As an overall theme, the present report aims to illuminate the differences that existed between children travelling in the Titanic disaster in 1912; and how these differences correlated with their chances of surviving the disaster.

More specifically, the following independent inquiries (that relate to the overall theme and give valuable universal insights) and research questions (more specific and hypothesis-driven) will be answered:

1. What was the mean age of children (defined as being less than 12 years old), $\bar{x}_{childrenAge}$, in the Titanic embarkation, with an $(1 - \alpha)\%$ confidence level? ($\alpha = 0.05$)
2. Was there a statistically significant difference between the mean age of children who survived ($\bar{x}_{survivingChildrenAge}$) and the mean age of children who did not survive ($\bar{x}_{nonSurvivingChildrenAge}$)? What would be the effect size?
3. What was the probability of a passenger to survive given that he was a child and was travelling in third class $P(Survive = 1 \mid Age < 12 \ \& \ Pclass = 3)$? What if instead of third class he was travelling in first class $P(Survive = 1 \mid Age < 12 \ \& \ Pclass = 1)$?
4. How many trials would it have taken for a parent to ensure an escape route for her child in first class in comparison to the ones travelling in third class.
5. How many trials would a parent in third class would have needed to ensure making it likely with a 90% of probability that at least three of their children would survive, in comparison to the amount of trials parents in first class needed?

Important notice: Refer to each section for context on the underlying assumptions and conventions being made.

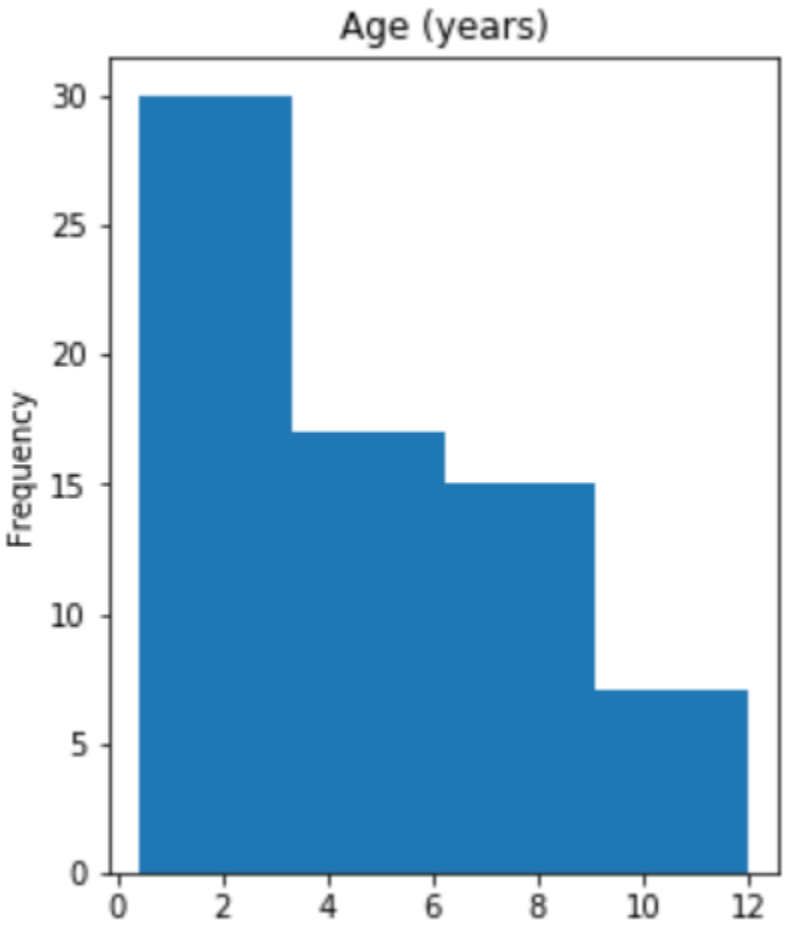
1. Independent Inquiry: What was the mean age of children (defined as being less than 12 years old), $\bar{x}_{childrenAge}$, in the Titanic embarkation, with an $\alpha = 0.05$ level of significance?

Assumptions or definitions:

- Children are defined as passengers whose age is less than or equal to 12 years.

First, let us define the sample size n as the *subsample size of only children* taken from the originally available sample from the population in the Titanic (originally not only children).

The distribution of children age follows this pattern.



As it can be observed, the distribution is not normal. However, the principles of the central limit theorem allow us to estimate the mean of the population of children, $\mu_{childrenAge}$, as the sampling distribution of the sample mean is, conceptually, normal.

$$confidenceInterval = \bar{x} \pm z * SE$$

Where,

$\bar{x}_{childrenAge}$ is the sample mean, the point estimate.

z is the z-score that corresponds to the $\left(1 - \frac{\alpha}{2}\right)$ $percentile$.

