1. Introduction

The sinking of the Titanic is one of the most in/famous shipwrecks. On April 15, 1912, during the first voyage, it sank after colliding with an iceberg. This tragedy shocked the international community and led to better safety regulations for ships.

One of the main reasons that this catastrophe led to such loss of life is that there were not enough lifeboats. Although we cannot deny the element of luck involved, some groups of people were more likely to survive compared to others.

1.1 About the dataset

The dataset of this investigation contains information from 891 of the 2224 passengers and crew on board. Variables included are:

- ✓ Survival (0 = No; 1 = Yes)
- ✓ pclass = Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
- ✓ Name
- ✓ sex Gender
- ✓ Age
- ✓ Sibsp = Number of Siblings/Spouses Aboard
- ✓ parch = Number of Parents/Children Aboard
- ✓ Ticket Number
- ✓ Passenger Fare
- ✓ Cabin
- ✓ embarked Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

2. Data Analysis

2.1 Initial questions

Throughout this analysis where are going to answer the following questions and visualize them.

- ✓ Is the rule 'women and children first' true in this case?
- ✓ Does man with children or spouse had more chances of survival?
- ✓ Is there a trend in survivors' class and gender?
- ✓ Do higher class passengers have more chances of survival?
- ✓ What age is giving the person best chance to survive?

2.2 Investigating data

Let's start by exploring the data set & if needed fix identified problems as well as spot any surprising data points. To answer the questions, we have posed, we won't need all fields so the data set will be limited to Survived, Name (this variable is quite arbitrary & analysis can be done without it. Nevertheless I have decided to keep it in case we'd like to put a name against our numbers.), Pclass, Sex, Age, SibSp and Parch.

Out of the seven selected fields, five are numeric so the describe function provides a good summary of them. There doesn't seem to be any abnormalities except that there seems to be quite a few of missing values in the Age variable (179 out of 891). If we look at the counts by gender:

male 577

female 314

Name: Sex, dtype: int64

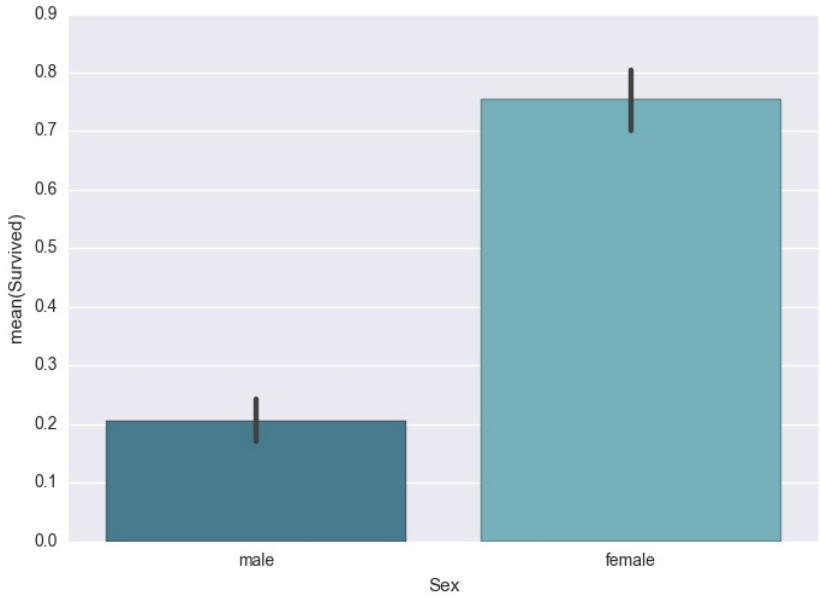
It seems there aren't any missing values, so the main problem is within the age variable. Given this fact I will move forward with a list wise deletion because it doesn't make sense to input the missing age within any statistical tests.

