6 April 2024

# Statistical Analysis of Passing Rate and Game Outcomes from Bundesliga


Prepared by :
**Saijal Singhal**

# Index


| Sr No. | Content | | Page No. |
|---|---|---|---|
| 1. | Introduction | | 3 |
| 2. | Research Questions and Tasks | | 4 |
| 3. | Methods Applied | | 5 |
| | 3.1 | Shapiro-Wilk Test | 5 |
| | 3.2 | Two-Sample T-Test | 6 |
| | 3.3 | Mann-Whitney U-Test | 8 |
| | 3.4 | Cohen’s D | 10 |
| 4. | Evaluation | | 11 |
| | 4.1 | Task 1 | 11 |
| | 4.2 | Task 2 | 11 |
| | 4.3 | Task 3 | 12 |
| | 4.4 | Task 4 | 14 |
| | 4.5 | Task 5 | 16 |
| | 4.6 | Task 6 | 17 |
| 5. | Summary | | 19 |
| 6. | Bibliography | | 20 |

# Introduction

Soccer, also known as football, captivates billions of fans worldwide. With its simple rules and fast-paced action, it transcends cultural barriers to become a global phenomenon. Beyond the entertainment value, soccer is a multi-billion dollar industry, with leagues generating vast revenue through broadcasting rights. Understanding the factors influencing success in this highly competitive sport is not just a matter of fan curiosity, but holds significant strategic and financial implications.

One crucial aspect of soccer is passing, the art of accurately transferring the ball between players. Passing not only facilitates ball movement but also creates scoring opportunities and disrupts the opponent's game plan. However, the exact relationship between passing and winning outcomes remains a subject of debate. This project delves into this very question: **Does passing rate significantly influence the outcome of a soccer game?** This report analysed data on passing rates and game results to uncover potential correlations.

The analysis revealed a **statistically significant difference** in passing rates between winning and losing teams. Teams with higher passing rates emerged victorious more often, suggesting a positive link between efficient passing and securing wins. However, the effect size, a measure of the strength of this association, was small. This implies that successful soccer gameplay likely involves a complex interplay of factors beyond just passing proficiency.

While passing seems to be a contributing factor, the story doesn't end there. The investigation identified no significant difference in passing rates between winning and tie games. This hints that other factors beyond passing excellence might play a more prominent role in determining the outcome of a tie game.

This report delves deeper into these findings by following the structure –

- Research Questions and Tasks involved
- Understanding statistical methods that will be used
- Implement the various statistical tools using Python to analyse our data
- Evaluate and summarize the results from the statistical test performed
- Conclusion

## Research Questions and Tasks

This report investigates passing rates to understand its impact in soccer matches. This statistical analysis aims to address the following research questions –

**Research Question 1** – Is there a statistically significant difference in the average passing rate between winning and tie games?

**Research Question 2** – Is there a larger difference in passing rates within winning games compared to tie games?

The following tasks will focus on answering the above questions –

1. Data cleaning and preparations – This task will involve reading and cleaning the data to address missing values.
2. Data re-structuring – To restructure data such that it is not only easy to understand but also clearly states the outcome, i.e., win/lose/tie, of soccer matches played in pairs. This will in turn keep the inferencing of statistical methods straight forward.
3. Exploratory Data Analysis – To carry out basic statistics (mean, median, standard deviation) for passing rates in different groups (winners, losers, ties). Along with this, basic visualization graphs to understand the spread of data.
4. Check for normality – To assess the normality of passing rates within each group (winners, losers, ties) which is an important assumption for parametric tests in statistics.
5. Hypothesis Testing – To define the null and alternative hypothesis clearly and perform hypothesis testing, if –
    - Data is deemed normal (or if violations is not severe) then to perform parametric two-sample t-test to compare passing rates between relevant groups.
    - Data is not normal them, to perform Mann-Whitney U test as a non-parametric alternative for comparing passing rates.
6. Effect size calculation – To quantify the magnitude of the observed difference in passing rates.

# Methods Applied

Statistical analysis operates with a framework known as hypothesis testing. This framework provides a structured approach for making evidence-based decisions based on data. Hypothesis testing or significance testing gives us this structured approach to assess claims involving a group. This testing involves setting up two-key components:

1. Null hypothesis ($H_0$) – This is the statement about the data that we aim to test. It typically represents the scenario where there is "no effect" or "no difference" between groups.
2. Alternative hypothesis ($H_1$) – This statement directly contradicts the null hypothesis. It reflects what we hope to find evidence for, suggesting an "effect" or a "difference" between groups.

