Univariate Analysis of Titanic Dataset

Rajput, Kaustubh Rajendra

University of the Cumberlands

Course number: MSDS-530-A01

Afshin Zarenejad

January 12, 2025

Data Distribution and Summary Measures Analysis

Introduction

This document provides a comprehensive analysis of the Titanic dataset with a focus on data distribution and summary measures across different variable types. The analysis aims to uncover trends and relationships influencing passenger survival rates.

Nominal Variables

Survived: The survival rate of passengers is approximately 40%. This indicates that a large proportion of passengers did not survive the Titanic disaster. This variable can be analyzed by comparing survival rates across other categorical variables (e.g., Sex, Embarked).

Sex: The male-to-female ratio is roughly 2:1. This indicates that more male passengers were aboard compared to females. Analyzing this with survival data could help identify gender-based survival trends.

Embarked: Most passengers embarked from Southampton, which is expected as the Titanic was primarily sailing from England. This could provide insights into regional survival differences.

Cabin, Ticket, Name, PassengerId: These variables are unique identifiers or categorical variables, often used for grouping but not providing much insight individually without further context or aggregation.

Ordinal Variables

Pclass: This shows that the majority of passengers were from 3rd class, followed by 1st class and then 2nd class. The ordinal nature of this variable allows for analysis of survival rates by class, potentially showing how class affected the likelihood of survival.

Interval Variables

BirthYear: As an interval variable, this provides the passenger's age when the Titanic sank. However, since it’s an interval variable without a true zero point, its analysis focuses on the age distribution of passengers.

Ratio Variables

Fare: The average fare was around 32.2, but the median is much lower at 14.45, suggesting a skewed distribution. The mode (8.05) further supports that a large number of passengers paid a lower fare. The wide standard deviation (49.67) and variance (2466.67) suggest a significant spread in fare values.

SibSp: The mean is 0.52, indicating that most passengers traveled alone or with one sibling/spouse. The mode (0) shows that many passengers did not have siblings/spouses aboard.

Parch: The mean is 0.38, with a mode of 0, suggesting most passengers did not travel with parents or children.

Influence of Data Types on Analysis

Nominal Data: These variables (e.g., Survived, Sex, Embarked) are categorical, and the analysis focuses on frequencies and proportions. You can compare the distribution of nominal variables with survival rates or other key outcomes. For example, you might want to compare the survival rate of male vs. female passengers or passengers from different embarkation points.

Ordinal Data: Pclass is an ordered categorical variable, and its analysis can reveal trends based on rank. For instance, higher-class passengers (Pclass 1) might have had a higher survival rate than those in Pclass 3. Since the intervals between classes are not equal, non-parametric statistical methods are typically used.

Interval/Ratio Data: Fare, SibSp, and Parch are numerical variables where the difference between values is meaningful. For ratio variables, both differences and ratios are meaningful, so you can perform more complex statistical analyses such as calculating means, standard deviations, and performing regression analysis. These variables can also help in exploring relationships with survival, such as whether higher fares or traveling with family members increased the chances of survival.

Reflection: Importance of Understanding Data Types

Understanding the type of data is critical in choosing the appropriate descriptive statistics and analysis techniques. For nominal and ordinal data, categorical analysis methods (e.g., frequency distributions, chi-square tests) are suitable, while interval and ratio data require measures of central tendency (mean, median, mode) and dispersion (standard deviation, variance) to understand distributions. Recognizing the type of data helps avoid inappropriate statistical methods, leading to more accurate and meaningful insights. For instance, using the mean for ordinal data could be misleading, whereas the median or mode is more appropriate.

By selecting the right statistical methods, data analysts can uncover trends, relationships, and insights that drive informed decision-making. This process is fundamental to solving real-world problems, such as understanding passenger survival rates based on different factors in the Titanic dataset.