Project Part 3

Julia Burek

Data Summary

The data set I am using for my project is called "Survival of Passengers on the Titanic." The data set contains information on the survival status, sex, age, and passenger class of 1309 passengers in the Titanic disaster of 1912. In total, there were 2240 passengers on the Titanic and more than 1500 died (Reference 1). The Titanic sinking is a very historic, disastrous event. This data frame contains 1309 observations with several variables, including name, survived status, sex, age, and passenger class. The data set was begun by various researchers. The main source of this data is from the Encyclopedia Titanica, an online source of information about the Titanic. The data set was begun by various researchers. One of the original sources is Eaton & Haas (1994) "Titanic: Triumph and Tragedy", Patrick Stephens Ltd, which includes a passenger list created by many researchers and edited by Michael A. Findlay (Reference 2). Other researchers have updated and improved the titanic data frame to include previously missing data and new variables. For the purposes of my project, however, I will be using the older, smaller data set, TitanicSurvival. I chose to use this data set because I thought the variables included were sufficient for my data exploration. The number of observations was also large enough to run a statistical test.

TitanicSurvival is a data frame with 1309 observations on the following 4 variables: survived (no or yes), sex (female or male), age (in years–and for some children, fractions of a year), passengerClass (1st, 2nd, or 3rd). Each observation or row represents a different passenger that contains information about the associated passenger. The data set contains each of their names if applicable. For the purpose of my project, I want to explore the age and survived variables. There is missing data about the age of some passengers, so I will remove these observations and create a new data set that now contains 1046 observations.

```
# Removing Observations with Age Missing
Titanic <- read.csv("~/Downloads/TitanicSurvival.csv")
Titanicnew <- na.omit(Titanic)
```

Question and Test

In the historic tragedy of the sinking of the Titanic in 1912, many passengers lost their lives. Some, however, did survive. There were numerous factors that helped a passenger's likelihood of survival including sex, age, and passenger class. Women and children, for example, were prioritized in the rescue efforts. Due to this, I want to explore this idea of survival rates through a statistical test. My question is: Did children have a greater rate of survival than adults on the Titanic? I want to compare the proportion of children (below 18) who survived to the proportion of adults (18 and older) who survived. I want to see if the proportion of children who survived is higher than the proportion of adults who survived. In order to do this, I will perform a Two-Proportion Z-test. The null hypothesis is that the proportions are the same. The alternative is that the proportion of children survivors is greater than the proportion of adult survivors. The assumptions for the Two-proportion Z-test are met. The normal approximation to the binomial model assumption is met as the minimum np is 48 and the minimum n(1-p) is 46, both of which are greater than 10. The observations in the data set are independent from one another as they each represent a different passenger on the Titanic. Each sample size is more than 30 (94 for children and 952 for adults). Because there were 2240 passengers total on the Titanic, but only 1046 observations (after omitting NAs for age) were used from this data set for the test, I can say that the data sample I use for the test is a simple random sample from all Titanic passengers.

```
# Transforming age variable from numeric to character
Titanicnew$age <- as.character(Titanicnew$age)

# Subsetting data to child and adult passengers
TitanicYoung <- Titanicnew[Titanicnew$age<'18', c(1:5)]
TitanicOld <- Titanicnew[Titanicnew$age>='18', c(1:5)]

# Subsetting data to survivors
YoungSurv <- TitanicYoung[TitanicYoung$survived=='yes', c(1:5)]
OldSurv <- TitanicOld[TitanicOld$survived=='yes',c(1:5)]

# Two-Prop Z-Test
Test <- prop.test(x=c(nrow(YoungSurv), nrow(OldSurv)), n=c(nrow(TitanicYoung),
                 nrow(TitanicOld)), alternative='greater')
Test
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(nrow(YoungSurv), nrow(OldSurv)) out of c(nrow(TitanicYoung), nrow(TitanicOld
## X-squared = 4.0307, df = 1, p-value = 0.02234
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.01795295 1.00000000
## sample estimates:
##    prop 1    prop 2
## 0.5106383 0.3981092
```

Conclusions

The p-value that resulted from the test is 0.02234 which is less than the 0.05 significance level. Because of this, we reject the null hypothesis that the proportion of children survivors and the proportion of adult survivors is the same. There is significant evidence to support that the proportion of children survivors is greater than the proportion of adult survivors. This goes to show that children were prioritized in the rescue efforts and were more likely to survive the sinking than adults. There were a lot less children than adults on the Titanic, but their proportion of survival (0.5106) was still significantly higher than that of adults (0.3981).

Although my data set contained 1309 passengers, and I ran the test on 1046 of these observations, there were 2240 passengers in total on the Titanic. Because of this, we can cautiously generalize the results of the test to all 2240 passengers on the Titanic. In this case, it is reasonable to say, with caution, that there is evidence that the proportion of children survivors is greater than the proportion of adult survivors across all passengers on the Titanic.

References

1. https://www.history.com/topics/early-20th-century-us/titanic
2. http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/titanic.html