However, simply stating these hypotheses isn't enough. To make data-driven decisions within hypothesis testing, we rely on evidence from the data itself. We establish a significance level (α), typically set at 0.05, which represents the threshold for the probability of making a Type I error.

| | Null hypothesis (Ho) is TRUE | Null hypothesis (Ho) is FALSE |
|---|---|---|
| Reject null hypothesis (Ho) | Type I error False Positive | Correct Outcome True Positive |
| Fail to reject null hypothesis (Ho) | Correct Outcome True Negative | Type II error False Negative |

To assess the evidence against the null hypothesis, we employ statistical tests. Each test calculates a specific statistic that reflects the strength of this evidence. Additionally, these tests provide a p-value, which represents the probability of observing a test statistic as extreme or more extreme than the calculated value, assuming the null hypothesis is true.

Strong evidence contradicting the null hypothesis is indicated by a low p-value (usually less than α), which may lead to the alternative hypothesis' rejection. By analyzing these test statistics and p-values, we can make statistically sound conclusions about the data and draw inferences about the relationships or differences we are investigating.

The following sections will explore various statistical tests used in research, delving deeper into their specific calculations, workings, advantageous and disadvantages –

### 3.1 Shapiro-Wilk Test [1]

This is a quantitative test that compliments the visualization of normality. an analysis of the hypothesis that determines whether a sample is representative of a normal distribution. In other words, it evaluates whether a data set is normally distributed or not. The following steps are involved mathematically [2] [3] –

$$Let, the\ sample\ data\ as\ X_1, X_2, X_3, \ldots, X_n, where\ n\ is\ the\ sample\ size.$$

$$Shapiro - Wilk\ test\ statistic, W = \frac{(\sum_{i=1}^{n} a_i X_i)^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$X_i - i^{th}\ order\ statistic\ (ordered\ statistic\ values)$$

$$\bar{X} - Sample\ mean$$

$a_i -$ *constants obtained from the covariance matrix of the order statistics of a sample from a normal distribution*

Step 1 – To arrange the sample data in ascending order and compute Sample mean ($\bar{X}$) and Variance ($S^2$) of the sample data.

Step 2 – To calculate the coefficients ($a_i$) which depend on the sample size and are obtained from precomputed tables.

Step 3 – To calculate the test statistic ($W$) by plugging in the values of ordered sample data, mean, and coefficients into the formula.

Step 4 – Compare $W$ with Critical Value from statistical tables (based on significance level and sample size).

Step 5 – To make a decision:

1. If $W$ is less than the critical value, fail to reject the null hypothesis (data is normally distributed).
2. If $W$ is greater than the critical value, reject the null hypothesis (data is not normally distributed).

Shapiro-Wilk test is more powerful than other normality tests, especially with smaller sample sizes. The Shapiro-Wilk test has minimal assumptions compared to some normality tests. It primarily relies on the data being random and independent. It can be applied to various data types, including continuous and ordinal data.

The Shapiro-Wilk test can be overly sensitive to sample size. With very large datasets, even minor deviations from normality might lead to significant p-values. While software handles calculations, the underlying formula for W is mathematically complex.

**2. Two-Sample T-Test**

A statistical tool used to compare the means of two independent groups which determines if the observed difference in means is likely due to random chance or reflects a genuine difference between the populations from where the samples were drawn. The two-sample t-test relies on the t-distribution, a bell-shaped curve similar to the normal distribution but with slightly fatter tails. To understand how two-sample t-test works, the following method is used [4]_

Step 1 – To calculate the two-sample means followed by calculating the difference between the estimates. Here, the estimate calculated is our best guess of the true difference between means.

$$\bar{X}_1 \; and \; \bar{X}_2 - Sample \; means \; for \; group \; 1 \; and \; 2, respectively$$

$$n_1 \; and \; n_2 - Sample \; sizes \; for \; group \; 1 \; and \; 2, respectively$$

Step 2 – To estimate the standard error of difference between sample means under the null hypothesis of no difference. This will in turn give us an idea of how much sampling variation we should expect to observe in estimated difference if there were actually no difference between the means.

If both groups have equal sample sizes and variance,

$$Standard \; error = s_p \sqrt{\frac{2}{n}}$$

$$s_p \; (pooled \; standard \; deviation) = \sqrt{\frac{{s_{X_1}}^2 + {s_{X_2}}^2}{2}}$$

$$s_{X_1} and \; s_{X_2} - \; Sample \; standard \; deviations \; (unbiased \; estimators) \; for \; group \; 1 \; and \; 2, respectively$$

But if both groups have equal or unequal sample sizes and similar variances [5],

$$Standard \; error = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$s_p(pooled \; standard \; deviation) = \sqrt{\frac{(n_1 - 1){s_{X_1}}^2 + (n_2 - 1){s_{X_2}}^2}{n_1 + n_2 - 2}}$$

$$s_{X_1} and \; s_{X_2} - \; Sample \; standard \; deviations \; (unbiased \; estimators) \; for \; group \; 1 \; and \; 2, respectively$$

Here, $s_p$ is defined such that its square is an unbiased estimator of the common variance, whether or not the population means are the same.

