the area of interest and contains a quantitative value in terms of integers ranging from 0 to 8.

Parch (Numeric): Again, this is also another variable among areas of interest and contains a quantitative value in terms of integers ranging from 0 to 9.

Ticket (Categorical): This variable represents the Ticket number of the Passenger and hence, discarded from our experiment.

Fare (Numeric): This variable represents the continuous variable and measured in Pounds.

Cabin (Categorical): This is also one of the variables among the area of interest and we can fetch Cabin information from the values. Ex: C103 situated in the 'C' section and so on.

Embarked (Categorical): This represents the city name and has the following 3 levels/types:

- S -> Southampton
- C -> Cherbourg
- Q -> Queenstow

The Titanic dataset is readily available over the internet, but we are going to download our dataset from a very reliable source i.e. kaggle.com, a subsidiary of Google LLC.

Link: https://www.kaggle.com/c/titanic/data/

I believe the source collected the data from the company who issued the Titanic ship ticket to the customers where customers must have had filled their details mentioned above.

One challenge after getting the dataset is that we have approx. 1309 data points (passengers) in our dataset and the dataset have some missing values which need to be handled before experimenting, hence the following operations need to be carried out to make data ready for the experiment.

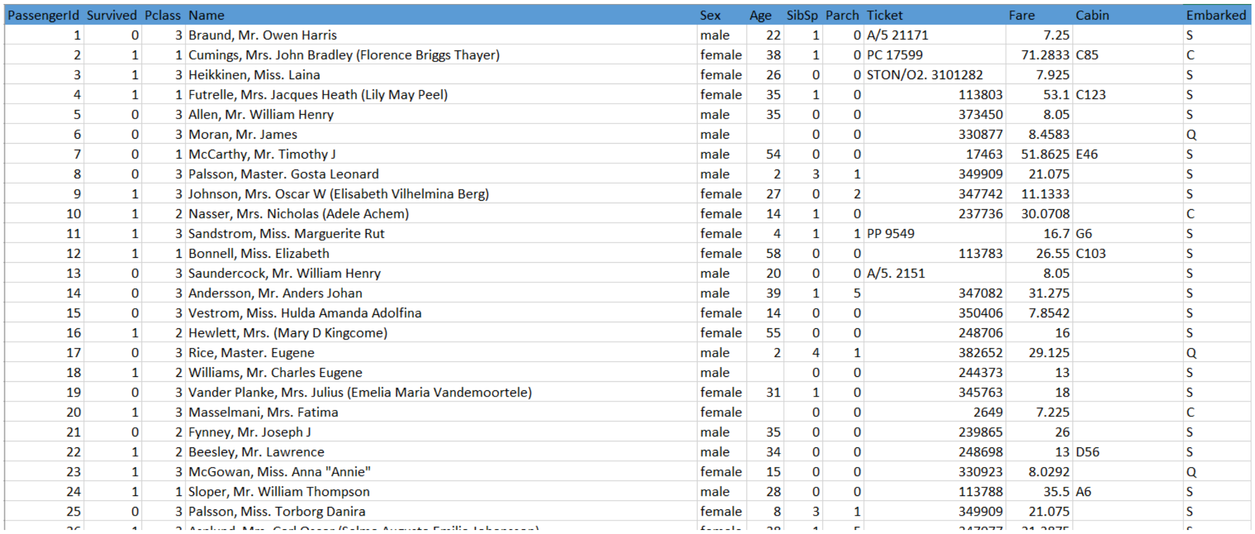
Imputation: It is a process of handling missing values in the dataset by taking the mean, median, etc. for that column data.

We will be using the complete dataset which has around 1309 rows in total.

Titanic Dataset

The below dataset has the following properties:

- Total rows: 1309
- Total Columns: 12
- Data and Variables are mentioned below in the snapshot.



PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
13	0	3	Saundercock, Mr. William Henry	male	20	0	0	A/5. 2151	8.05		S
14	0	3	Andersson, Mr. Anders Johan	male	39	1	5	347082	31.275		S
15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14	0	0	350406	7.8542		S
16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55	0	0	248706	16		S
17	0	3	Rice, Master. Eugene	male	2	4	1	382652	29.125		Q
18	1	2	Williams, Mr. Charles Eugene	male		0	0	244373	13		S
19	0	3	Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)	female	31	1	0	345763	18		S
20	1	3	Masselmani, Mrs. Fatima	female		0	0	2649	7.225		C
21	0	2	Fynney, Mr. Joseph J	male	35	0	0	239865	26		S
22	1	2	Beesley, Mr. Lawrence	male	34	0	0	248698	13	D56	S
23	1	3	McGowan, Miss. Anna "Annie"	female	15	0	0	330923	8.0292		Q
24	1	1	Sloper, Mr. William Thompson	male	28	0	0	113788	35.5	A6	S
25	0	3	Palsson, Miss. Torborg Danira	female	8	3	1	349909	21.075		S

Figure 1: Titanic Dataset

- As the hypothesis is focused on the number of people who have survived/deceased, we can infer the following counts from the above data.

Survived	Count
No	815
Yes	494

Analysis #1 - Gender

In this analysis we are going to perform the below hypothesis.

Null Hypothesis: The proportion of females onboard who survived the Titanic incident was equal than the proportion of males onboard.

Alternate Hypothesis: The proportion of females onboard who survived the Titanic incident was higher than the proportion of males onboard.

Mathematical Notation

H0: count (female who survived) = count (male who did not survive)

Ha: count (female who survived) >= count (male who did not survive)

The **significance level** is 0.05 for the above hypothesis.

Data Analysis on the Titanic dataset

It is a good practice to know about the before performing any experiment/hypothesis. As we are trying to get the relation between the Gender group and the survival rate, we performed data analysis on the Titanic dataset and found the astonishing results.

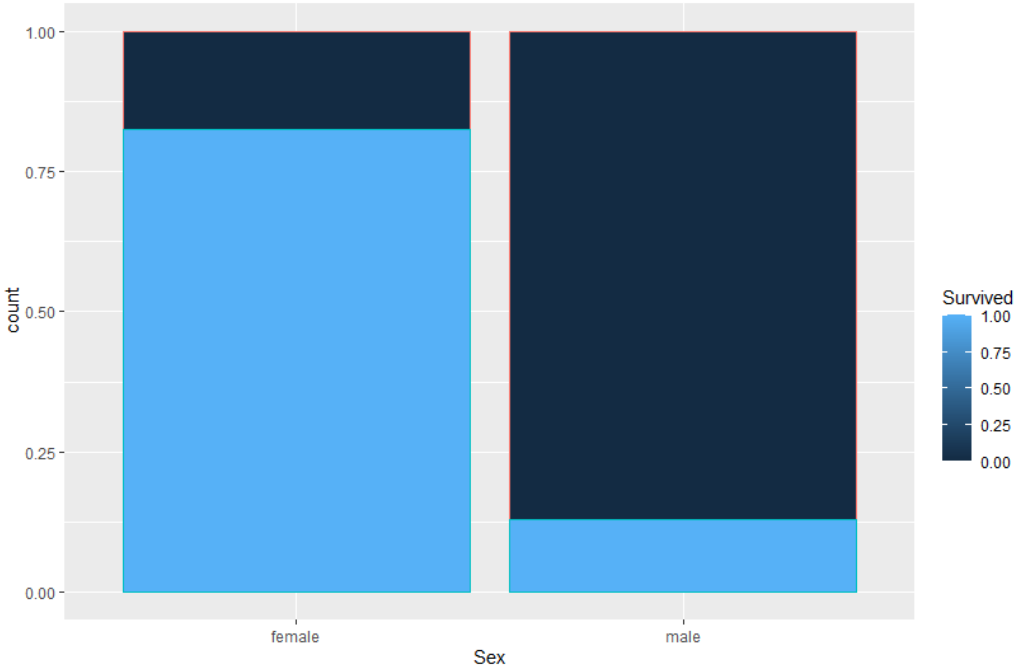


Figure 2: Proportion of the females and male survival ratio.

