Analysis of Airplane Crashes and Fatalities: A Statistical Exploration

Ramkumar Perumalagaram Subramanian

Bellevue University

DSC 530 - Data Exploration and Analysis

Fadi Alsaleem

March 1, 2025

Analysis of Airplane Crashes and Fatalities: A Statistical Exploration

Statistical and Hypothetical Question

The primary objective of this study was to explore trends and factors influencing airplane crashes and fatalities. The statistical question that guided the analysis was: “Has the survival rate of airplane crashes improved over time?” Additionally, a secondary hypothesis was tested to determine whether airplane crashes are evenly distributed across different countries, using a Chi-Square test.

Exploratory Data Analysis (EDA) Outcomes

The dataset contained records of airplane crashes from 1908 to present, with variables such as Date, Location, Operator, Aboard, Fatalities, and Ground fatalities. Through histograms and descriptive statistics, we identified that most crashes had low fatalities, but a long tail of high-fatality events existed. A Probability Mass Function (PMF) revealed that military crashes tend to have a higher likelihood of complete fatalities compared to commercial crashes. A Cumulative Distribution Function (CDF) demonstrated that most crashes resulted in low fatalities, with a small percentage being catastrophic.

A Chi-Square Test confirmed that certain countries experience significantly more crashes, rejecting the assumption of uniform crash distribution. Finally, a linear regression analysis of Fatalities vs. Aboard showed a strong positive correlation, suggesting that larger flights tend to result in higher fatalities, though other factors such as safety measures and aircraft type likely contribute.

Missed Opportunities in the Analysis

While the analysis covered a broad range of statistical techniques, weather conditions, aircraft age, and maintenance records were not included due to dataset limitations. These factors could provide better insights into crash severity and survival rates. Additionally, analyzing crash cause summaries could have identified patterns related to human error, technical failures, or external conditions.

Potentially Useful Variables

Variables such as flight phase (takeoff, landing, mid-air), weather conditions at the time of crash, aircraft age, and air traffic density could have significantly improved the analysis. For example, incorporating weather conditions could help determine if bad weather significantly impacts crash survival rates.

Incorrect Assumptions

One of the initial assumptions was that fatalities follow a normal distribution; however, the data was highly skewed, requiring log-transformation or alternative distribution fitting. Additionally, assuming crashes were uniformly distributed across countries was incorrect, as confirmed by the Chi-Square test.

Challenges and Areas for Further Understanding

One challenge was handling highly skewed data with extreme outliers. Future analysis should consider non-linear models to better capture relationships. Another challenge was ensuring proper statistical significance interpretations, especially in hypothesis testing and p-values.

Conclusion

The analysis confirmed that survival rates have improved over time, but there are still disparities in crash distributions and fatality causes. Future research should integrate external factors such as weather, aircraft maintenance, and regulatory measures to create a more comprehensive safety analysis.

References

Airplane crashes and fatalities. (n.d.). <https://www.kaggle.com/datasets/thedevastator/airplane-crashes-and-fatalities/data>