Step 3 – To calculate the test statistic (t-statistic or t-score) which is the core of this test that represents difference between the groups mean relative to the standard error. It provides the number of standard deviations that the observed mean difference falls within –

$$t = \frac{Difference \; between \; sample \; means}{Standard \; error \; of \; the \; Difference} = \frac{\bar{X}_1 - \bar{X}_2}{Standard \; error}$$

Step 4 – To compare the test statistic to the theoretical predictions of the t-distribution for assessing the statistical significance of the observed difference. P-value is to calculate the probability that we would have observed a difference between means with a magnitude as large as, or larger than, the observed difference, if the null hypothesis were true.

Decision rule – If the p-value is less than significance value ($\alpha$), we reject the null hypothesis and conclude there's a statistically significant difference between the means of the two groups. Conversely, a p-value

greater than significance value ($\alpha$) suggests we fail to reject the null hypothesis, meaning the observed difference might be due to random chance.

Unlike some statistical tests that require large datasets, the two-sample t-test can be reasonably reliable even with moderate sample sizes, provided the data meets the assumptions. The t-test directly compares the means of two groups, which is often the primary question of interest in many research scenarios. It helps determine if an intervention, treatment, or other factor has a statistically significant effect on the average outcome within a group.

The t-test relies on specific assumptions about the data, such as normality and homogeneity of variances. Violations of these assumptions can affect the accuracy of the results. A statistically significant difference found by the t-test doesn't necessarily imply causation. Other factors might influence the observed difference. The t-test is a starting point for further investigation, not a definitive answer to cause-and-effect relationships.

**3. Mann-Whitney U-Test** [6]

The Mann-Whitney U-test is a non-parametric test used to compare the medians of two independent groups. This test is used to determine whether the total ranks in two independent groups are significantly different. The null hypothesis ($H_0$) for the Mann-Whitney U test is that the ranks in two groups are equally dispersed. The alternative hypothesis is that the ranks in two groups are not equally dispersed. Breakdown of the math involved is –

Step 1 – For each group arrange all data points from both groups together in ascending order.

Step 2 – Assigning a rank (from 1 to the total number of observations) to each data point based on its position in the combined order. If there are ties (multiple data points with the same value), assign them the average rank for that position.

Step 3 – Assign points when a score in one group outranks scores in another group and calculate the sum (denoted by T) for each group.

$$n_1 \text{ and } n_2 - \text{Sample sizes for group } 1 \text{ and } 2, \text{respectively}$$

$$\text{Rank } (r_i) \text{ to each data point } (z_i)$$

$$T_1 = \sum r_i \text{ for all } z_i \text{ belonging to group } 1 \ (X)$$

$$T_2 = \sum r_i \text{ for all } z_i \text{ belonging to group } 2 \ (Y)$$

Step 4 – Under $H_0$, we expect ranks to be evenly distributed between the two groups. The expected rank sum for each group under $H_0$ is:

$$E(T_1) = n_1 * \frac{(n_1 + n_2 + 1)}{2}$$

$$E(T_2) = n_2 * \frac{(n_1 + n_2 + 1)}{2}$$

Step 5 - For large samples ($n_1$ and $n_2$ both greater than 20), the sampling distribution of the U statistic can be approximated by a normal distribution. The U statistic itself can be calculated in two ways but typically the smaller value of U is used:

$$U_1 = n_1 n_2 - T_1$$

$$U_2 = n_1 n_2 - T_2$$

Step 6 – To calculate the mean under $H_0$ –

$$\mu = \frac{n_1 n_2}{2}$$

Step 7 – To calculate the standard deviation under $H_0$ –

$$\sigma = \sqrt{n_1 n_2 \frac{(n_1 + n_2 + 1)}{12}}$$

Step 8 – Finally, the test statistic (z-score) is calculated as –

$$z = \frac{U - \mu}{\sigma}$$

Step 9 – P-value represents the probability of observing a U statistic as extreme or even more extreme than the one calculated, assuming $H_0$ is true.

Decision Rule – If the p-value is less than alpha, reject $H_0$ and conclude a statistically significant difference between the medians of the two groups. If the p-value is greater than alpha, fail to reject $H_0$, suggesting the medians might be similar.

The Mann-Whitney U test is a non-parametric test, meaning it doesn't require the data to follow a specific normal distribution. This makes it a robust choice for analyzing data that might be skewed or have outliers, unlike parametric tests like the two-sample t-test which assume normality. The test directly compares the medians of two independent groups, which can be a more informative measure of central tendency than the mean, especially for skewed data. Compared to parametric tests, the Mann-Whitney U test has fewer assumptions. It primarily requires independent samples and ordinal data (data with a rankable order),

Since the Mann-Whitney U-test relies on ranks, it discards some information contained in the actual data values. It might be less powerful than a parametric test if your data truly follows a normal distribution. This test can only compare the medians of two independent groups. If you have more than two groups to compare, you'll need to use a different non-parametric test like the Kruskal-Wallis's test. Mann-Whitney U-test only tells whether there's a statistically significant difference between the medians. It doesn't quantify the magnitude of that difference.

