Statistical Analysis of the *Tips* Dataset

StatBridge Studio

Abstract

This report presents a comprehensive statistical analysis of the *Tips* dataset. The analysis includes descriptive statistics, data visualization, and inferential statistical tests. The primary aim is to understand the relationship between the total bill and the tip amount, and to explore differences in tipping behavior across different demographic groups.

1 Introduction

The *Tips* dataset, available in the *seaborn* library, contains information on the tips received by waiters and waitresses at a restaurant, along with various other factors such as the total bill, the day of the week, and whether the customer was a smoker. The dataset includes data from 244 transactions.

This report aims to provide a detailed statistical analysis of the dataset, including the exploration of key descriptive statistics, the visualization of data distributions, and the application of inferential statistical tests to investigate differences in tipping behavior.

2 Data Description

The dataset consists of the following variables:

- **total_bill**: The total bill (including tip) in USD.
- **tip**: The tip amount in USD.
- **sex**: The gender of the person paying the bill (Male/Female).
- **smoker**: Whether the person is a smoker (Yes/No).
- **day**: The day of the week (Thur, Fri, Sat, Sun).

- **time**: The time of day (Lunch/Dinner).
- **size**: The size of the party.

3 Methodology

The analysis was conducted using Python, with libraries such as *pandas* for data manipulation, *seaborn* for data visualization, and *scipy* for statistical tests. The steps involved in the analysis include:

1. **Data Cleaning and Preparation**: Handling missing values and outliers.
2. **Descriptive Statistics**: Calculating measures of central tendency and dispersion, such as mean, median, and standard deviation.
3. **Data Visualization**: Using histograms, box plots, and scatter plots to explore data distributions and relationships.
4. **Inferential Statistics**: Conducting t-tests to compare mean tip amounts across different groups.

The Python code for loading the dataset and calculating descriptive statistics is shown below:

```
import pandas as pd
import seaborn as sns

# Load the dataset
data = sns.load_dataset('tips')

# Descriptive statistics
descriptive_stats = data.describe()
print(descriptive_stats)

```

Listing 1: Loading the Dataset and Descriptive Statistics

4 Results

4.1 Descriptive Statistics

The descriptive statistics for the key variables are summarized in Table 1.

	Count	Mean	Std Dev
Total Bill	244	19.79	8.90
Tip	244	2.99	1.38
Size	244	2.57	0.95

Table 1: Summary Statistics of Total Bill, Tip, and Size

4.2 Data Visualization

The distribution of total bills is illustrated in Figure 1. The corresponding Python code used to create this histogram is:

```
    import matplotlib.pyplot as plt

    plt.figure(figsize=(10, 6))
    sns.histplot(data['total_bill'], kde=True, color='skyblue',
bins=20)
    plt.title('Distribution of Total Bill', fontsize=16)
    plt.xlabel('Total Bill', fontsize=12)
    plt.ylabel('Frequency', fontsize=12)
    plt.grid(True)
    plt.show()

```

Listing 2: Distribution of Total Bill

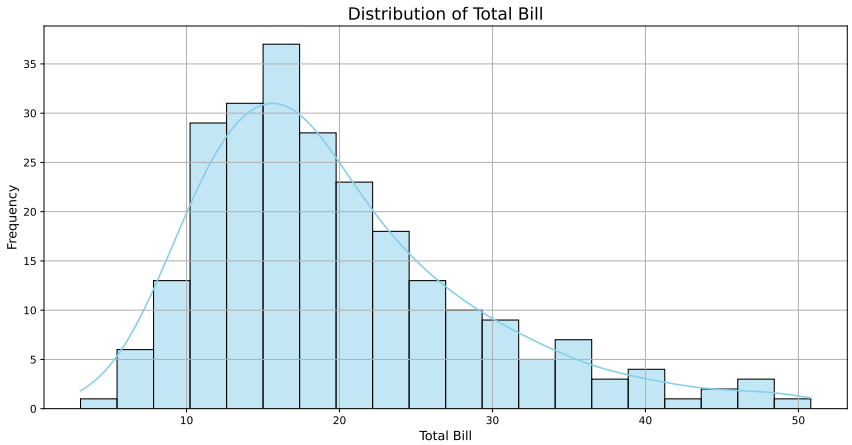

Figure 1: Distribution of Total Bill

Figure 1 shows the distribution of the total bill amounts. The distribution

is right-skewed, with most bills falling below $30.

