**LAB QUESTION:**

Did being a passenger on the Titanic, Lusitania, or Estonia relate to your chances of survival?

**SUMMARY OF INFERENCE:**

> #--------------------#

> # SDS 301 #

> # Final Project #

> #--------------------#

>

> # LAB QUESTION:

> # Did being a passenger on the Titanic, Lusitania, or Estonia relate

> # to your chances of survival?

>

> # Create ship name variable

> est$Ship[est$Survived == 1] <- "Estonia"

> est$Ship[est$Survived == 0] <- "Estonia"

> tit$Ship[tit$Survived == 1] <- "Titanic"

> tit$Ship[tit$Survived == 0] <- "Titanic"

> est$Ship[est$Survived == 1] <- "Estonia"

> est$Ship[est$Survived == 0] <- "Estonia"

> lus$Ship[lus$Survived == 1] <- "Lusitania"

> lus$Ship[lus$Survived == 0] <- "Lusitania"

>

> # Merge datasets

> ships <- rbind(tit,lus,est)

>

> # Save new dataset as a CSV file

> write.csv(ships, "shipwreckdata.csv",row.names = FALSE)

>

>

> # Relabel survival status

> ships$SurvivedNew[ships$Survived == 1] <- "Lived"

> ships$SurvivedNew[ships$Survived == 0] <- "Died"

>

> # Two-way frequency table:

> shipSurvival <- table(ships$SurvivedNew, ships$Ship)

> shipSurvival

Estonia Lusitania Titanic

Died 852 1193 809

Lived 137 763 500

>

> # Column proportions:

> propcolumn <- prop.table(shipSurvival,2)

> propcolumn

Estonia Lusitania Titanic

Died 0.8614762 0.6099182 0.6180290

Lived 0.1385238 0.3900818 0.3819710

>

> # Stacked bar chart:

> barplot(propcolumn, main='Survival by Shipwreck',

+ xlab="Ship", ylab="Proportion", legend=T,

+ xlim=c(0,5))

>

> # Expected counts of chi-squared test:

> chisq.test(shipSurvival)$expected

Estonia Lusitania Titanic

Died 663.5181 1312.2764 878.2055

Lived 325.4819 643.7236 430.7945

>

> # Chi-squared test of independence:

> chisq.test(shipSurvival, correct=T)

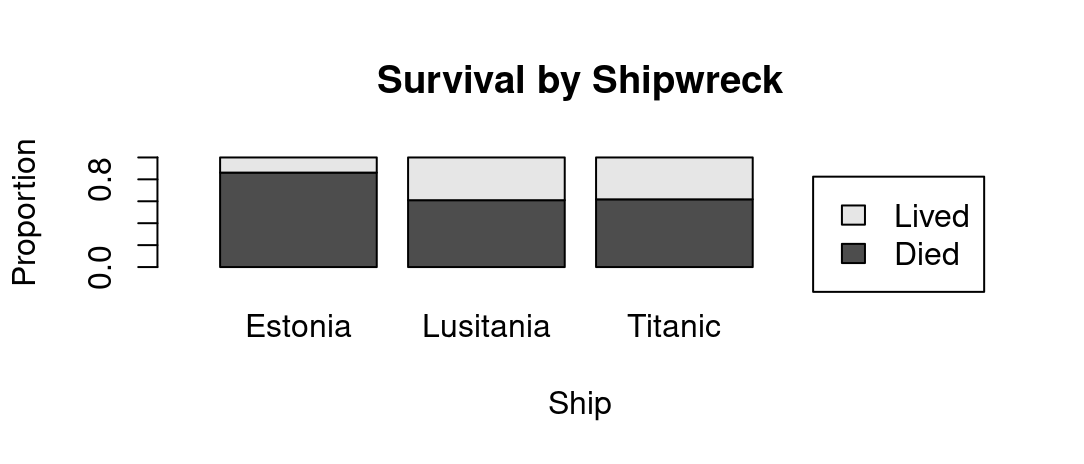
Pearson's Chi-squared test

data: shipSurvival

X-squared = 212.2, df = 2, p-value < 2.2e-16

**GRAPHICAL OUTPUT:**

*Stacked proportion bar chart of observed survival status by ship:*



**LAB QUESTION ANSWER:**

There is significant evidence to suggest that being a passenger on the Titanic, Lusitania, or Estonia is related to one’s chances of survival. However, what is surprising is that the Estonia has a higher mortality rate than both the Titanic and Lusitania despite occurring years after the two incidents, contrasting with my initial prediction. This conclusion was found through a chi-square test of independence.

To arrive at this conclusion, the null and alternative hypotheses for this comparison must be stated. Whether or not a passenger survived a ship disaster and the ship they were on are independent in the null hypothesis - meaning that there are no differences in the conditional distributions for each level of one of the categorical variables. The two categorical variables are related and dependent on each other in the alternative hypothesis - meaning that the conditional distribution of one variable differs for each level of the other variable.

As such, the assumptions in the criteria for the chi-square test of independence must be checked. It will be assumed that the samples from the dataset of passengers on a ship disaster were collected at random. Furthermore, it will be assumed that the samples of passengers are independent from one another, which is probable given the circumstances of the lab question. Finally, it will be assumed that the sample sizes included in this analysis are sufficient and large enough as in the two-way table of the expected counts between the two categorical variables generated, all expected counts are at least 5.

Because of this, we are able to proceed with our analysis. From the stacked bar chart generated from the proportion table of the observed survival status by ship, it is evident that there is a distinct relationship between a passenger’s survival status and their ship, especially when visually comparing the survival rates on the Estonia to the Lusitania and Titanic. As such, the categorical variables are likely related to one another. However, to confirm this, a chi-square test of independence must be done.

When computed, the appropriate chi-square statistic is as 212.2. And with the degrees of freedom computed, 2, the p-value of this analysis can be identified as less than a significance level of 0.05, computed as 2.2e-16.

Thus, we reject the null hypothesis because there is evidence for a significant difference between the conditional distributions for each level of one of the categorical variables. The chi-square statistic can be found within the tails of the chi-square distribution past the chi-square critical value attributed to the distribution with 2 degrees of freedom, 5.991, and a confidence level of 95%.