**4. Cohen's D**

When comparing the results from two independent groups, understanding the magnitude of the observed difference is crucial. While simply looking at the raw difference in means can be informative, Cohen's D provides a more interpretable measure of effect size. Expressed in standard deviation units, Cohen's d allows us to assess the practical significance of the observed difference between the means of the two groups we're comparing.

It is popular metric for quantifying effect size when comparing the means of two independent groups following these steps –

Step 1 – To calculate mean values of group 1 and 2 respectively along with their sample standard deviations.

$$n_1 \, and \, n_2 - Sample \, sizes \, for \, group \, 1 \, and \, 2, respectively$$

$$M_1, M_2 - Mean \, values \, of \, group \, 1 \, and \, 2, respectively$$

$$s_1, s_2 - Sample \, standard \, deviations \, of \, group \, 1 \, and \, 2, respectively$$

Step 2 – Since the variances (${S_1}^2 and \, {S_2}^2$) might differ slightly, we estimate a common standard deviation ($s_p$) for both groups, considering their sample sizes. This $s_p$ is called the pooled standard deviation –

$$s_p(pooled \, standard \, deviation) = \sqrt{\frac{(n_1 - 1){s_1}^2 + (n_2 - 1){s_2}^2}{n_1 + n_2 - 2}}$$

Step 3 – To calculate the core statistic that represents the difference between the group means standardized by the pooled standard deviation. It tells us how many standard deviations the observed difference in means falls under [7] –

$$d = \frac{(M_1 - M_2)}{s_p}$$

Interpretation – Cohen's general guidelines for interpreting the magnitude of Cohen's d:

$$0.2 - Small \, effect \, size$$

$$0.5 - Medium \, effect \, size$$

$$0.8 - Large \, effect \, size$$

Above is the sample version of Cohen's D [8], using the sample mean ($M$) and pooled standard deviation ($s_p$), is more commonly employed for real-world research as it leverages the data from samples to estimate the effect size. The population version of Cohen's D [9], using the population mean ($\mu$) and population standard deviation ($\sigma$), is rarely used in practice because it requires knowing the population mean and standard deviation, which are often unavailable. Essentially, the concept of Cohen's d remains the same (standardizing the mean difference), but the formula is adjusted based on whether we have access to population parameters or rely on sample estimates. Statistical software typically calculates Cohen's d based on the sample statistics.

## Evaluation *[Code can be found here.]*

**4.1 Task 1** – Data cleaning and preparations

- Aim – To prepare the data for analysis by addressing any missing values present.
- Method –
    - Missing values can introduce biases and inaccuracies in statistical analysis. This can be handled by employing the panda's library for data cleaning tasks.
    - By using `isnull().sum()` method missing values were assessed in the dataset.
- Result –
    - The initial dataset contained 2 missing values in the $passing\_quote$ column and 2 missing in the $winner$ column. Therefore, they were removed using the `dropna()` function in panda's library.
    - `reset_index(drop=True)` was also used to reset and reassign consecutive indices after removing the missing rows.
    - Now, the final dataset contains a complete set of observations for all variables.

**4.2 Task 2** – Data re-structuring

- Aim – This task aimed to restructure the data to facilitate analysis of passing rates in relation to game outcomes (win, lose, tie).
- Method – The original data structure didn't explicitly represent game outcomes. Below are reasons why restructuring was necessary –
    - **Handles Ties**: The processed data separates winners and losers clearly, and ties are identified as a separate category. This avoids complications in interpreting passing rate differences between winners and losers.
    - **Independent Samples**: Each row in the processed data represents a single game (one winner vs. one loser) or a single tie.

  By employing panda's library for data manipulation, the following steps were performed –
    - **Winner Binary Encoding:** The winner column was transformed into a binary variable (1 for 'Yes' and 0 for others) using a lambda function.
    - **Function for Game Data Merging:** A custom function `merge_game_data()` was created to handle each unique $game\_id$. This function takes a $game\_id$ as input and retrieves the corresponding rows (representing the two teams) from the original data. It extracts the passing rates for both teams and assigns the winner based on the original $winner$ column (either Team 1, Team 2, or Tie).
    - **Iterative Processing and Feature Creation:** Lists were created to store passing rates for Team 1, Team 2, and winner labels for each game. A loop iterated through unique $game_id$s, calling

the `merge_game_data()` function for each ID to obtain the passing rates and winner of that game. These values were then appended to the respective lists.

- **Data frame Creation:** Separate DataFrames were created for Team 1 passing rates, Team 2 passing rates, and winners.
- **Data Concatenation and Saving:** The features DataFrame (containing Team 1 and Team 2 passing rates) was concatenated with the target DataFrame (containing winner labels) along the axis 1. This combined DataFrame was then saved as a new CSV file ("processed_data.csv") containing the restructured data.

