Titanic Survival Analysis

A Comprehensive Statistical Report

July 25, 2025

Contents

1 Project Overview

This project explores the factors influencing passenger survival aboard the RMS Titanic using statistical and inferential methods. Data preprocessing, exploratory data analysis (EDA), hypothesis testing, and confidence intervals were applied to derive meaningful conclusions. The analysis was conducted in both Microsoft Excel and Python for maximum accessibility and reproducibility.

2 Module 1: Data Preparation and Cleaning

Objective: Prepare a clean dataset suitable for analysis.

Key Steps:

- Removed irrelevant features (`Cabin`, `Ticket`).
- Handled missing values:
  - `Age` imputed using the median due to right-skewness.
  - `Embarked` filled using mode.
- Encoded categorical variables (`Sex`, `Embarked`, `Pclass`).

Outcome: A well-structured, cleaned dataset ready for further statistical analysis.

3 Module 2: Descriptive Statistics & Visualization

Descriptive Summary:

Variable	Mean	Median	Mode	Std. Dev.
Fare	32.20	14.45	8.05	49.69
Age	29.70	28.00	24.00	14.53

Table 1: Summary Statistics for Fare and Age

Insights:

- **Fare** is heavily right-skewed due to high outliers (first-class passengers).
- **Median** is a better measure of central tendency for skewed data.
- **Age** is roughly normally distributed.

Visualizations: Histograms and box plots revealed outliers and data spread effectively.

4 Module 3: Survival Rate by Gender and Class

Survival Rate by Class

Class	Survival Rate
1st	63%
2nd	47%
3rd	24%

Survival Rate by Gender

Gender	Survival Rate
Female	74%
Male	19%

Interpretation: Significant survival bias based on gender and class suggests social and structural inequalities during evacuation.

5 Module 4: Correlation Analysis

Variable Pair	Pearson Correlation (r)
Fare vs. Survival	+0.26
Age vs. Survival	–0.08

Table 2: Correlation Coefficients

Interpretation:

- **Fare** positively correlates with survival—supporting the idea that wealthy passengers had better survival chances.
- **Age** shows minimal impact on survival.

6 Module 5: Estimating Survival Rate with Sampling

Task: Randomly sample 50 passengers and estimate survival rate.

Observation:

- Sample estimates varied (e.g., 36% to 44%).
- Sampling bias evident in small datasets.
- Reinforces the importance of larger, randomized samples.

7 Module 6: Hypothesis Testing – Gender Bias in Survival

Hypothesis:

- H_0: No survival difference between men and women.
- H_1: Women had a significantly higher survival rate.

Statistical Techniques

- **MLE Estimates:**
  - Female: 0.74
  - Male: 0.19
- **Two-sample T-test:** p-value $< 0.001 \Rightarrow$ Reject H_0

MLE vs Bayesian:

- MLE uses observed data alone to find the most likely parameter.
- Bayesian estimation incorporates prior knowledge and updates beliefs.

8 Module 7: Confidence Intervals & Hypothesis Testing

Survival Probability (95% CI): [34.5%, 42.1%]
T-Test Result: Gender-based survival rates are statistically different. p-value < 0.05.
Conclusion: Hypothesis that women had higher survival rate is accepted.

9 Module 8: A/B Testing and Final Report

A/B Testing Between 1st and 3rd Class:

- 1st Class Survival Rate: 63%
- 3rd Class Survival Rate: 24%
- p-value $< 0.01 \Rightarrow$ Statistically significant difference.

Conclusion: Passenger class played a major role in survival.

10 Final Summary and Insights

Aspect	Insight
Gender	Women had far higher survival rates than men.
Class	1st class passengers were prioritized.
Fare	Positively correlated with survival.
Age	Slight negative correlation with survival.
Visualization	Box plots and histograms essential for interpretation.
Hypothesis Tests	Confirmed gender and class-based disparities.

Table 3: Key Findings

11 Skills Demonstrated

- Data cleaning and wrangling
- Descriptive and inferential statistics
- Correlation and confidence intervals
- Hypothesis testing (T-test, MLE)
- A/B testing and sample estimation
- Visual analytics (Excel & Python)

Final Notes

This comprehensive report reflects not only individual outcomes aboard the Titanic but highlights systemic inequalities in crisis situations. By combining statistical rigor and visual tools, we derived historically consistent insights with modern data science methods.