[redacted]

Intro to Probability & Statistics

Project 8

Dr. Nicholas Jacob

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
13	0	3	Saundercock, Mr. William Henry	male	20	0	0	A/5. 2151	8.05		S
14	0	3	Andersson, Mr. Anders Johan	male	39	1	5	347082	31.275		S
15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14	0	0	350406	7.8542		S
16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55	0	0	248706	16		S
17	0	3	Rice, Master. Eugene	male	2	4	1	382652	29.125		Q
18	1	2	Williams, Mr. Charles Eugene	male		0	0	244373	13		S
19	0	3	Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)	female	31	1	0	345763	18		S
20	1	3	Masselmani, Mrs. Fatima	female		0	0	2649	7.225		C
21	0	2	Fynney, Mr. Joseph J	male	35	0	0	239865	26		S
22	1	2	Beesley, Mr. Lawrence	male	34	0	0	248698	13	D56	S
23	1	3	McGowan, Miss. Anna "Annie"	female	15	0	0	330923	8.0292		Q
24	1	1	Sloper, Mr. William Thompson	male	28	0	0	113788	35.5	A6	S
25	0	3	Palsson, Miss. Torborg Danira	female	8	3	1	349909	21.075		S
26	1	3	Asplund, Mrs. Carl Oscar (Selma Augusta Emilia Johansson)	female	38	1	5	347077	31.3875		S
27	0	3	Emir, Mr. Farred Chehab	male		0	0	2631	7.225		C
28	0	1	Fortune, Mr. Charles Alexander	male	19	3	2	19950	263	C23 C25 C27	S
29	1	3	O'Dwyer, Miss. Ellen "Nellie"	female		0	0	330959	7.8792		Q
30	0	3	Todoroff, Mr. Lalio	male		0	0	349216	7.8958		S

Though this is a condensed version, the complete version of the dataset above can be found on github, originally from Kaggle.com. It is a representation of the unique circumstances for individuals that were aboard the Titanic when it sunk. The data collected includes names of the passengers, ticket number, survival, sex, ticket class, fare, number of family members present on the ship, and boarding location. The abbreviations used in this dataset are described below.

This dataset uses categorical variables to describe the population of passengers of the titanic. The ordinal data from this set is used to identify names of passengers and sex. The quantitative data represents the age and number of relatives of each passenger. The survival record, number of relatives, and class would be considered discrete data, whereas the fare is continuous.

The titanic shipwreck has always intrigued me. I am very impressed that there was a record kept of each passengers' name, survival, and even as specific as their ticket number. Although I find the ticket number to be irrelevant data to build on, I think it is one of the most meaningful pieces of data to record. After viewing this dataset, it is interesting to notice the relationship between ticket class and survival of the passenger. I was expecting to find a closer relationship between age and survival. These struck me as interesting statistics that I would like to explore more.

From the data provided, I would like to not only notice the relationship between class ranks and survival, but I would really like to find the averages of survival for each class and compare each one. The dataset also does not clarify how many of the passengers were staff or employees of the ship, and I would like to find the average survival rate of that group as well. The data in this dataset is not clearly pointing to any conclusions, just simply providing the information that was recorded or accessible after the crash. In terms of data collected, it is just the "tip of the iceberg," if you will.

Variable	**Definition**	**Key**
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q =Queenstown, S = Southampton

Embarkation port	Frequency	Relative Frequency
Cherbourg	4	0.13
Queenstown	4	0.13
Southhampton	22	0.73
Total	30	1

Age	Frequency	Relative Frequency
0-15	7	0.29
16-30	6	0.25
31-45	8	0.33
46-60	3	0.08
Total	24	1

The first table provides insight to the boarding location of each passenger. The boarding location provides insight to how long the passengers had been aboard the titanic and help in the process of tracking down where everyone came from before boarding the ship. This data shows that most passengers boarded at the Southhampton port.

The second table shows the various ages of passengers aboard the ship. Although most ages are recorded, some were unobtainable and prevent a completely accurate summary of passenger age. As shown by the data, the smallest age group present is the 46-60. The younger ages make up the majority of the population, ranging from ages 1-45.

