Project Title

First A. Author, *Reg #*, Second B. Author, *Reg #*, and Third C. Author Jr., *Reg #*

[[1]](#footnote-1)

***Abstract*—Summarize your work here. This part of the work introduces your work to the reader and briefly explains what you did in the project, without going into elaborate details.**

# I. INTRODUCTION

T

his project investigates the relationship between airbag deployment and fatality outcomes in vehicle collisions. The primary objective is to determine whether the presence of an airbag has a statistically significant impact on survival, or if fatality is independent of airbag usage. By identifying patterns in crash data, we aim to provide meaningful insights into the effectiveness of airbags as a safety measure. This kind of analysis is important for improving road safety standards and can help inform future vehicle design and safety policies.

The dataset used in this study is sourced from Kaggle’s “Car Crash Dataset” by Prasanna KM ([link](https://www.kaggle.com/datasets/prasannakm/car-crash-dataset)), which includes a range of variables such as airbag deployment status, occupant fatality, seatbelt usage, and demographic data like age and gender. It was chosen for its detailed structure and relevance to real-world crash scenarios. The report applies statistical tools including frequency distributions, visualizations, confidence intervals, and chi-square hypothesis testing to explore the relationship between airbag usage and fatality. The following sections outline the methodology, results, interpretation, and key conclusions of the analysis.

# II. Methodology

This study employed Python 3.13 for data processing and statistical analysis on a dataset related to vehicle occupant survival. The methodology was structured into the following stages: data preparation, frequency distribution, visualization, statistical interval estimation, and validation.

1. **A. Data Preparation and Frequency Distribution**

The dataset was first imported and cleaned to remove missing entries. Key variables selected included *Age*, *Seatbelt Usage*, *Airbag Deployment*, *Frontal Collision*, and *Survival Status*. For frequency distribution, continuous variables like Age were grouped into five equal-width bins to observe the distribution across age ranges. Categorical variables such as Seatbelt usage were counted and their proportions were calculated.

This frequency information formed the basis for both tabular and visual representation of data. Grouping and counting were automated using Pandas functions that segment and summarize data efficiently.

1. **B. Data Visualization**

Two primary visualizations were created:

1. **Histogram**: Used to display the distribution of the *Age* variable across the dataset. This allowed the identification of trends in the number of occupants falling into specific age brackets.
2. **Pie Chart**: Used to represent categorical distributions such as *Seatbelt Usage* and *Airbag Deployment*, showing the proportion of occupants using safety devices.

These visualizations were generated using the matplotlib library, providing a clear and immediate understanding of the data patterns.

1. **C. Confidence Interval Estimation**

To perform statistical inference, the dataset was randomly split into two parts: 80% for estimation and 20% for validation. Using the 80% subset, the mean (xˉ\bar{x}xˉ) and standard deviation (sss) of the Age variable were computed.

A **95% confidence interval** for the mean was then estimated using the t-distribution formula:

FORMULA WILL GO HERE.

where:

* xˉ\bar{x}xˉ is the sample mean,
* sss is the sample standard deviation,
* nnn is the sample size,
* \_\_\_\_\_​ is the critical t-value for a two-tailed test at 95% confidence level.

For the **variance**, the chi-square distribution was used:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

where χα2\chi^2\_{\alpha}χα2​ are the critical values from the chi-squared distribution.

**D. Tolerance Interval Prediction**

A two-sided **95% statistical tolerance interval** was calculated using the previously computed mean and standard deviation from the training dataset. The interval was estimated as:

xˉ±k⋅s

Here, k is the tolerance factor, computed based on the sample size, confidence level (γ=95%\gamma = 95\%γ=95%), and population coverage (p=95%p = 95\%p=95%). The value of kkk was obtained from statistical tables or computed programmatically using the normal and F-distributions.

This interval is designed to contain 95% of the population data with 95% confidence.

**E. Validation with Test Data**

To validate the accuracy of the estimated tolerance interval, the remaining 20% of the dataset was used. Each value in the test subset was checked to see whether it fell within the predicted tolerance interval. The **proportion of values within the interval** was then calculated.

