Data Science Lab-I Report
Frequency Analysis and Probability in Data Science

Student ID: U23AI118

July 31, 2025

Contents

1 Introduction

This report presents a comprehensive analysis of the Titanic dataset focusing on frequency analysis, probability calculations, and correlation analysis. The objective is to gain hands-on experience with statistical concepts and data visualization techniques using Python's pandas, seaborn, and matplotlib libraries.

1.1 Dataset Overview

The Titanic dataset contains demographic and survival information of passengers aboard the RMS Titanic. Key variables analyzed include:

- **class**: Passenger class (First, Second, Third)
- **sex**: Gender of passengers
- **survived**: Survival status (0 = No, 1 = Yes)
- **age**: Age of passengers
- **fare**: Ticket fare

2 Part I: Frequency Analysis

2.1 Task 1: Frequency Table of Categorical Variable

The analysis begins with creating a comprehensive frequency table for the 'class' variable.

```
import seaborn as sns
import pandas as pd
df = sns.load_dataset('titanic')

print("Frequency Table")
df.head()

print("\n1. Absolute Frequencies:")
abs_freq = df['class'].value_counts().sort_index()
print(abs_freq)

print("\n2. Relative Frequencies (%):")
rel_freq = (df['class'].value_counts(normalize=True).sort_index() *
    100).round(2)
print(rel_freq)

print("\n3. Cumulative Frequencies:")
cum_freq = df['class'].value_counts().sort_index().cumsum()
print(cum_freq)

frequency_table = pd.DataFrame({
    'Absolute Frequency': abs_freq,
    'Relative Frequency (%)': rel_freq,
    'Cumulative Frequency': cum_freq
})

print("\n4. Complete Frequency Table:")
```

```
print(frequency_table)
```

Listing 1: Data Loading and Frequency Table Creation

2.2 Results

Table 1: Frequency Table for Passenger Class

Class	Absolute Frequency	Relative Frequency (%)	Cumulative Frequency
First	216	24.24	216
Second	184	20.65	400
Third	491	55.11	891

Key Findings:

- Third class passengers represent the majority (55.11%) of the dataset
- First class passengers comprise 24.24% of the total
- Second class passengers are the smallest group at 20.65%

3 Part II: Joint, Marginal, and Conditional Probabilities

3.1 Task 2: Two-Way Contingency Table

A contingency table was constructed to analyze the relationship between passenger sex and survival status.

```
print("Contingency Table (sex vs survived):")

contingency_table = pd.crosstab(df['sex'], df['survived'], margins=True
    , margins_name='Total')
contingency_table.columns = ['Survived = 0', 'Survived = 1', 'Total']
print(contingency_table)
```

Listing 2: Contingency Table Creation

Table 2: Contingency Table: Sex vs Survived

Sex	Survived = 0	Survived = 1	Total
Female	81	233	314
Male	468	109	577
Total	549	342	891

3.2 Task 3: Probability Calculations

```
print("Probability Calculations:")

total = contingency_table.loc['Total', 'Total']

# 1. Joint Probability
joint_female_survived = contingency_table.loc['female', 'Survived = 1']
p_joint = joint_female_survived / total
print(f"P(Sex = female, Survived = 1) = {joint_female_survived}/{total}
    = {p_joint:.4f}")

# 2. Marginal Probabilities
female_total = contingency_table.loc['female', 'Total']
p_female = female_total / total
print(f"P(Sex = female) = {female_total}/{total} = {p_female:.4f}")

survived_total = contingency_table.loc['Total', 'Survived = 1']
p_survived = survived_total / total
print(f"P(Survived = 1) = {survived_total}/{total} = {p_survived:.4f}")

# 3. Conditional Probabilities
p_survived_given_female = joint_female_survived / female_total
print(f"P(Survived = 1 | Sex = female) = {joint_female_survived}/{
   female_total} = {p_survived_given_female:.4f}")

p_female_given_survived = joint_female_survived / survived_total
print(f"P(Sex = female | Survived = 1) = {joint_female_survived}/{
   survived_total} = {p_female_given_survived:.4f}")
```

Listing 3: Probability Calculations

3.3 Probability Results

Table 3: Calculated Probabilities

Probability Type	Value
Joint Probability	
P(Sex = female, Survived = 1)	0.2615
Marginal Probabilities	
P(Sex = female)	0.3524
P(Survived = 1)	0.3838
Conditional Probabilities	
P(Survived = 1 — Sex = female)	0.7420
P(Sex = female — Survived = 1)	0.6813

Interpretation:

- 26.15% of all passengers were female survivors
- 74.20% of female passengers survived

- 68.13% of survivors were female
- Women had significantly higher survival rates than the overall population