2.3 Answers to initial questions

1. Is the rule 'women and children first' true in this case?

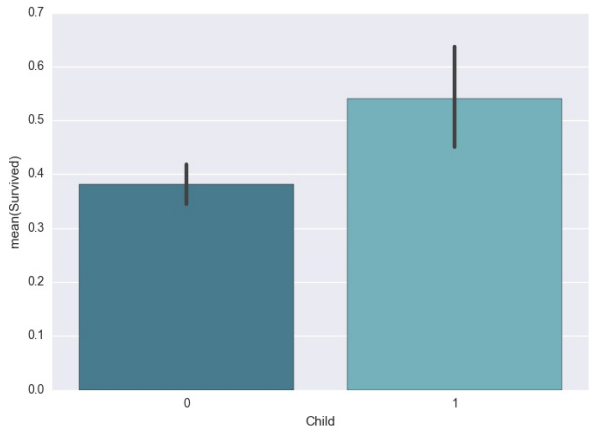
   In my opinion the fastest way to answer this question it to review how are survivors distributed across gender and age. To quickly mark whether the passenger is a child or not, we are going to add a flag Child with 0 value for age equal or greater than 18 and 1 if less.

   Let's first look at the survivals by gender:



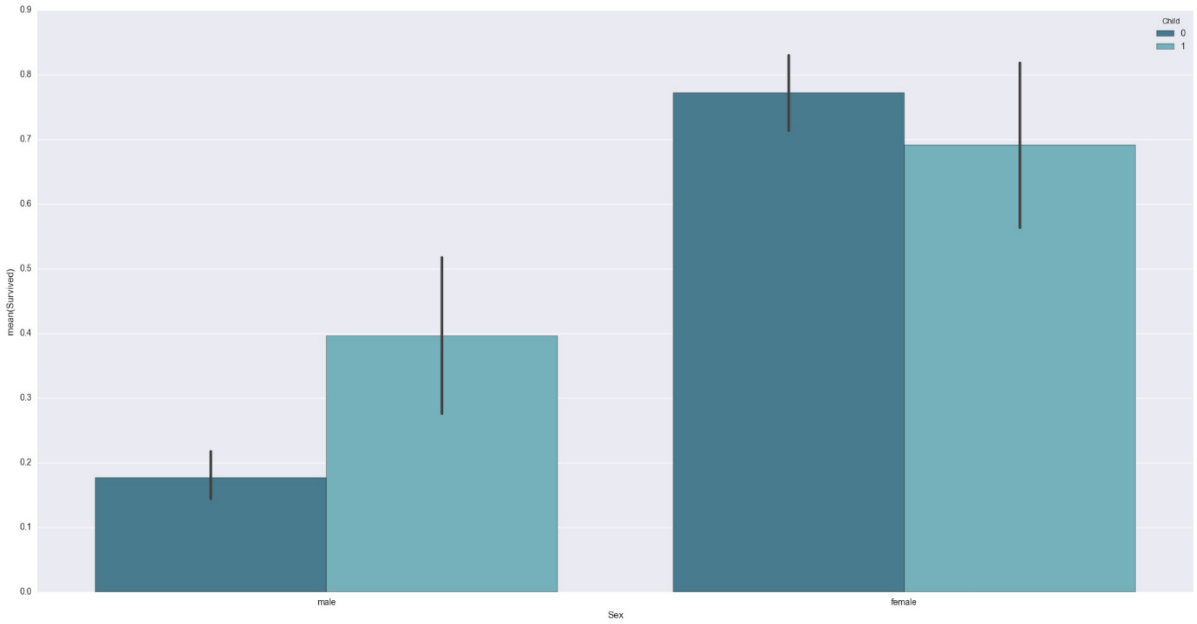
   It is quite obvious that woman's mean survival is much higher compared to men. We can conclude with a good amount of confidence that females were saved with priority compared to men. So, yes – women first.

Now, let's review the picture of children vs adults. If we look at the plot below it is evident that the children have bigger chances compared to adults. It is quite interesting however that the difference in survival is not that large as in the gender distinction.