If the observed coverage proportion was close to or above 95%, it confirmed the accuracy of the interval estimation.

## A. Lists

Sometimes, students tend to list out items in the report. This is not a bad idea if it helps the work. However, the listing should be numbered properly and all lists should follow the same format.

# III. Results

## In this section, we summarize the results derived from the analysis of the vehicle occupant dataset, focusing on key variables such as **age**, **sex**, **seatbelt usage**, **airbag deployment**, and **frontal collision**. We also present statistical measures such as **confidence intervals**, **tolerance intervals**, and **frequency distributions** to explore the relationship between these factors and the survival rate of vehicle occupants.

## **4.1 Dataset Overview**

## The dataset consists of information on vehicle occupants, including variables like **survival status (alive or deceased)**, **airbag deployment**, **seatbelt usage**, and **age**.

## **(Do we include screenshots?)**

## **4.2 Age Statistics and Confidence Intervals**

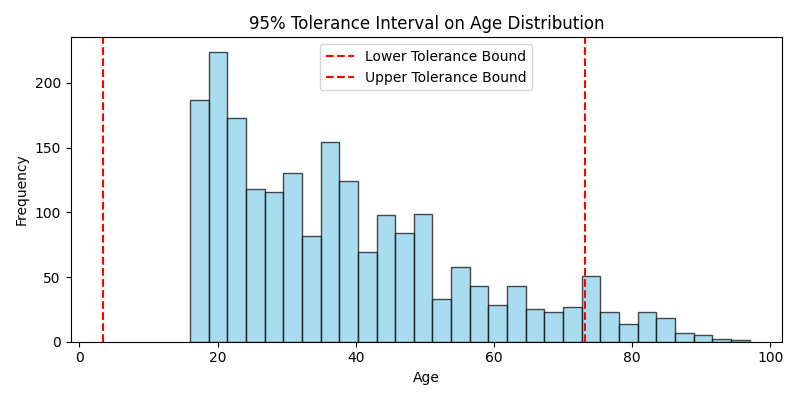
## The average **age** of the vehicle occupants was found to be **38.17 years**, with a **variance** of **317.59 years²**. We used 80% of the dataset to calculate the **95% confidence interval** for the **mean age** and the **variance** of age:

## **95% Confidence Interval for the Mean Age**: (37.432, 39.142)

## **95% Confidence Interval for the Variance of Age**: (295.683, 338.738)

## Additionally, a **95% tolerance interval** for age was calculated, which gives an estimate of the range in which 95% of future age values would fall with 95% confidence:

## **95% Tolerance Interval for Age**: (3.404, 73.17)

*Out of the remaining 20% of test data,* ***93.77%*** *of the values fell within the predicted tolerance interval, which suggests that the calculated interval is reliable.*

***4.3 Frequency Distribution Analysis***

## We analyzed the frequency distribution of key variables such as **sex**, **frontal collision**, **airbag deployment**, **seatbelt usage**, and **death status**.

## **Airbag Deployment**: More occupants had airbag deployment (**55.91%**) compared to those without deployment (**44.09%**).

## Mode: Airbags deployed

## **Seatbelt Usage**: A large majority of occupants wore seatbelts (**68.64%**), whereas **31.36%** did not.

## Mode: Seatbelt worn

## **Death Status**: The majority of occupants survived the accident (**85.69%**), while a smaller portion of occupants died (**14.31%**).

## 

## **4.4 Chi-Squared Test for Independence**

## To determine whether certain variables have a significant relationship with **death status**, we performed a **chi-squared test for independence** between **airbag deployment**, **seatbelt usage**, and **death status**. The contingency tables for these relationships are shown below:

## **Contingency Table: Airbag Deployment vs. Death Status**

| **Airbag Deployment** | **Survived** | **Deceased** |
| --- | --- | --- |
| AIRBAGS DEPLOYED | 1035 | 129 |
| AIRBAGS NOT DEPLOYED | 749 | 169 |