4 Part III: Correlation Analysis

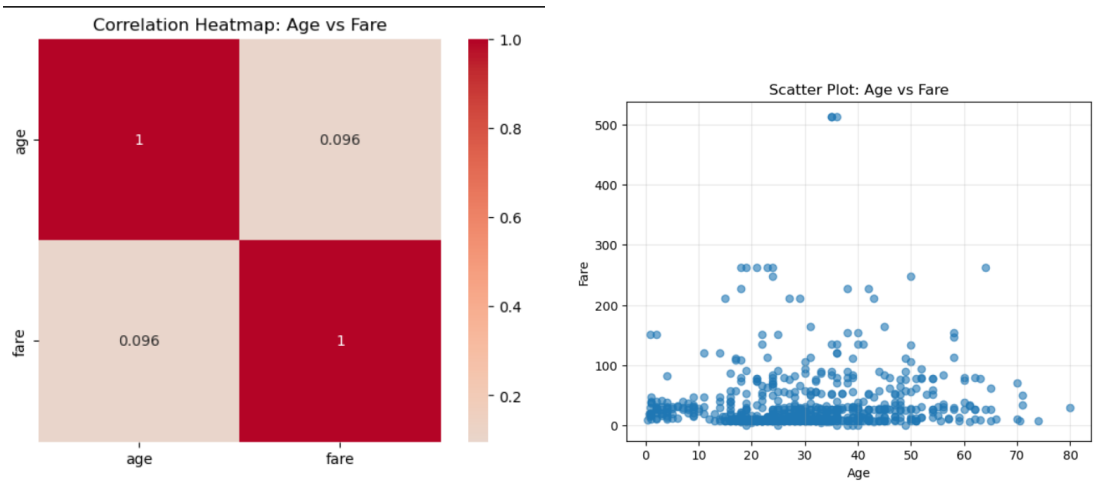
4.1 Task 4: Numerical Correlation Between Age and Fare

```
import matplotlib.pyplot as plt

print("Numerical Correlation")
print("Missing values before cleaning:")
print(f"Age: {df['age'].isnull().sum()}")
print(f"Fare: {df['fare'].isnull().sum()}")

df_clean = df[['age', 'fare']].dropna()
print(f"\nDataset size after cleaning: {len(df_clean)} rows")

correlation = df_clean['age'].corr(df_clean['fare'])
print(f"Pearson correlation between age and fare: {correlation:.4f}")

# Create visualizations
fig, axes = plt.subplots(1, 2, figsize=(15, 5))

# Correlation heatmap
correlation_matrix = df_clean[['age', 'fare']].corr()
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', center=0,
            square=True, ax=axes[0])
axes[0].set_title('Correlation Heatmap: Age vs Fare')

# Scatter plot
axes[1].scatter(df_clean['age'], df_clean['fare'], alpha=0.6)
axes[1].set_xlabel('Age')
axes[1].set_ylabel('Fare')
axes[1].set_title('Scatter Plot: Age vs Fare')
axes[1].grid(True, alpha=0.3)

plt.show()
```

Listing 4: Correlation Analysis and Data Cleaning

4.2 Data Cleaning Results

- **Missing values:** Age: 177, Fare: 0
- **Dataset size after cleaning:** 714 rows
- **Pearson correlation coefficient:** 0.0961



(a) Correlation Heatmap (b) Scatter Plot: Age vs Fare

Figure 1: Correlation Analysis Visualizations

4.3 Task 5: Correlation Interpretation

```
print("Correlation Interpretation")
print(f"Correlation coefficient: {correlation:.4f}")

if correlation > 0:
    direction = "positive"
    print(f"Direction: {direction} - As age increases, fare tends to
        increase")

abs_corr = abs(correlation)
if abs_corr >= 0.8:
    strength = "very strong"
elif abs_corr >= 0.6:
    strength = "strong"
elif abs_corr >= 0.4:
    strength = "moderate"
elif abs_corr >= 0.2:
    strength = "weak"
else:
    strength = "very weak"
print(f"Strength: {strength} correlation")

print(f"\nSign indication:")
if correlation > 0:
    print("Positive sign (+): Variables move in the same direction")
    print("- When one variable increases, the other tends to increase")
```

Listing 5: Correlation Interpretation

Correlation Analysis Results:

- **Correlation coefficient:** 0.0961
- **Direction:** Positive - As age increases, fare tends to increase slightly
- **Strength:** Very weak correlation
- **Interpretation:** There is minimal linear relationship between age and fare

