# Exploring the Relationship between Passenger Class and Survival during the sinking of the Titanic

Anthony Wen

2024-02-14

Please complete the following:

- Address each of the bolded prompts or questions below.
- Compile the document into a multipage PDF file
- Submit to Gradescope and paginate individual questions correctly
- Please leave lines containing “\” in the file, as these force linebreaks

## Preamble

There is some debate in statistical circles as to how to treat the survival data from the sinking of the Titanic. Some individuals view this data as representing a fixed event, with determined outcomes, describing a specific population exactly. Some consider it a single observation of potential ship sinkings of the time. For the purposes of this exercise we will consider the outcome of the sinking of the Titanic as a random observation of all possible outcomes, where the specifics of any individual’s survival are dependent on many complex factors.

In such a context we test hypotheses against the hypothetical potential outcomes that could have occured under slightly different circumstances. In such a context we can compare this specific outcome against other potential outcomes utilizing the process of hypothesis testing to address questions about the underlying “population” of possible outcomes.

## Question 1 - Defining Hypotheses [3 points]

### Using the Titanic dataset provided in the variable `titanic.data`, construct a table of the `Class` and `Survived` variables. Call this variable `titanic.table`

```
# Construct a table of the titanic.data variables Class and Survived
titanic.table= table(titanic.data$Class, titanic.data$Survived)

print(titanic.table)
```

```
## 
##         No Yes
##   1st  122 203
##   2nd  167 118
##   3rd  528 178
##   Crew 673 212
```

We wish to conduct a chi-squared test for independence hypothesis test for this data set (and for these variables).

### What are the null and alternative hypotheses for this test? Be sure to define these in the specific context of this problem.

$H_0$: “Null Hypothesis: There is no relationship between the passenger class and survival rate on the Titanic”

$H_a$: “Alternative Hypothesis: The survival rate on the Titanic depends on the passenger class”

## Question 2 - Hypothesis Test [2 points]

### Using the functions available in R, conduct a chi-squared test for independence test on the table previously created. Save this test as the variable `titanic.test`

```
# Save the results of the chisq.test in the variable titanic.test
titanic.test= chisq.test(titanic.table)
```

Consider the following table calculated for the hypothesis test.

```
print(titanic.test$expected)
```

```
##
##             No       Yes
##   1st  220.0136 104.98637
##   2nd  192.9350  92.06497
##   3rd  477.9373 228.06270
##   Crew 599.1140 285.88596
```

### Explain as if explaining to someone with moderate statistical knowledge what the table of values displayed above represents, contextualized to this specific hypothesis. Also, identify why (under the null hypothesis being true), the observed data does not match the table seen above (aside from not being whole numbers).

“The table displays based on the fact that the survival rate on the titanic is independent of the class they are in. Aka we are operating under the fact that the Null Hypothesis is true. Based on this, the table displays what the distribution of survivors and non-survivors across different classes on the Titanic.

the observed data does not match the expected as one shows more survivors in certain classes than others when the other shows survival rates to be similar across all the classes. This difference probably happened because the null hypothesis did not consider real-world factors like the access to lifeboats, passenger emotions and skills, and things like that. ”

## Question 3 - P-value and conclusions [3 points]

```
print(titanic.test$p.value)
```

```
## [1] 4.999928e-41
```

The code above reports the p-value for the hypothesis test.

### Explain as if explaining to someone with moderate statistical knowledge what the p-value represents in the context of the specific hypotheses being tested here.

“The p-value basically represents the probability of getting the observed results if the null hypothesis was true. In our hypothesis, it would mean the value/measure would show how likely it is for there to be extreme differences in survival rates and classes on people on the titanic if these two things actually have no correlation with each other. Since in this case the p-value is so low, it would basically means that the chance of observing significant differences in survival rates in different classes would be very unlikely.”

### Based on the results of this study, what do you conclude from your hypothesis test? Be sure to frame your response in the context of the specific hypotheses being tested here.

“Since the p-value is so small, we can basically reject the null hypothesis. The differences in survival rates among different classes are probably not random. There is also definetly a correlation between class and survival rate on the Titanic. For example, the dataset would more likely suggest that first class passengers would have a higher chance of survival as they have better access to lifeboats and other real-world factors”

## Question 4 - Other Considerations [2 points]

```
print(table(titanic.data$Class, titanic.data$AgeOrSex))
```

```
## 
##        Child Female Male
##   1st      6    144  175
##   2nd     24     93  168
##   3rd     79    165  462
##   Crew     0     23  862
```

### Consider the table of values above, which looks at the relationship between class and age or sex (for adults). Contextualizing with what was seen in the analysis during class, how might we consider the information above as it relates to the hypothesis test that we just conducted? IE: What relationships between variables must be considered for a more in depth analysis?

“We must consider the fact that there could be a significant impact on the relationship between class and survival if we had to take into account age, sex, and class. For example, we can see that at the third class – where survival rates are the lowest – there were more males in the crew. This could be due to the fact where the status-quo has a rule of”women’s first” onto the life boat, which causes men having a less likely time and chance to survive. There are other variables (mostly social and real-life related) like this that we could take into consideration to understand the survival rates, which also gives us a more “real” analysis of the data when it comes to thinking about the relationship between class, age, and sex on the survival rates in titanic.”