The variation in tips across different days of the week is illustrated in Figure 2. The corresponding Python code is:

```
    plt.figure(figsize=(10, 6))
    sns.boxplot(x='day', y='tip', data=data, palette='Set2')
    plt.title('Tips by Day of the Week', fontsize=16)
    plt.xlabel('Day of the Week', fontsize=12)
    plt.ylabel('Tip', fontsize=12)
    plt.grid(True)
    plt.show()

```

Listing 3: Boxplot of Tips by Day of the Week

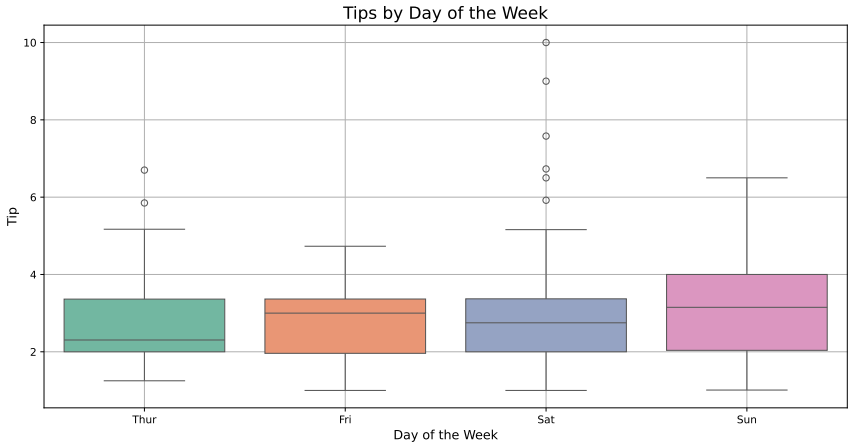

Figure 2: Tips by Day of the Week

Figure 2 reveals that tips tend to be higher on weekends (Saturday and Sunday) compared to weekdays.

The relationship between the total bill and the tip amount is shown in Figure 3, with the corresponding Python code:

```
    plt.figure(figsize=(10, 6))
    sns.regplot(x='total_bill', y='tip', data=data, scatter_kws
    ={'s':50}, line_kws={'color':'red'})
    plt.title('Relationship between Total Bill and Tip',
    fontsize=16)
    plt.xlabel('Total Bill', fontsize=12)
    plt.ylabel('Tip', fontsize=12)
    plt.grid(True)
```

```
    plt.show()

```

Listing 4: Scatter plot of Total Bill vs Tip

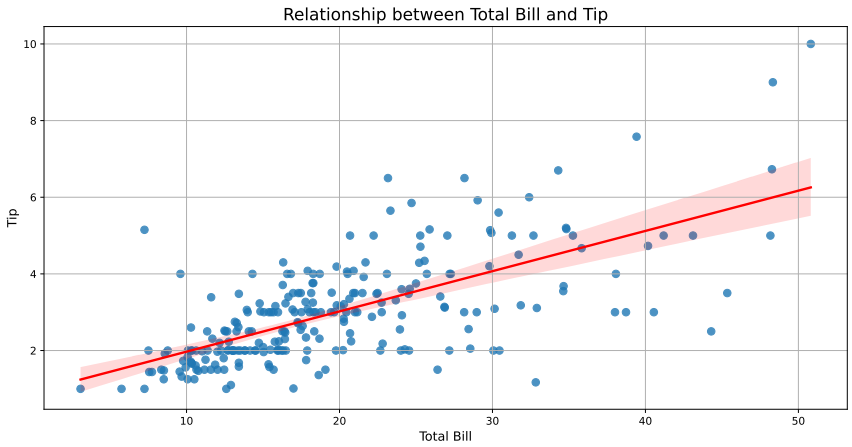

Figure 3: Relationship between Total Bill and Tip

Figure 3 demonstrates a positive correlation between the total bill and the tip amount, indicating that higher bills are generally associated with higher tips.

4.3 Inferential Statistics

A two-sample t-test was conducted to compare the mean tips given by male and female customers. The test resulted in a t-statistic of 1.39 and a p-value of 0.1665. The following Python code snippet demonstrates how the t-test was performed:

```
    from scipy import stats

    # Extracting tips based on gender
    male_tips = data[data['sex'] == 'Male']['tip']
    female_tips = data[data['sex'] == 'Female']['tip']

    # T-test
    t_stat, p_value = stats.ttest_ind(male_tips, female_tips)
    print(f'T-statistic: {t_stat:.2f}, P-value: {p_value:.4f}')

```

Listing 5: Two-sample t-test

Since the p-value is greater than 0.05, we do not reject the null hypothesis. This indicates that there is no statistically significant difference in the mean tip amounts between male and female customers.

5 Discussion

The analysis reveals several insights into tipping behavior. The average tip is significantly influenced by the total bill amount, as shown by the positive correlation observed in Figure 3. Additionally, tips are generally higher on weekends, possibly due to different dining behaviors or customer expectations.

The t-test analysis indicates no significant difference in tipping behavior between genders, which suggests that other factors might play a more critical role in determining tip amounts.

6 Conclusion

This report provided a thorough analysis of the *Tips* dataset, exploring descriptive statistics, visualizations, and inferential statistics. The findings highlight key trends and differences in tipping behavior based on various factors such as gender and day of the week.

Further analysis could delve deeper into the impact of other variables, such as the time of day or party size, on tipping behavior. Additionally, a more detailed exploration of potential outliers and their impact on the analysis could provide more nuanced insights.

7 References

- Seaborn Library: `https://seaborn.pydata.org/`
- Pandas Documentation: `https://pandas.pydata.org/`
- Scipy Documentation: `https://scipy.org/`