- Result – This new structure allows for direct analysis of passing rates in relation to game outcomes. The original data structure was transformed into a new format with the following features:
  - Team 1 Passing Rate: Represents the passing rate of Team 1 in each game/tie.
  - Team 2 Passing Rate: Represents the passing rate of Team 2 in each game/tie.
  - Winner: Categorical variable indicating the winner ('Team 1', 'Team 2', or 'Tie') for each game/tie.

**4.3 <u>Task 3</u>** – Exploratory Data Analysis

- Aim – This task aimed to explore the distribution and characteristics of passing rates within different winner groups (winners, losers, ties) using descriptive statistics and visualizations.
- Method – Exploratory Data Analysis (EDA) helps us understand the central tendency (mean, median), spread (standard deviation), and general distribution of the data within each group. This knowledge lays the groundwork for further statistical tests and interpretations. The following steps were performed using Pandas and Matplotlib –
  - **Descriptive Statistics**: The `describe(include='all')` function was used on the processed data to obtain summary statistics (mean, median, standard deviation, minimum, maximum) for all variables (team1_rate, team2_rate, winner)
  - Average Passing Rates by Winner Category: Separate calculations were performed to analyze passing rates within winning and tie games:

    <u>Winners</u>: Identified rows where the winner wasn't a tie and created a new DataFrame containing these "winners" data. Then, iterated through each row to determine the winning team's passing rate (either Team 1 or Team 2) and added it to a list. Finally, the average passing rate for winning teams was calculated.

    <u>Ties</u>: Similarly, rows where the winner was a tie a separate DataFrame containing these "tie games" data was created. Then combined the Team 1 and Team 2 passing rates for all tie games into a single series and calculated the average passing rate for ties.

- **Visualizations**: Boxplots were created to visualize the distribution of passing rates for Team 1 and Team 2 within each winner category. Boxplots provide insights into the center (median), spread (interquartile range), and potential outliers in the data.

- Result –
  - The output of `processed_data.describe(include='all')` summarizes the key statistics for all variables as shown below –

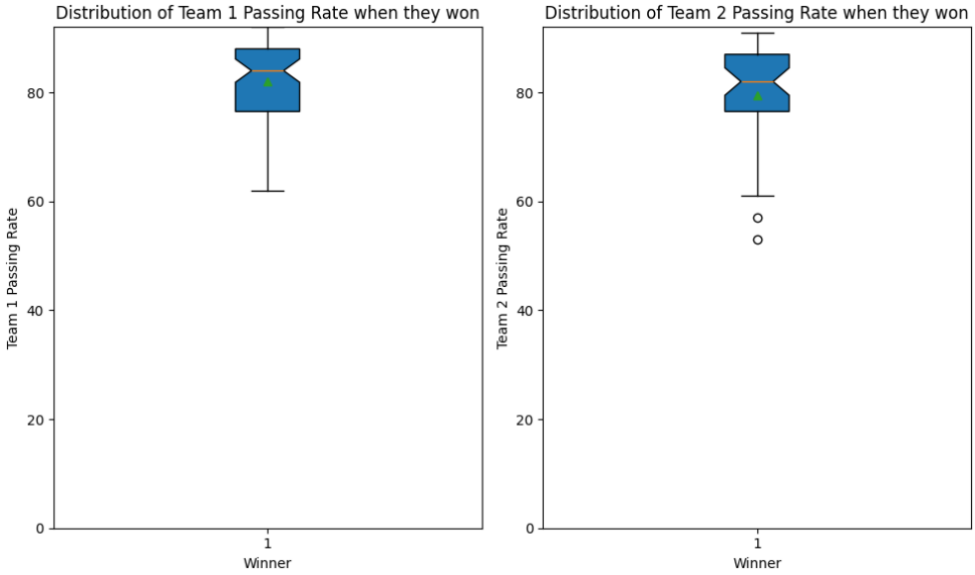
| Variable | team1_rate | team2_rate | winner |
|---|---|---|---|
| Count | 152.000000 | 152.000000 | 152 |
| Unique | NaN | NaN | 3 |
| Top | NaN | NaN | Team 1 |
| Freq | NaN | NaN | 71 |
| Mean | 80.585526 | 78.776316 | NaN |
| Standard Deviation | 6.800316 | 7.022384 | NaN |
| Min | 59.000000 | 53.000000 | NaN |
| 25% | 76.000000 | 75.000000 | NaN |
| 50% | 82.000000 | 79.000000 | NaN |
| 75% | 86.000000 | 84.000000 | NaN |
| Max | 92.000000 | 91.000000 | NaN |

  - Average Passing Rates –

    Winners – The average passing rate for winning teams was found to be 81.08

    Ties – The average passing rate for games ending in a tie was found to be 78.2

  - Boxplot –




    Team 1 Passing Rates –

    Winners – The distribution appears to be slightly skewed to the right, with a longer tail extending towards higher passing rates.