$SE = \frac{\sigma_{x_{childrenAge}}}{n^2}$ is the **Standard Error of the *Sampling Distribution of the Sample Mean***. Notice that $\sigma_{x_{childrenAge}}$ is used instead of $\sigma_{\mu_{childrenAge}}$, since we only have a sample to work with. This is acceptable since the distribution of a representative sample should resemble the one of the population, and therefore the standard deviation. (Standard deviation of the population and/or sample are not to be confused with the concept of standard error of the sampling distribution of the sample mean.)

Estimate of children's mean Age in Titanic, with a 95% confidence level (see Appendix A, *'Mean-Age-Of-Children-Groups-Estimation'* for computations):

$$\mu_{childrenAge} = 4.77 \pm 0.8\ years$$

2. Research question: Was there a statistically significant difference between the mean age of children who survived ($\bar{x}_{survivingChildrenAge}$) and the mean age of children who did not survive ($\bar{x}_{nonSurvivingChildrenAge}$)? What would be the effect size?

   - $H_A : \bar{x}_{survivingChildrenAge} \neq \bar{x}_{nonSurvivingChildrenAge}$
   - $H_0 : \bar{x}_{survivingChildrenAge} = \bar{x}_{nonSurvivingChildrenAge}$

For answering this question, we must perform a hypothesis test for the difference between two means often called the two-sample t-test. Particularly, a two-tailed test since an extreme value on either side of the sampling distribution, of the mean difference of means, would imply rejecting the null hypothesis.

We will proceed with a significance level of $\alpha = 0.05$

First, we may find the standard error SE of the sampling distribution of the mean difference of sample means as follows,

$$SE = \sqrt{\frac{\sigma^2_{survivingChildrenAge}}{n_{survivingChildrenAge}} + \frac{\sigma^2_{nonSurvivingChildrenAge}}{n_{nonSurvivingChildrenAge}}}$$

Second, we may approximate the degrees of freedom by taking the smaller of ($n_{survivingChildrenAge} - 1$) and ($n_{nonSurvivingChildrenAge} - 1$

Third, we may compute the test statistic by using the following equation for determining a t statistic. (Notice brackets to indicate absolute value).

$$t = \frac{\left[\bar{x}_{nonSurvivingChildrenAg} - \bar{x}_{survivingChildrenAge}\right]}{SE}$$

By assessing the probability associated with the t statistic with any sort of statistical software, we will find the probability of observing a sample statistic as (or more) extreme as the test statistic, which will be the p-value. In particular, since this is a two-tailed hypothesis test, we must be careful of inserting the right parameters in the t distribution calculator of function and/or using the result adequately (Please

review Appendix A, ‘SignificanceTest-Of-Surviving-and-NonSuriving-ChildrenMeanAge-Difference’ to see computations)

With this method, the P-value turns to be:

$$pvalue = 0.0145$$

Which is far less than the significance level $\alpha = 0.05$, meaning that **the difference of means is statistically significant.**

Concerning the effect size, we may use Cohen’s d to measure it as follows.

$$d = \frac{\bar{x}_{nonSurvivingChildrenAge} - \bar{x}_{survivingChildrenAge}}{pooled\ standard\ deviation}$$

Where the pool standard deviation may be computed as

$$SDpooled = \sqrt{\frac{\sigma^2_{survivingChildren}\left(n_{suvivingChildren} - 1\right) + \sigma^2_{nonSurvivingChildren}\left(n_{nonSuvivingChildren} - 1\right)}{n_{survivingChildren} + n_{nonSurvivingChildren} - 2}}$$

With this method, Cohen’s d turns to be:

$$d = 0.65$$

Which may **be considered an effect size from *Medium* to *Large***, according to the classification provided by (Sawilowsky, 2009).

Please review Appendix A, ‘SignificanceTest-Of-Surviving-and-NonSuriving-ChildrenMeanAge-Difference’ to see each of the computations referenced on this section.

3. Independent Inquiry: What was the probability of a passenger to survive given that he was a child and was travelling in third class $P(Survive = 1 \mid Age < 12\ AND\ Pclass = 3)$? What if instead of third class he was travelling in first class $P(Survive = 1 \mid Age < 12\ AND\ Pclass = 1)$?

We know that joint dependent probabilities may defined as follows,

$$P(Survive = 1 \mid Age < 12\ AND\ Pclass = 3) * P(Age < 12\ AND\ Pclass = 3) = P(Survive = 1\ AND\ Age < 12\ AND\ Pclass = 3)$$

Which, according to Bayes theorem

$$P(Survive = 1 \mid Age < 12\ AND\ Pclass = 3) = \frac{P(Survive = 1\ AND\ Age < 12\ AND\ Pclass = 3)}{P(Age < 12\ AND\ Pclass = 3)}$$

Since we have a sample with information about the passengers that include their survival, age, and class; we may compute the joint probabilities of $P(Survive = 1\ AND\ Age < 12\ AND\ Pclass = 3)$ and $P(Age < 12\ AND\ Pclass = 3)$ by simply computing the proportion of passengers that meet such criteria in the sample.