Gender	Count
Male	843
Female	466

Survived?/Gender	Female	Male
No	81	734
Yes	385	109

Statistical Method and Results

Pearson's chi-squared test is a statistical test performed on the categorical data to evaluate how likely it is that any observed difference between the different groups. As we have a categorical data with 2 level/group for our analysis, we choose to perform Pearson chi-square test for this experiment. Following are the results directly coming from the test.

X-squared	df	p-value
617.31	1	<< 0.0001

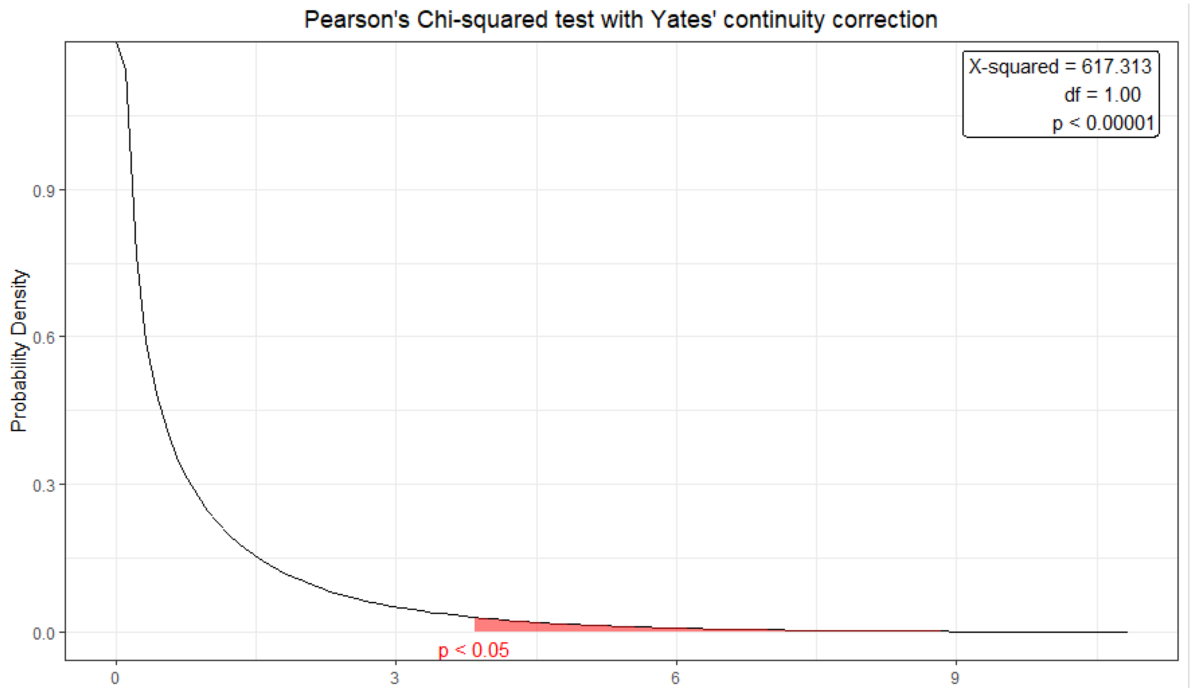


Figure 3: Distribution of chi-square statistic and p-value.

Conclusion

- As the p-value is much lesser than the significance value, we can reject the null hypothesis and conclude that the proportion of females onboard who survived the Titanic incident was higher than the proportion of males onboard.
- From the bar graph, we can conclude that the survival of females was much larger than the male, that is why the p-value is very very small in this case. Hence, supporting the hypothesis test.

Analysis #2 - Pclass

In this analysis we are going to perform the below hypothesis.

Null Hypothesis: The proportion of first-class passengers onboard who survived the Titanic incident was equal than the proportion of other class passengers onboard.

Alternate Hypothesis: The proportion of first-class passengers onboard who survived the Titanic incident was higher than the proportion of other class passengers onboard.

Mathematical Notation

H0: count (first-class passengers who survived) = count (other class passengers who did not survive)

Ha: count (first-class passengers who survived) >= count (other class passengers who did not survive)

The **significance level** is 0.05 for the above hypothesis.