	Survived	Did not survive	**Total**
3rd class	7	12	19
1st & 2nd class	8	3	11
Total	15	15	

This two way table breaks down the number of survivors based on which class the passenger belonged to. I chose to group the 1st & 2nd class as their numbers were smaller that the 3rd class, even when grouped together. This data suggests that a larger number of 3rd class passengers did not survive in comparison to both the 1st and 2nd class passengers combined.

The average cost of Fare for tickets onto the Titanic was $28.83, with a standard deviation of 46.21.

5 point summary:

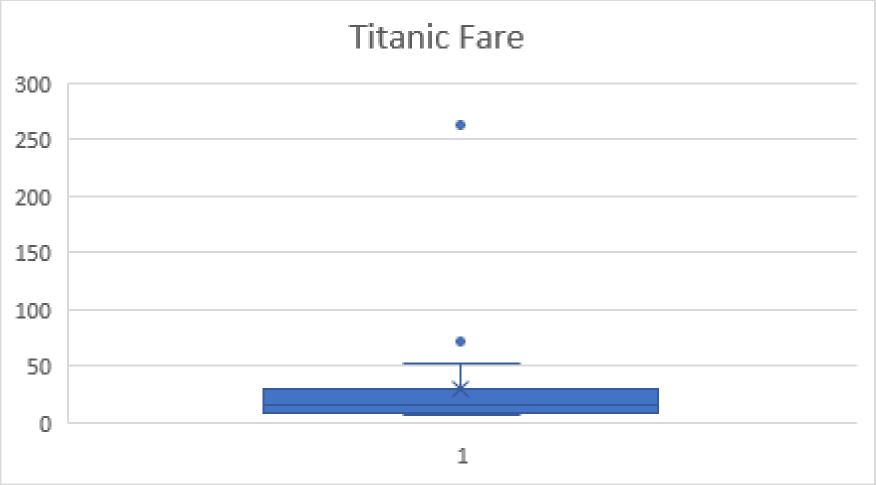
Minimum – 7.225

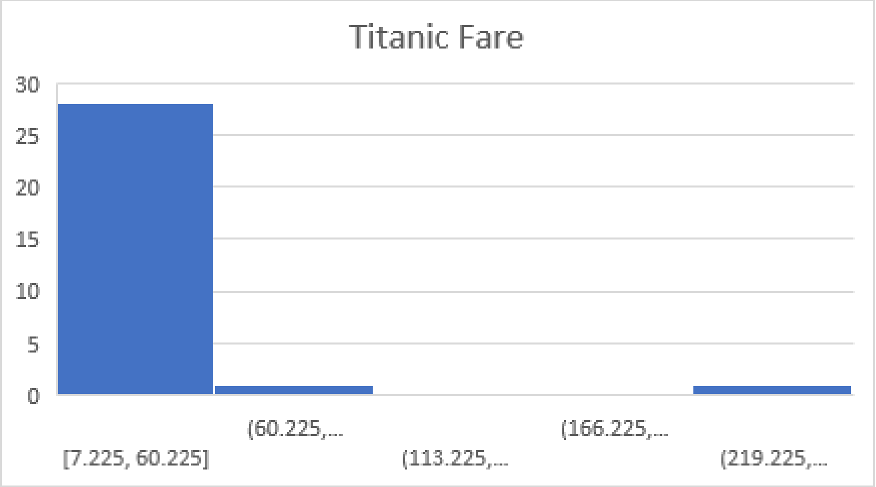
Q1 – 8.034

Median – 16.34

Q2 – 29.83

Maximum – 263





The box plot for this data is small and difficult to read due to the relatively close Q1 and Q3, paired with an extreme outlier. It does accurately represent the 5 point summary provided above the tables. The outliers are 71 and 263. The outliers exist because of luxurious packages which granted access to extravagant delicacies that only a few could afford. The drastic difference caused by the outlier can be explained by the difference in experiences of the most expensive ticket and the cheapest.