Ties – The distribution seems more symmetrical (less skewed) compared to winners, with the centre (median) closer to the middle of the box.

Team 2 Passing Rates –

Winners – The distribution for winners might be similar to Team 1 winners, with a possible positive skew and a longer tail towards higher rates.

Ties – Similar to Team 1 in ties, the distribution for Team 2 in ties might be more symmetrical with data spread across a wider range.

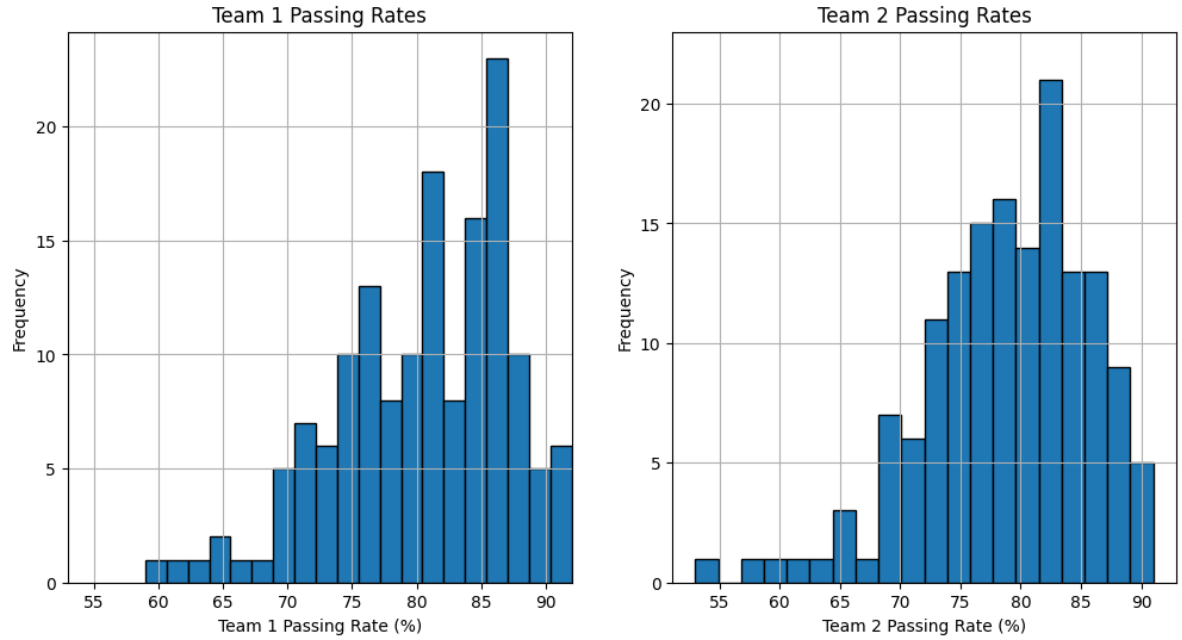
Spread (IQR) – The boxplots for winners might have a smaller interquartile range (IQR) compared to ties, indicating that the middle 50% of data points are more concentrated for winning teams.

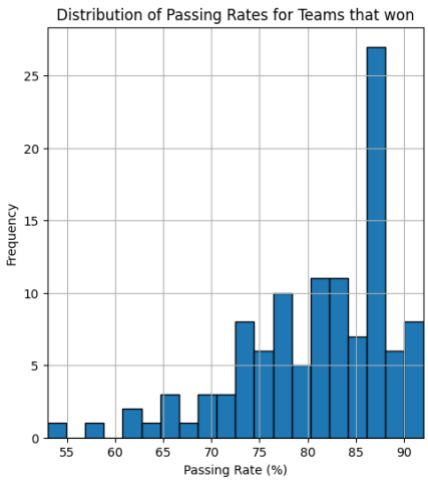
**4.4 Task 4** – Check for normality

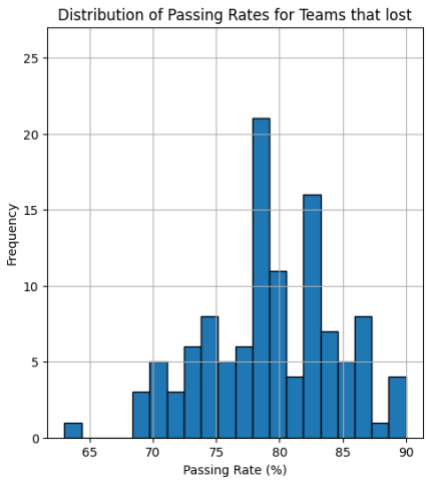
- Aim – This section examines whether the passing rates in data follow a normal distribution.
- Method –
  - Visualization – Histograms were created to visualize the distribution of passing rates. These plots allow us to assess visually if the data appears symmetrical (bell-shaped), skewed, or has outliers. Additionally, QQ plots were also created
  - Shapiro-Wilk Test – The Shapiro-Wilk test, a statistical test for normality, was performed.