Data Analysis on the Titanic dataset

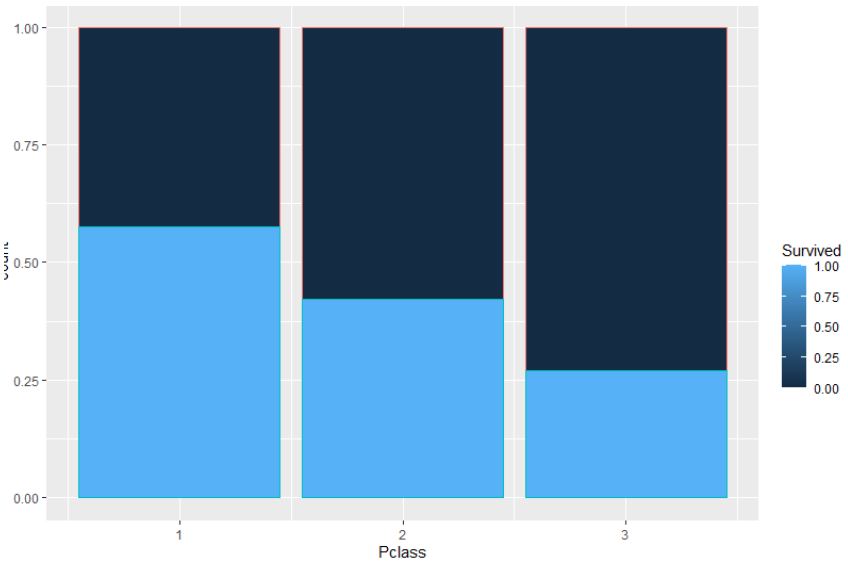


Figure 4: Proportion of the passenger class for each survival category.

Class	Count
1st	323
2nd	277
3rd	709

Survived?/Class	1st	2nd	3rd
No	137	160	518
Yes	186	117	191

Statistical Method and Results

Let’s perform the **Pearson's Chi-squared test** on the above hypothesis.

X-squared	df	p-value
91.724	2	<< 0.0001

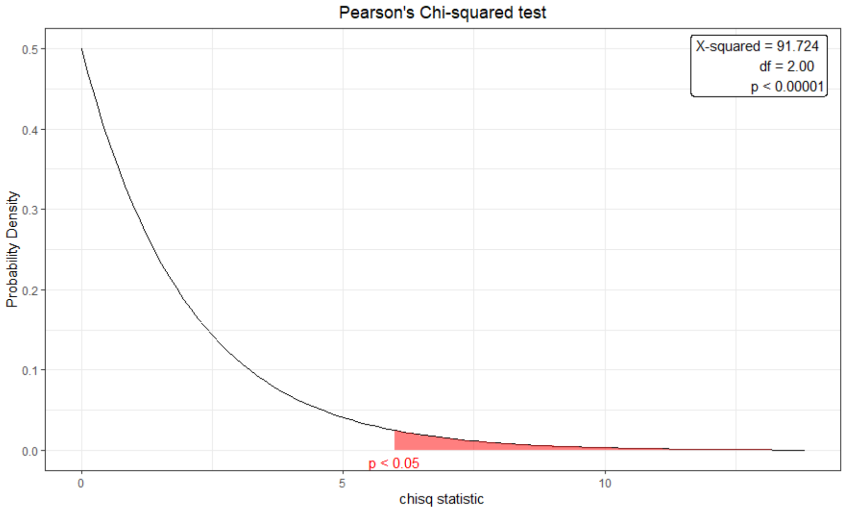


Figure 5: Distribution of chi-square statistic and p-value.

Conclusion

- As the p-value is much lesser than the significance value, we can reject the null hypothesis and conclude that the proportion of first class passengers onboard who survived the Titanic incident was higher than the proportion of other class passengers (second and third) onboard.
- From the bar graph, we can conclude that the survival of first-class passengers was much larger than other class passengers, that is why the p-value is very very small in this case. Hence, supporting the hypothesis test.

Analysis #3 - Age

In this analysis we are going to perform the below hypothesis.

Null Hypothesis: The mean age of passengers onboard who survived the Titanic incident was equal to the mean age of passengers onboard.