The histogram represents the ticket fare between 7.225-60.225 was very popular, with only a limited number of passengers paying more for their ticket. This graph is skewed right.

Quantitative Hypothesis

$H_{0:}\ \mu = 30.$

Null: The average age of passengers will be 30.

$H_{a:}\ \mu \neq 30.$

Alternative: The average age of passengers will not equal 30.

I made the hypothesis after reviewing the ages provided in the original dataset. Although some passengers were very young, and some were older adults, 30 seemed to be a good assumption. I also believe that age 30, give or take a few years, is a reasonable mean age to afford and attend a cruise such as the titanic was.

Categorical Hypothesis

$H_{0:}\ P > 32\%$ *(Overall survival)*

Null: The proportion of 1st class survivors will be greater than the proportion of overall survival percentage.

$H_{a:}\ P < 32\%$ *(overall survival)*

Alternative: The proportion of 1st class survivors will be less than the proportion of overall survival percentage.

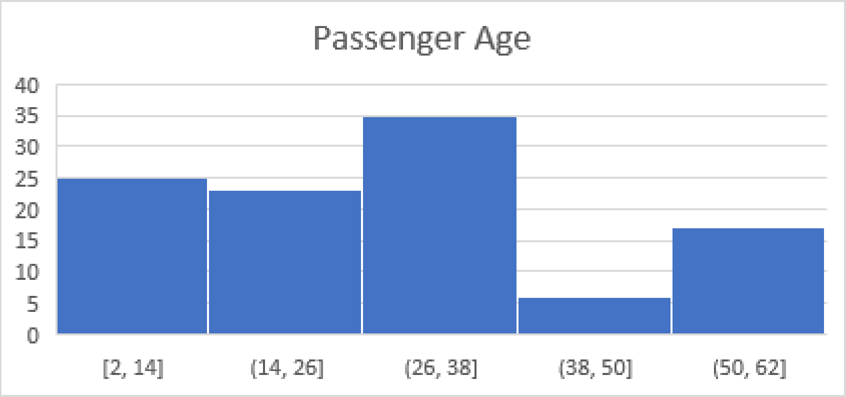
I made my hypothesis based off the knowledge that the 1st class was provided cabins in the upper deck and levels, which meant a shorter trip to safety boats. Also, some 1st class passengers were viewed as a priority to staff, even amid an emergency which provided extra opportunity for safety. My two way table provided insight for this hypothesis.

Quantitative Hypothesis

Standard Error: 3.07

95% CI: between 19.65 and 34.16

The Histogram below was produced from the bootstrap data for the testing of my quantitative hypothesis, It shows a large favor in the ages 28-41, while the mean of bootstrap data is 26, as shown in the attached datasheet. These averages are very close to my hypothesis age of 30, but not quite specific enough to prove my hypothesis without doubt. Therefore, my null hypothesis is rejected.

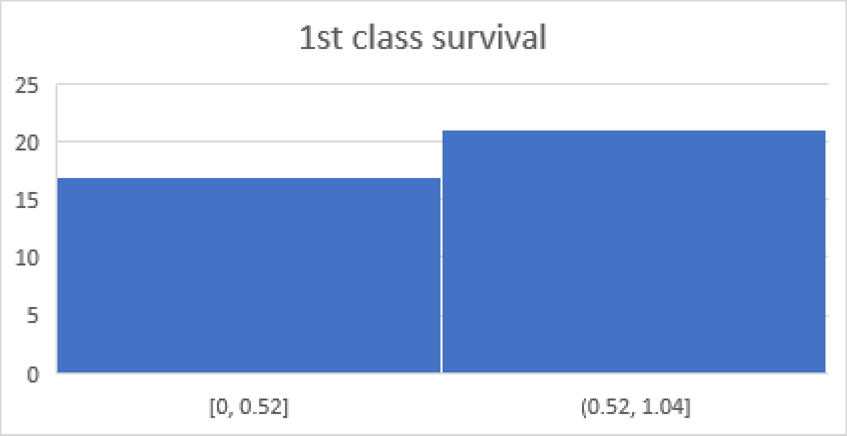


Categorical Hypothesis

Standard Error: 0.22

95% CI: between 33% and 99%

The histogram below was produced from the bootstrap data for the testing of my categorical hypothesis. The histogram, along with the bootstrap average of 66% of first class passengers survival, the data shows very clearly the average is much higher than the overall survival of 32%. Therefore, my null hypothesis does not fail.