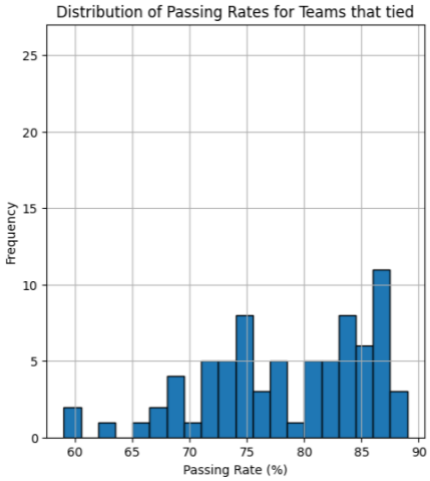
Both visualization and Shapiro wilk test was performed on the following groups -

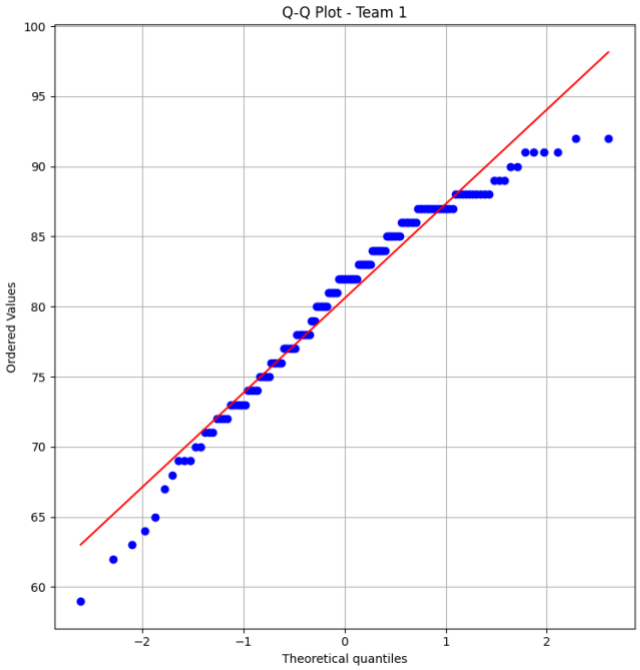
  - Team 1 passing rates
  - Team 2 passing rates
  - Winners' passing rates (combined Team 1 and Team 2)
  - Losers' passing rates (combined Team 1 and Team 2)
  - Ties' passing rates (combined Team 1 and Team 2)
- Result –

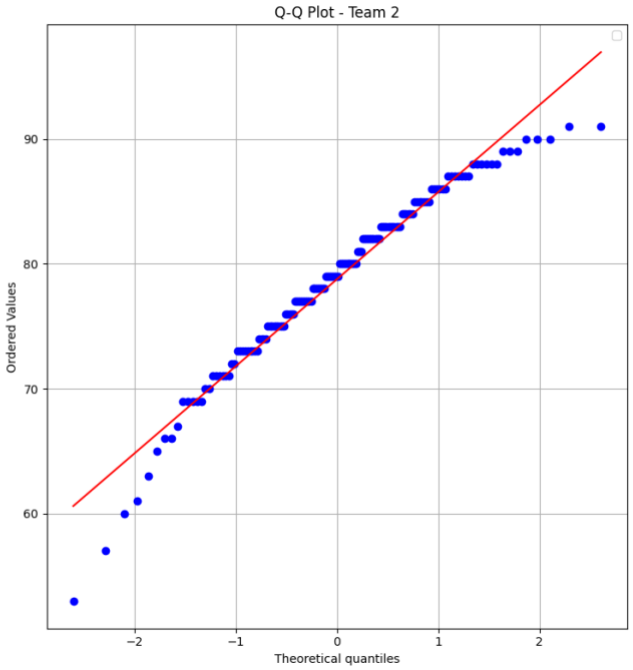


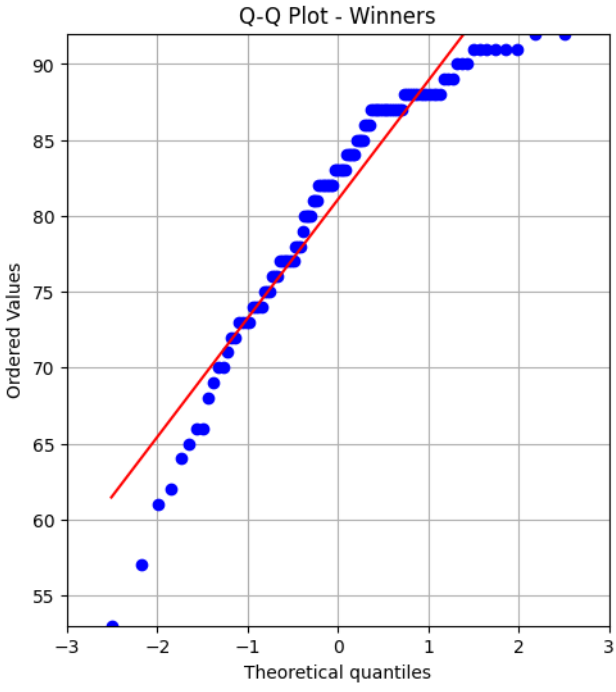

Distribution of Passing Rates for Teams that won
Frequency
Passing Rate (%)