Alternate Hypothesis: The mean age of passengers onboard who survived the Titanic incident was not equal to the mean age of passengers onboard.

Mathematical Notation

H0: count (mean age of passengers who survived) = count (mean age of passengers who did not survive)

Ha: count (mean age of passengers who survived) != count (mean age of passengers who did not survive)

The **significance level** is 0.05 for the above hypothesis.

Data Analysis on the Titanic dataset

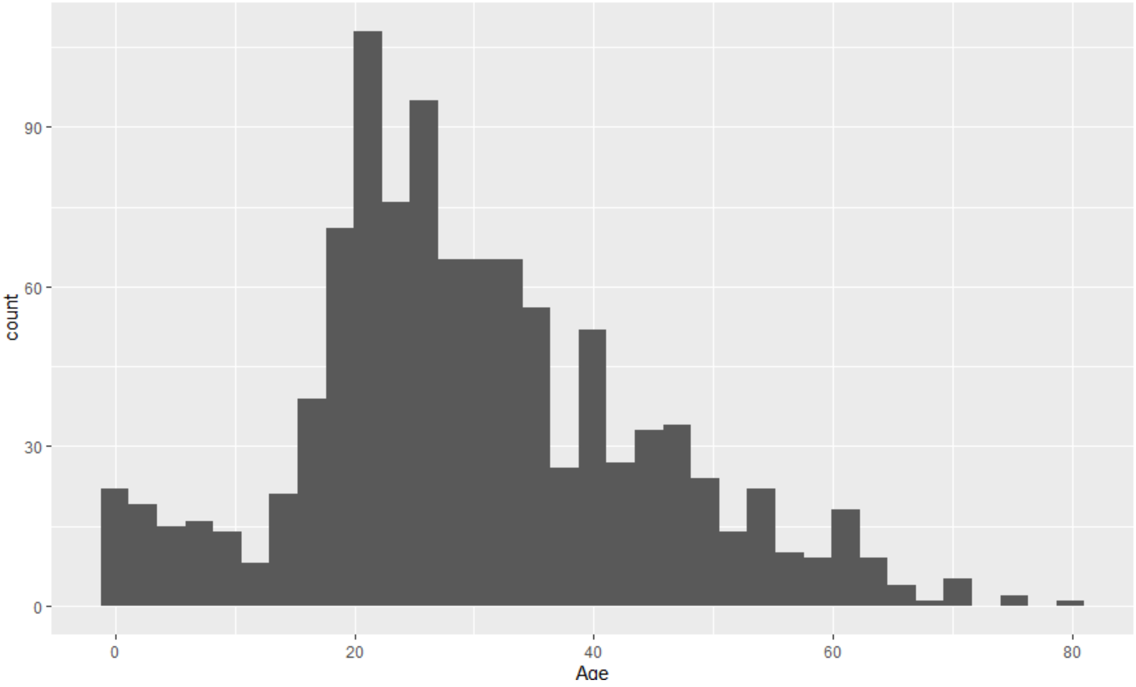


Figure 6: Histogram for the Age of all the passengers.