Categorical Hypothesis Testing by Formula

My initial categorical hypothesis will remain the same while testing using formulas.

H_0: P > 32% (Overall survival)

Null: The proportion of 1st class survivors will be greater than the proportion of overall survival percentage.

H_a: P < 32% (overall survival)

Alternative: The proportion of 1st class survivors will be less than the proportion of overall survival percentage.

The Standard Error produced by this hypothesis testing technique is 0.1904 and my Z statistic is 1.837868.

The 95% confidence interval is between 0.317196 and 1.022804.

After examination of the 95% confidence intervals, the Null Hypothesis would fail as the lower bound is below the hypothesized proportion of 32%.

In comparison to the 95% interval produced by bootstrapping (0.33 and 0.99), the results are very similar. The lower bound of the 95% interval from formula testing is only .02 more specific, although the upper bound was larger. Although the confidence intervals were similar, they did not result in the same conclusion pertaining to the hypothesis.

The formulas I used to provide these statistics are as follows;

SE = sqrt (p(1 − p) nc + p(1 − p) nn) & Z = point estimate − null value SE

Quantitative Inference with Formulas

The Quantitative hypothesis I will be testing is the same as stated earlier,

$H_{0:}$* μ = *30.

Null: The average age of passengers will be 30.

$H_{a:}$* μ ≠ *30.

Alternative: The average age of passengers will not equal 30.

To calculate the 95% Confidence interval for this hypothesis, I first had to find the standard deviation, standard error, and T value.

SD = 15.98635

SE = 3.19727

T = -0.87575

These values were used in the formula to find my 95% CI which is between 20.93335 and 33.46665.

The formula used to find this is, point estimate ± z* × SE

These statistics, in comparison to my bootstrapping statistics, vary only slightly. The 95% confidence interval range was higher from bootstrapping. As 30 falls between both confidence intervals, my hypothesis passed using the bootstrap technique but failed using inference formulas.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

To create conditional probabilities, I used the two way table previously presented in project 2 (shown below) and the conditional probability formula (shown above).

	Survived	Did not survive	**Total**
3rd class	7	12	19
1st & 2nd class	8	3	11
Total	15	15	30

1. The probability of 3rd class given they did not survive. **P = 12/15 ≈ 80%**
2. The probability of 1st and 2nd class given they did survive. **P = 8/15 ≈ 53%**

To analyze my initial assumption that the class rank affected the survival of passengers, both of my conditional probabilities focused on the probability of survival being related to their class.

The first conditional probability exemplifies the probability of 3rd class passengers that did not survive from the sample population used throughout the project. The overall percentage was approximately 80%, meaning that 12 out of 15 3rd class passengers did not survive the shipwreck. This statistic shows the disadvantage that 3rd class passengers faced, in terms of survival chances. I came to this conclusion by finding the intersection of the 3rd class probability and did not survive probability, then dividing by the given statistic.

The second conditional probability exemplifies the probability of the combined total of 1st and 2nd class survivors from the sample population. The overall percentage was 53%, meaning that 8 out of 15 upper class passengers survived the shipwreck. This statistic goes to prove that over half of the 1st and 2nd class passengers from the sample population survived, which cannot be said for the 3rd class. I came to this conclusion by locating the intersection of 1st and 2nd class probability and survived probability, then dividing by the given statistic.

I believe it is important to note that the overall population of 3rd class population was much higher than the other classes combined, but there were still more survivors from the upper classes. In conclusion of my conditional probability analysis, I believe the statistical testing used throughout this project validates that the chance of survival has a clear relationship with the passenger class.