Analyzing Survivors of the Titanic Passenger List

Scott Miner

Colorado State University – Global Campus

**Analyzing Survivors of the Titanic Passenger List**

The author used SAS to analyze the passenger list of the Titanic. The following is the author’s first-time performing a hypothesis analysis via SAS, where the author solely devised the analyses and hypotheses.

**Brief Description**

The author found the dataset on the Tableau Public Resources webpage. The dataset contains all the known passengers of the Titanic. The dataset includes information like passengers’ names, whether passengers survived, passengers’ age and fare, destination, and cabin information. The author imported the dataset from a .CSV file. In all, the dataset contains 1,309 observations and 14 variables (“Resources | Tableau Public,” n.d.).

Although smaller than many datasets available online, the dataset is larger than many available in text. Additionally, in the author’s opinion, the dataset lends itself well to analysis, offering a structure like that of many datasets found in Elliott and Woodward’s 2016 book *SAS Essentials: Mastering SAS for data analytics.*

**Analysis**

Specifically, the raw .CSV file contains two categorical variables: “SEX”and “SURVIVED.” *“*SEX” has the potential to contain the values “male” or “female,” while “SURVIVED” has the potential to contain the values “1” (for survived) or “0” (for died). With these two categorical variables, we can construct a 2 x 2 cross-tabulation table, conduct a test of independence, and calculate a chi-square statistic to determine whether there is an association between passengers’ gender and passengers’ ability to survive the events that sunk the Titanic on May 31, 1911 (Elliott & Woodward, 2016).

We can state the null and alternative hypotheses for the test of independence as follows:

Gender and survival status are independent (the proportions of those who died and those who survived are the same amongst males and females)

Gender and survival status are not independent (the proportions of those who died and those who survived are different amongst males and females)

We can use the PROC FREQ procedure in SAS to create a 2 x 2 cross-tabulation table:A portion of the results are shown in Figure 1:

| **Table of sex by survived** | | | |
| --- | --- | --- | --- |
| **sex** | **survived** | | |
| **Frequency Percent Row Pct Col Pct** | **0** | **1** | **Total** |
| **female** | 127 9.70 27.25 15.70 | 339 25.90 72.75 67.80 | 466 35.60 |
| **male** | 682 52.10 80.90 84.30 | 161 12.30 19.10 32.20 | 843 64.40 |
| **Total** | 809 61.80 | 500 38.20 | 1309 100.00 |

Figure 1. Cross-tabulation table of sex by survived

Out of 1,309 observations, 466 (35.6%) were female, while 843 (64.4%) were male. Likewise, 809 (61.8%) passengers died, while 500 (38.2%) passengers survived. Specifically, 127 (9.7%) female passengers died, 339 (25.9%) female passengers survived, 682 (52.1%) male passengers died, and 161 (12.3%) male passengers survived.

Of females, only 27.25% died, while 72.75% survived. Of males, 80.9% died, while 19.1% survived. Of those who died, 15.7% were female, while 84.3% were male. Of those who lived, 67.8% were female, and 32.2% were male. From this, we can see that males were more likely to die than females and that we are more likely to observe a deceased male passenger than any other result.

Figures 2, 3, and 4 present visual depictions of the above statistics using a frequency plot, a mosaic plot, and a stacked horizontal bar chart, respectively.



Figure 2. Frequency plot of the distribution of gender by survival status



Figure 3. Mosaic Plot for the distribution of sex by survival status



Figure 4. Horizontal stacked bar chart for the distribution of sex by survival status

We can create a table displaying expected results, assuming no association between either sex or survival status. The results are displayed in Figure 5.



| **Table of sex by survived** | | | |
| --- | --- | --- | --- |
| **sex** | **survived** | | |
| **Frequency Expected** | **0** | **1** | **Total** |
| **female** | 127 288 | 339 178 | 466 |
| **male** | 682 521 | 161 322 | 843 |
| **Total** | 809 | 500 | 1309 |

Figure 5. Cross-tabulation table of sex by survival status with expected results

Out of 466 observations, we would expect to see 288 deceased female passengers and 178 living female passengers if there were no association between passengers’ gender and passengers’ life status. What we see, on the other hand, is that more female passengers survived (*n*=339) than died (*n*=127). The expected results are used in tandem with the observed results to calculate the chi-square statistic along with the corresponding degrees of freedom and *p-*values (Chandrayan, 2019). The output is shown in Figure 6.

| **Statistic** | **DF** | **Value** | **Prob** |
| --- | --- | --- | --- |
| **Chi-Square** | 1 | 365.8869 | <.0001 |
| **Likelihood Ratio Chi-Square** | 1 | 372.9213 | <.0001 |
| **Continuity Adj. Chi-Square** | 1 | 363.6179 | <.0001 |
| **Mantel-Haenszel Chi-Square** | 1 | 365.6074 | <.0001 |
| **Phi Coefficient** |  | -0.5287 |  |
| **Contingency Coefficient** |  | 0.4674 |  |
| **Cramer's V** |  | -0.5287 |  |

Figure 6. Chi-square statistics and p-values

The chi-square value is 365.89 and p < 0.0001. The p-value for all computed chi-square statistics is less than 0.0001. These results tell us to reject the null hypothesis of independence between variables and conclude there is evidence for an association between the variables gender and survival status for passengers onboard the Titanic.