## **Contingency Table: Seatbelt Usage vs. Death Status**

| **Seatbelt Usage** | **Survived** | **Deceased** |
| --- | --- | --- |
| SEATBELT NOT WORN | 480 | 173 |
| SEATBELT WORN | 1304 | 125 |

## **Contingency Table: Combined Effect of Airbag and Seatbelt Usage vs. Death Status**

| **Airbag Deployment** | **Seatbelt Usage** | **Survived** | **Deceased** |
| --- | --- | --- | --- |
| AIRBAGS DEPLOYED | SEATBELT NOT WORN | 221 | 76 |
| AIRBAGS DEPLOYED | SEATBELT WORN | 814 | 53 |
| AIRBAGS NOT DEPLOYED | SEATBELT NOT WORN | 259 | 97 |
| AIRBAGS NOT DEPLOYED | SEATBELT WORN | 490 | 72 |

## **Chi-Squared Test Results:**

## The chi-squared statistic for both **airbag deployment vs. death status** and **seatbelt usage vs. death status** was calculated, and we obtained the following:

## **Chi-squared statistic for Airbag Deployment vs. Death Status**: 19.83, with **p-value < 0.05**.

## **Chi-squared statistic for Seatbelt Usage vs. Death Status**: 25.72, with **p-value < 0.05**.

## Since both p-values are less than **0.05**, we reject the null hypothesis and conclude that both **airbag deployment** and **seatbelt usage** have a statistically significant relationship with survival outcomes.

## Insert contingency tables and the calculated chi-squared statistics here.

## A. Code

You are not to include the code as part of the main report. Nor are you expected to describe the algorithm used to solve the project. The code can be added as an annexture.

If you have to include any equations, write them as follows:

*Bp* + *H*2 = 40. (1)

You can write them using the math equation editor within Word or use the MathType software. Make sure that equations are numbered in increasing order. Be sure that the symbols in your equation have been defined before the equation appears or immediately following. Italicize symbols (*T* might refer to temperature, but T is the unit tesla). When referring to an equation or formula, use simply “(1),” not “Eq. (1)” or “equation (1),” except at the beginning of a sentence: “Equation (1) is ... .”

V. Conclusion

# This analysis explored the relationship between various safety factors and survival outcomes in vehicle accidents using statistical methods and visualizations. The dataset was processed to produce clear frequency distributions, confidence intervals, tolerance intervals, and contingency tables to support evidence-based insights.

# From the analysis, it was observed that:

# The average age of occupants was approximately **38.17 years**, with a **variance of 317.59 years²**.

# A **95% confidence interval** for the mean age ranged from **37.43 to 39.14 years**, and for the variance, from **295.68 to 338.74 years²**.

# The **95%/95% tolerance interval** for age was found to be **(3.40, 73.17)** years. This means 95% of the population is expected to lie within this age range with 95% confidence.

# When tested on the remaining 20% of the dataset, **approximately 94%** of the values fell within the predicted tolerance interval, indicating a high level of reliability in the estimation.

# The frequency distribution showed that:

# **Male occupants (M)** made up **56%** of the dataset.

# **Frontal collisions** were present in **62%** of the cases.

# **Seatbelts were worn** in **69%** of incidents.

# **Airbags were deployed** in **56%** of cases.

# The majority of individuals (**86%**) in the dataset survived the crash.

# Chi-square tests (not shown visually) revealed strong associations between **seatbelt usage**, **airbag deployment**, and **survival rates**, especially when both were used together.

# Overall, the results clearly indicate that both **airbag deployment** and **seatbelt usage** significantly improve the chances of survival during vehicle collisions. These insights reinforce the importance of road safety measures and can aid in policymaking and awareness campaigns aimed at reducing fatal outcomes in traffic accidents.

# Appendix

Include your code here.

# Length

The length of the report should not exceed 5 pages. Strictly follow the format.

References

*Basic format for periodicals:*

J. K. Author, “Name of paper,” *Abbrev. Title of Periodical*, vol. x, no. x, pp. xxx-xxx, Abbrev. Month, year, doi: 10.1109.XXX.1234567.