Given the above two we can give a positive answer to the question – yes, the rule was observed & women/children were saved first.

But the graphs also pose an interesting possibility to extend the question further – does it mean that if a passenger is a child & female this will increase the chances to survive? We can quickly hypothesize around this with the help of the gender/survival plot divided by the children flag.



Despite our expectation the data shows that a female child has less chances to survive compared to an adult female. It is quite curious that for men trend is opposite – a boy has a bigger chance to get out alive compared to a male adult. But overall it seems that who was a child had bigger chances of survival. In other words, while the "rule" – women & children first was observed, most of the men on board died.

2. Does man with children or spouse had more chances of survival?

In order to explore this question, I will create a new category variable FamilyMan. The possible options would be Single, Husband & Father. The latter is equal to Husband with kids. Natural restriction on the data would be gender = male & child = 0, in other words we're going to work with a subset of male adults. Once the data refers only to male adults, there is no way to know for certain if he had a spouse or sibling, so for the sake of our analysis I will assume that the SibSp variable refers to spouse.

We can start by examining simple counts:

Bachelor 308

Husband 61

Father 26

Name: FamilyMan, dtype: int64

Looking at them it seems that the vast majority of our data seems to be consisting bachelors. Naturally we cannot rule out that our assumption above is inaccurate & the variable refers to siblings.

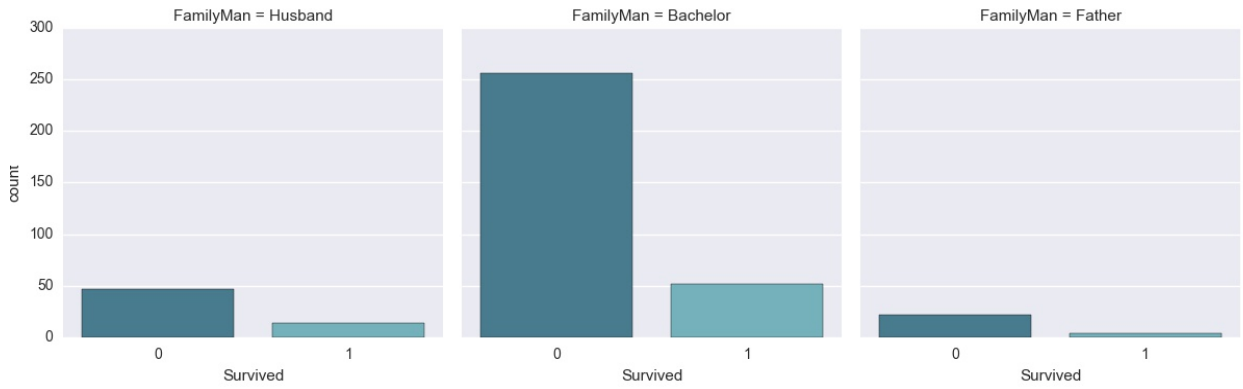
Next let's look at this percent wise, what percentage of men survived based on our categorization.