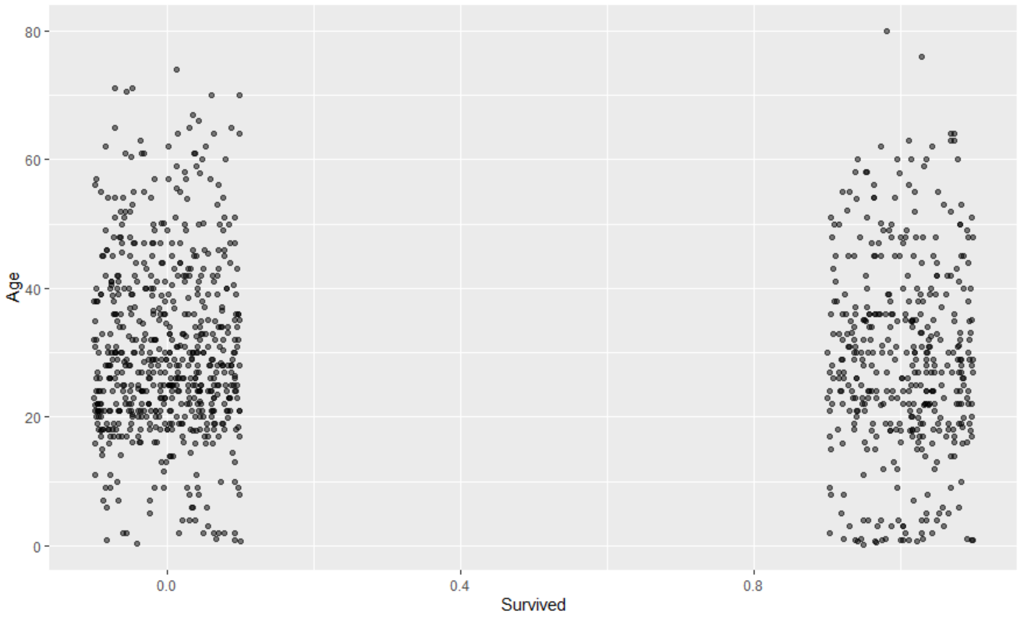


Figure 7: Scatter plot of Age vs. Survived

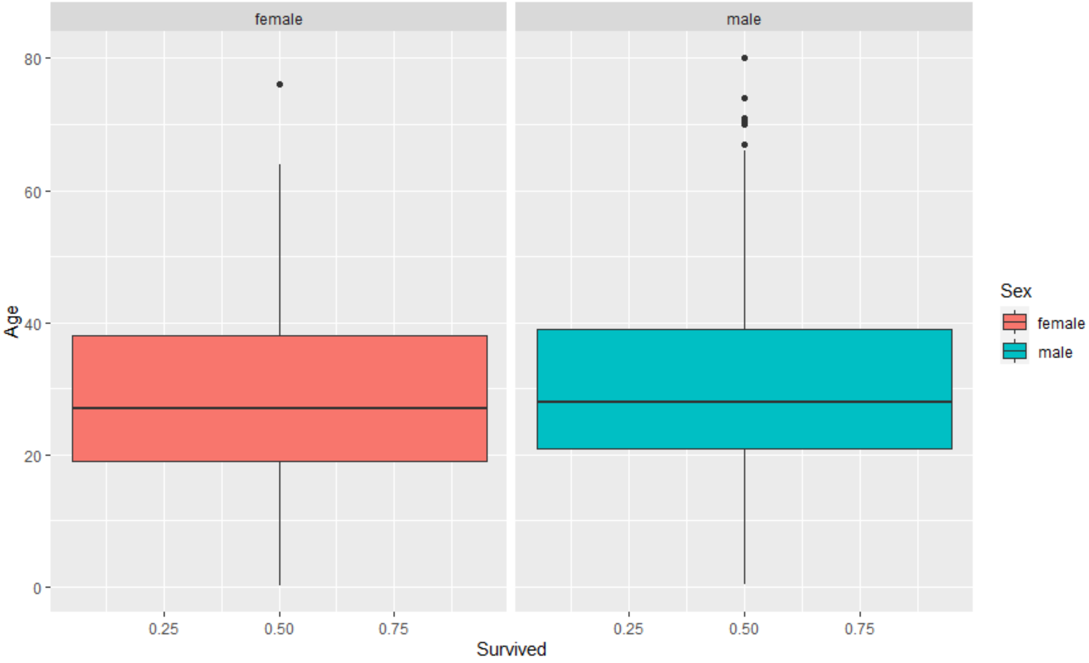


Figure 8: Box plot for Age vs. Survived

Statistical Method and Results

In statistics, **Welch Two Sample t-test** is a two-sample location test which is used to test the hypothesis that two populations that have equal means. As we have a continuous data for our experiment, we choose to perform Welch's t-test for this experiment. Following are the results directly coming from the test.

t	df	p-value
1.7012	939.63	0.08924

```
data:  Age.imp.mean by Survived
t = 1.7012, df = 939.63, p-value = 0.08924
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1978596  2.7740117
sample estimates:
mean in group 0 mean in group 1 
       30.36724        29.07917 
```