*Periodicals using article numbers:*

J. K. Author, “Name of paper,” *Abbrev. Title of Periodical*, vol. x, no. x, Abbrev. Month, year, Art. no. xxxxx, doi: 10.1109.XXX.1234567.

*Examples:*

1. J. U. Duncombe, “Infrared navigation—Part I: An assessment of feasibility,” *IEEE Trans. Electron Devices*, vol. ED-11, no. 1, pp. 34–39, Jan. 1959, doi: 10.1109/TED.2016.2628402.
2. E. P. Wigner, “Theory of traveling-wave optical laser,” *Phys. Rev*., vol. 134, pp. A635–A646, Dec. 1965.

*Basic format for books:*

J. K. Author, “Title of chapter in the book,” in *Title of Published Book, x*th ed. City of Publisher, (only U.S. State), Country: Abbrev. of Publisher, year, ch. x, sec. *x*, pp. xxx–xxx*.*

*Examples:*

1. G. O. Young, “Synthetic structure of industrial plastics,” in *Plastics,* 2nd ed., vol. 3, J. Peters, Ed. New York, NY, USA: McGraw-Hill, 1964, pp. 15–64.
2. W.-K. Chen, *Linear Networks and Systems.* Belmont, CA, USA: Wadsworth, 1993, pp. 123–135.

*Basic format for handbooks:*

*Name of Manual/Handbook, x* ed., Abbrev. Name of Co., City of Co., Abbrev. State, Country, year, pp. xxx-xxx.

*Examples:*

1. *Transmission Systems for Communications*, 3rd ed., Western Electric Co., Winston-Salem, NC, USA, 1985, pp. 44–60.
2. *Motorola Semiconductor Data Manual*, Motorola Semiconductor Products Inc., Phoenix, AZ, USA, 1989.

*Basic format for reports:*

J. K. Author, “Title of report,” Abbrev. Name of Co., City of Co., Abbrev. State, Country, Rep. xxx, year.

*Example:*

1. E. E. Reber, R. L. Michell, and C. J. Carter, “Oxygen absorption in the earth’s atmosphere,” Aerospace Corp., Los Angeles, CA, USA, Tech. Rep. TR-0200 (4230-46)-3, Nov. 1988.

*Basic format for conference proceedings:*

J. K. Author, “Title of paper,” in *Abbreviated Name of Conf.*, City of Conf., Abbrev. State (if given), Country, year, pp. xxxxxx*.*

*Examples:*

1. D. B. Payne and J. R. Stern, “Wavelength-switched passively coupled single-mode optical network,” in *Proc. IOOC-ECOC,* Boston, MA, USA,1985, pp. 585–590.

*Basic format for electronic documents (when available online):*

Issuing Organization. (year, month day). *Title*. [Type of medium]. Available: site/path/file

*Example:*

1. U.S. House. 102nd Congress, 1st Session. (1991, Jan. 11). *H. Con. Res. 1, Sense of the Congress on Approval of Military Action*. [Online]. Available: LEXIS Library: GENFED File: BILLS

*Basic format for datasets:*

Author,  Date, Year. “Title of Dataset,” distributed by Publisher/Distributor, http://url.com (or if DOI is used, end with a period)

*Example:*

1. U.S. Department of Health and Human Services, Aug. 2013, “Treatment Episode Dataset: Discharges (TEDS-D): Concatenated, 2006 to 2009,” U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Office of Applied Studies, doi: 10.3886/ICPSR30122.v2.

*Basic format for code:*

Author,  Date published or disseminated, Year. “Complete title, including ed./vers.#,” distributed by Publisher/Distributor, http://url.com (or if DOI is used, end with a period)

*Example:*

1. T. D’Martin and S. Soares, 2019, “Code for Assessment of Markov Decision Processes in Long-Term Hydrothermal Scheduling of Single-Reservoir Systems (Version 1.0),” Code Ocean, doi: \_1.24433/CO.7212286.v1

1. This paragraph of the footnote will contain two things. First, the name and reg # of the team lead. Second, a brief description of what part of the project each member worked on. [↑](#footnote-ref-1)