Husbands survived: 22.95

Bachelors survived: 16.88

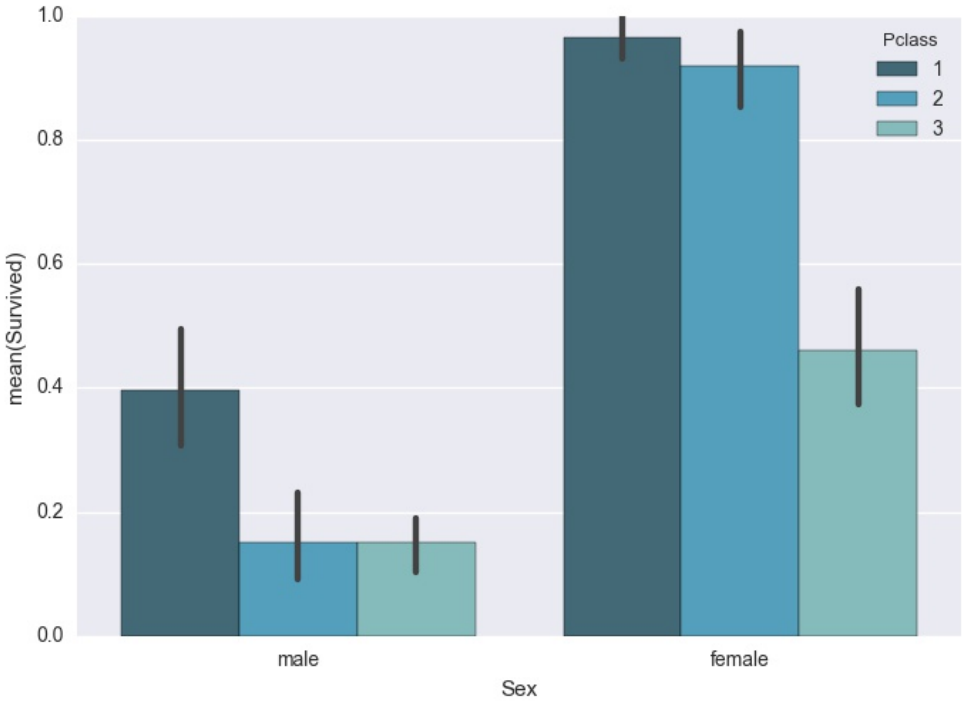
Fathers survived: 15.38

It seems that the highest percent of survivors is within the Husbands category. While a little surprising, this fact can be easily explained. A father would do everything possible to save his family, thinking of himself lastly. At the same time, Husbands have their beloved either on the ship or on shore waiting for them – big motive to live. The above picture can be easily illustrated on a plot, counting the number of survivals per category. While looking at it we need to keep in mind that most of the male adults in our sample fall within the Bachelor category.



3. Is there a trend in survivors' class and gender?

In order to give some insight to this question we'll need to plot the 3 variables – class, gender & survival together.



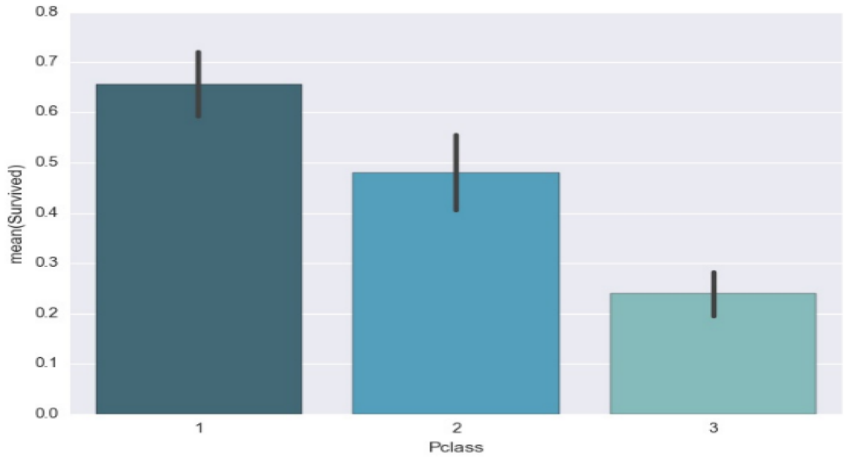
This graph speaks quite a lot already. I've already discussed the difference in the survive mean between men and woman in the first question. So now, let's focus on females first and explore the difference in the survival by class.

Females survived in first class are the highest percent close to 100%. Second class females who survived follow quite closely. If we add survival of 1st and 2nd classes they will double the mean of survival in 3rd class.

In case of men, the picture is very different. Only first class passengers are showing difference in survival – much higher compared to the other two classes.