Naturally, the probabilities of joint events are always less than the events alone. So, the values of the $P(Survive = 1\ AND\ Age < 12\ AND\ Pclass = 3)$ are $P(Age < 12\ AND\ Pclass = 3)$ will turn to be excessively small. This may seem counterintuitive, but it shouldn't. A way to see it may be to think: "What is the probability of finding someone with each of these characteristics together out of the full population?" It is indeed very small and not exactly what we are framing with the current research question.

What we frame with the research question involves the condition of *knowing already that we are dealing with a passenger that is a child and is travelling on a certain class.* For such reason, this probability will have a much more reasonable value. Still, mathematically it is obvious that when dividing two excessively small values it follows a ratio that may not be as excessively small.

By following the previous methodology, we get

$P(Survive = 1\ AND\ Age < 12\ AND\ Pclass = 3)$ = 0.05387 (excessively small, as expected)

$P(AND\ Age < 12\ AND\ Pclass = 3)$ = 0.02245 (excessively small, as expected)

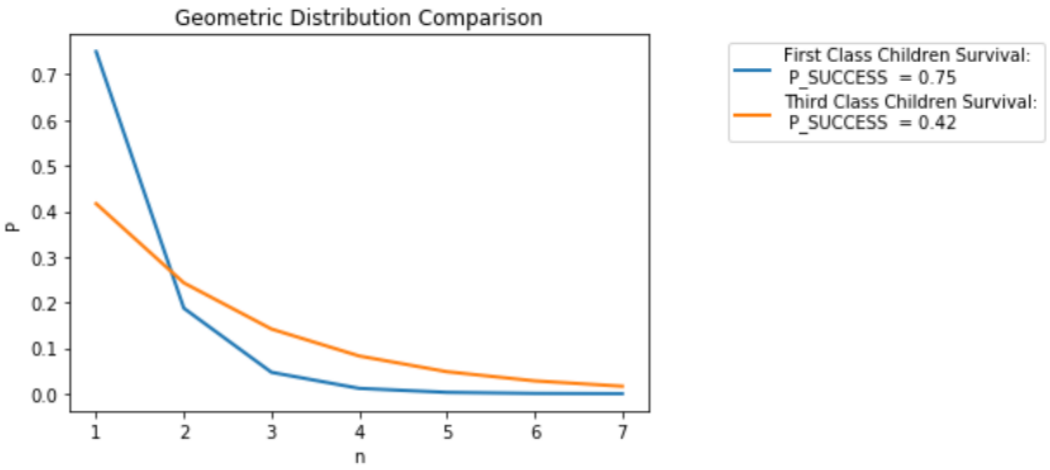
$$P(Survive = 1 \mid Age < 12\ AND\ Pclass = 3) = 0.4167$$ (tractable value, as expected)

Similarly, applying the same procedure with passengers in first class

$$P(Survive = 1 \mid Age < 12\ AND\ Pclass = 1) = 0.7500$$ (tractable value, as expected)

Please review Appendix A, ‘ConditionalProbability-PassgersSurvival-ByClassAndAge’ to see each of the computations referenced on this section.

4. Independent Inquiry:
    - For simplicity, let us:
        - Model the attempt of parents to ensure a boat to escape for the child, as a random Bernoulli trial X, where success is ensuring a boat, and failure is not.
        - Assume the probability of success $P_{success}$ of this random Bernoulli trials may resemble the probability of a child to survive, calculated as the conditional probability of surviving given that the passenger was a child and was travelling on a certain class $\Pr(Survive \mid Age < 12 \ \& \ Pclass = \ c)$
        - Assume $P_{success\ (X = x_i)}$ for each trial is Independent and Identically Distributed in comparison to any other trial.
    - To find:
        - The geometric distributions of these random Bernoulli variables, as a way to determine how many trials would it have taken for a parent to ensure an escape route for her child in first class in comparison to the ones travelling in third class.
    - Other strong assumptions and limitations:
        - Notice we assume that it makes no difference whether a child had one parent or two, siblings, or any other features
    - Still:
        - By considering how a highly relevant feature such as Pclass relates to survival, we can produce valuable insights.
    - Results:



Please review Appendix A.

5. Independent Inquiry:
    - Context for inspiration:
        - A parent may have more than one children, so how many trials would parent in third class would have needed to ensure making it likely with a 95% of probability that at least three of their children would survive, in comparison to the amount of trials parents in first class needed?
    - **For simplicity:**
        - Let us assume each trial may be identically and independently distributed.
        - Let us assume, for the sake of simplicity, that we may ignore that parents may not have at least three children. Although this is a strong assumption, it allows to prove, or disprove, the following point: The attempts that parents travelling in third class needed to ensure an escape route where more than the ones parents travelling in first class.
    - Results:
        - 6 trials for families in first class.
        - 13 trials for families in third class.

Please review Appendix A.