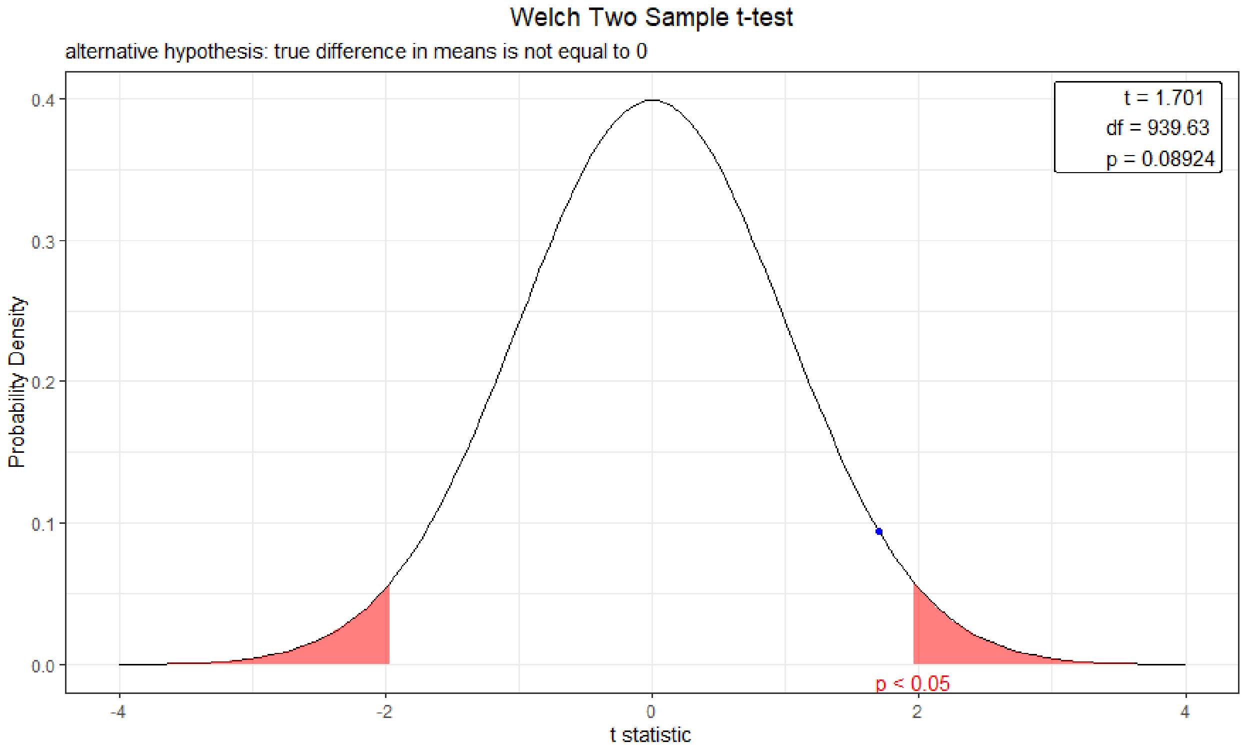


Figure 9: Distribution of common statistics and p-value

As we have the p-value is very close to the significance value, we are not confident in failing to reject the null hypothesis. Hence, to validate our findings, we also performed the Kruskal-Wallis rank sum test to ensure the above hypothesis results.

The **Kruskal–Wallis test** by ranks is a non-parametric method for testing whether samples originate from the same distribution. It is used for comparing two or more independent samples of equal or different sample sizes. Following are the results are directly coming from the test.

t	df	p-value
1.3734	1	0.2412

Conclusion

- As the p-value is larger than the significance value in both the test (t-test and Kruskal–Wallis test), we are failed to reject the null hypothesis and conclude that the mean age of passengers onboard who survived the Titanic incident was equal to the mean age of passengers onboard
- From the boxplot and scatter plots, we can roughly say that the distribution of age of the passengers who survived are somewhere same as the distribution of age of the passengers who did not survive.

R-code file

final code.txt

Attached as a separate file along with the submission report.