4. Do higher class passengers have more chances of survival?

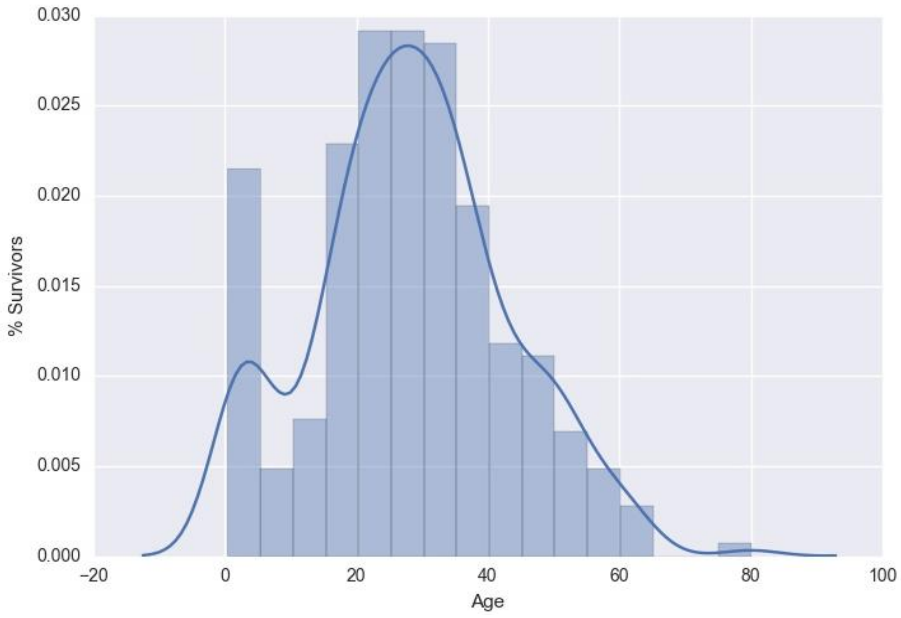
   To extend our last question further let's generalize on how the passenger class is impacting survival rate.



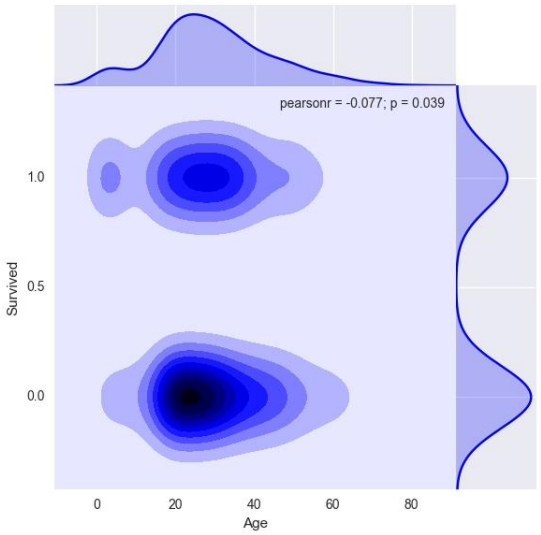
   As such the graph is speaking for itself, the ugly truth is – "first class first". The survival rate for first class passengers is much higher compared to both other classes. As expected the survival rate in the third class is the lowest given their place near the bottom of the ship.