Similarly, the results of Fisher’s Exact Test corroborate this evidence. We use Fisher’s Exact Test to calculate chi-square statistics when counts are small (Elliott & Woodward, 2016). We observe this output in Figure 7. Again, the results show p < 0.0001.

| **Fisher's Exact Test** | |
| --- | --- |
| **Cell (1,1) Frequency (F)** | 127 |
| **Left-sided Pr <= F** | <.0001 |
| **Right-sided Pr >= F** | 1.0000 |
|  |  |
| **Table Probability (P)** | <.0001 |
| **Two-sided Pr <= P** | <.0001 |

Figure 7. Fisher's Exact Test results

After finding evidence of an association between variables, we may want to determine the extent to which these variables influence each other. In other words, how much more likely were male passengers to die than female passengers?

To determine these results, we will be calculating the relative risk. When calculating the relative risk in a retrospective study the result is referred to as the Odds Ratio (OR) (Elliott & Woodward, 2016). We perform this calculation in SAS by supplying the RELRISK option to the TABLES statement of the PROC FREQ procedure:





Figure 8. Odds Ratio and Relative Risks calculation

An Odds Ratio with a value below 1 tells us that the odds of a female passenger dying onboard the Titanic is less than the odds of a male passenger dying. These results confirm our previous findings. What if we want to calculate the degree to which male passengers were more likely to die than female passengers? In this scenario, we take the inverse of the OR calculated above (1/0.0884=11.31) and conclude male passengers were 11.31 times more likely to die than female passengers. The 95% Confidence Limits provided in Figure 8 give ranges for our estimation. Taking the inverse of the lower limit OR (1/0.0677=14.77) and upper limit OR (1/0.1155=8.65), we can say male passengers were anywhere between 8.65-14.77 times more likely to die than female passengers. We can calculate these same values and confirm our results using SAS (Elliott & Woodward, 2016). We observe the results in Figure 9.





Figure 9. Changing the order of gender in the Odds Ratio analysis

In the first table in Figure 9, we see that the ordering of the genders has been reversed, with males now coming before females. Such a reversal allows us to compute the OR for males as compared to females. The results of our calculation confirm our analysis above: male passengers were, on average, 11.31 times more likely to die than female passengers, with ranges varying from 8.67 times more likely to 14.76 times more likely (Elliott & Woodward, 2016).

**Conclusion**

In this study, we analyzed the passenger list of the Titanic. Specifically, we created a 2 x 2 cross-tabulation table using the variables gender and survival status to assess whether any association exists between these two variables. We did find evidence of an association between gender and survival status. It appears more males died than females. One possible reason for this is that perhaps male passengers gave up their positions on lifeboats to ensure female passengers would survive. Although outside the realm of this study, it would be interesting to see if any correlations exist between passengers ages, fares, genders, and survival statuses. The author provides two depictions in figures 10 and 11 to aid in future research.

Figure 10. Distribution of age by gender and survival status



Figure 11. Distribution of fare by gender and survival status

These graphs visually depict distributions of ages and fares by gender and survival statuses, respectively. These depictions may be useful for future research as it relates to this dataset. The Passenger List of the Titanic appears to be a popular dataset for study, with analyses devoted to predicting the survival rates of passengers (Donges, 2018). The following code provides all the SAS code that was used in this study.





References

Chandrayan, P. (2019, July 11). Inferential Statistic: Understanding Hypothesis Testing Using Chi-Square Test. Retrieved November 10, 2019, from Medium website: <https://towardsdatascience.com/inferential-statistic-understanding-hypothesis-testing-using-chi-square-test-eacf9fcac533>

Donges, N. (2018, May 15). Predicting the Survival of Titanic Passengers. Retrieved November 10, 2019, from Medium website: <https://towardsdatascience.com/predicting-the-survival-of-titanic-passengers-30870ccc7e8>

Elliott, A. C., & Woodward, W. A. (2016). *SAS Essentials: Mastering SAS for data analytics* (Second edition). Hoboken, New Jersey: John Wiley and Sons, Inc.

Resources | Tableau Public. (n.d.). Retrieved November 9, 2019, from <https://public.tableau.com/en-us/s/resources>