5 Bonus Task: Survival Analysis by Class

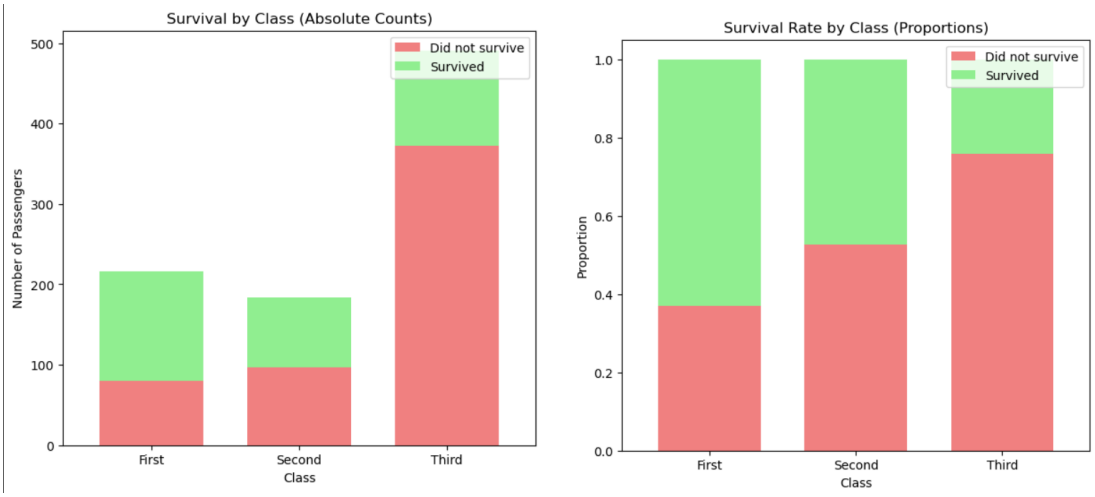
```
print("Survival Analysis by Class")

class_survival = pd.crosstab(df['class'], df['survived'], margins=True)
print("Contingency Table (Class vs Survived):")
print(class_survival)

survival_rates = pd.crosstab(df['class'], df['survived'], normalize='
    index') * 100
print(f"\nSurvival Rates by Class (%):")
print(survival_rates.round(2))

fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(15, 6))

# Stacked bar chart - absolute counts
class_survival_data = pd.crosstab(df['class'], df['survived'])
class_survival_data.plot(kind='bar', stacked=True, ax=ax1,
                         color=['lightcoral', 'lightgreen'], width=0.7)
ax1.set_title('Survival by Class (Absolute Counts)')
ax1.set_xlabel('Class')
ax1.set_ylabel('Number of Passengers')
ax1.legend(['Did not survive', 'Survived'], loc='upper right')

# Stacked bar chart - proportions
survival_rates_plot = pd.crosstab(df['class'], df['survived'],
    normalize='index')
survival_rates_plot.plot(kind='bar', stacked=True, ax=ax2,
                          color=['lightcoral', 'lightgreen'], width=0.7)
ax2.set_title('Survival Rate by Class (Proportions)')
ax2.set_xlabel('Class')
ax2.set_ylabel('Proportion')

plt.show()
```

Listing 6: Stacked Bar Chart for Survival by Class

5.1 Survival Analysis Results

Table 4: Survival Rates by Passenger Class

Class	Did not survive	Survived	Survival Rate (%)
First	80	136	63.0
Second	97	87	47.3
Third	372	119	24.2



(a) Absolute Counts (b) Survival Proportions

Figure 2: Survival Analysis by Class

Key Findings:

- **First class** had the highest survival rate at **63.0%**
- Second class had a moderate survival rate of 47.3%
- Third class had the lowest survival rate at 24.2%
- Clear inverse relationship: higher class passengers had better survival chances
- This likely reflects proximity to lifeboats and evacuation priority

6 Conclusions

This comprehensive analysis of the Titanic dataset reveals several important insights:

6.1 Statistical Findings

1. **Class Distribution:** Third class passengers dominated the ship (55.11%), followed by first class (24.24%) and second class (20.65%)
2. **Gender and Survival:** Women had significantly higher survival rates (74.20%) compared to men, demonstrating the "women and children first" protocol
3. **Age-Fare Correlation:** Very weak positive correlation (0.0961) suggests age and fare are largely independent variables
4. **Class-based Survival Disparity:** Strong socioeconomic bias in survival rates:
   - First class: 63.0% survival rate
   - Second class: 47.3% survival rate
   - Third class: 24.2% survival rate

6.2 Methodological Insights

- Successfully demonstrated frequency analysis techniques
- Applied probability theory to real-world disaster data
- Utilized correlation analysis to explore variable relationships
- Created effective data visualizations for complex relationships

6.3 Historical Context

The analysis confirms historical accounts of the Titanic disaster, showing clear evidence of:

- Social class privileges affecting survival outcomes
- Gender-based evacuation protocols
- Systematic disparities in access to life-saving resources

This project successfully demonstrates the power of statistical analysis in uncovering patterns and relationships within historical data, providing quantitative evidence for well-documented social phenomena during the Titanic disaster.