5. What age is giving the person best chance to survive?

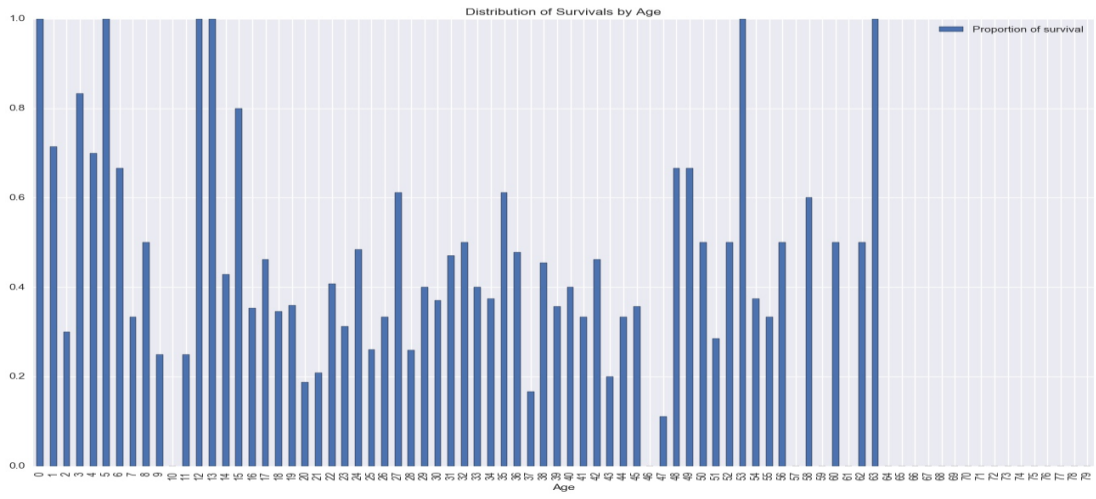
   Let's start by reviewing the distribution of survivors by age:



It is easy to spot that the most of the survivors fall within the age 20 – 40. But if we add non survivors as well?



Again it seems that most of the people who have not survived are again between 20 – 40. From both pictures we can't answer our question. In order to find a response, I will calculate the percentage of survivors per age.



The plot is quite crowded; however, we can see that there is 100% survival rate for some of the ages. Namely those are: 0, 5, 12, 13, 53, 63. But at the end we have to choose one age only and in order to do this I will compare the number of survivors in each category:

Max number of survivors: 7
Best age to survive: 0

It seems that the best chance to survive with the maximum number of survivors is between 0 – 1 years old.

3. Conclusions

In the previous section, I've assumed there is no difference between proportions for each sample. But what happens if this is not true?

In statistics there is a way to know if the present difference is significant and therefore the conclusions may not be true. With a two-proportion z-test we can determine this.

So for the first question, we saw through the graphs that survivors mean was higher for females & children opposed to males and adults.

Let's conduct a hypothesis test in order to see if the difference is significant. First males opposed to females:

$$H0: P_{male} - P_{female} = 0$$

$$HA: P_{male} - P_{female} \neq 0$$

all people: 714

females: 261

male: 453

We'll use two-proportion z-test as mentioned before with significance level $\alpha = 0.05$:

p: 0.41

SE: 0.038

z: -14.4

Given that z-score, the P-value is going to be extremely small (for sure smaller than our significance level 0.05). Thus we can consider to reject the null hypothesis & say that there is a significant difference between proportions.

We can apply the same scenario to the child & adults case:

total people: 714

children: 113

adults: 601

p: 0.406162464986

SE: 0.0503565958354

z: 3.15333854957

And again the P value is very small. So we can reject the null, thus the difference between populations is significantly different.

The same test can be done with many other iterrations (all where proportions are compared). The main conclusion is that the data set is not a good sample to perform analysis and generalize the conclusions on the entire population.

Sources

The list of sources used to complete this investigation is:

- ✓ Titanic dataset provided by Udacity (Data Analyst Nanodegree Project 2)
- ✓ Kaggle titanic competition page (https://www.kaggle.com/c/titanic)
- ✓ Seaborn statistical data visualization reference page (http://stanford.edu/~mwaskom/software/seaborn/)
- ✓ Pandas documentation (http://pandas.pydata.org/pandas-docs/stable/#)
- ✓ Test statistics for difference between proportions (http://stattrek.com/hypothesis-test/difference-in-proportions.aspx)
- ✓ https://web.stanford.edu/~mwaskom/software/seaborn/tutorial/categorical.html#distributions-of-observations-within-categories
- ✓ http://optional.is/required/2012/04/25/titanic-visualized/