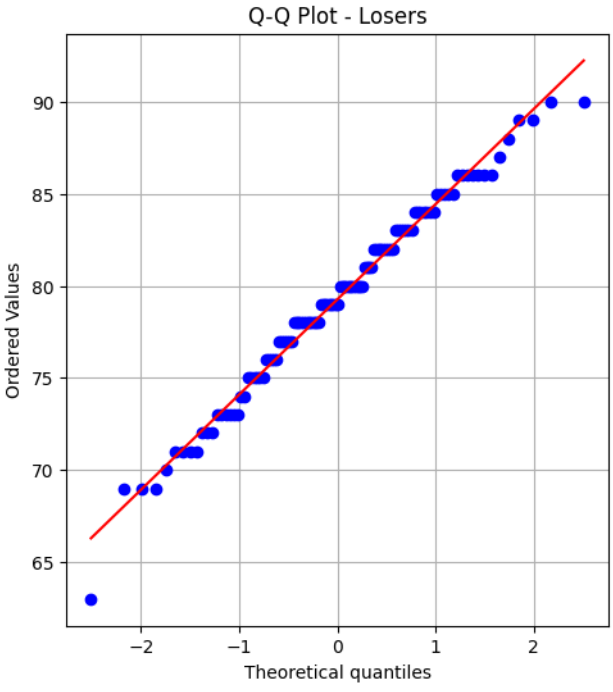

Distribution of Passing Rates for Teams that lost
Frequency
Passing Rate (%)


Distribution of Passing Rates for Teams that tied
Frequency
Passing Rate (%)


Q-Q Plot - Team 1
Ordered Values
Theoretical quantiles


Q-Q Plot - Team 2
Ordered Values
Theoretical quantiles

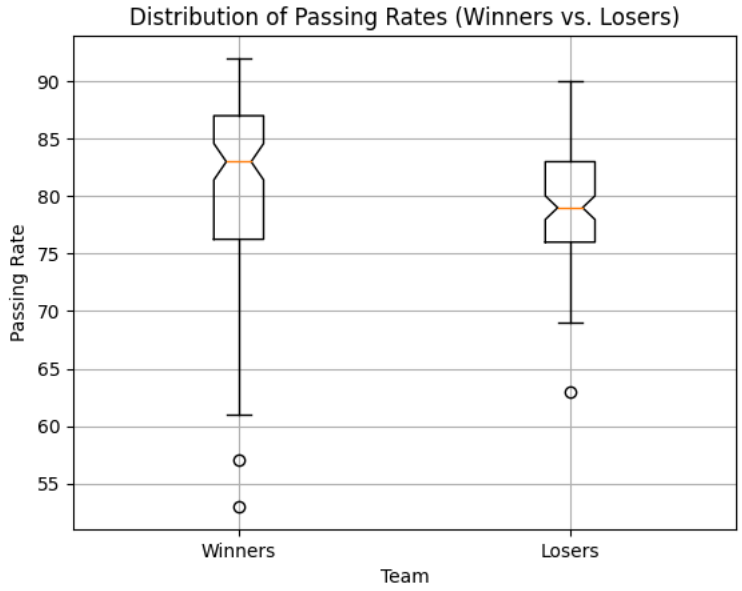

Q-Q Plot - Winners
Ordered Values
Theoretical quantiles


Q-Q Plot - Losers
Ordered Values
Theoretical quantiles

| Group | Shapiro-Wilk t-statistic | p-value |
|---|---|---|
| Team 1 passing rate | 0.958 | 0.000148 |
| Team 2 passing rate | 0.964 | 0.000515 |
| Winners' passing rates | 0.914 | 0.000002 |
| Losers' passing rates | 0.986 | 0.310275 |
| Ties' passing rates | 0.945 | 0.002580 |

**4.5 <u>Task 5</u> – Research Question 1 – Is there a statistically significant difference in the average passing rate between winning and losing games?**

- Method –
  - **Null Hypothesis ($H_0$)** – There is no statistically significant difference in the average passing rates between winning and losing teams.
  - **Alternate Hypothesis ($H_1$)** – There is a difference in passing rates between winners and losers.
  - **Two-Sample T-Test (Parametric)** – A two-sample t-test was initially conducted to compare the average passing rates of winning and losing teams. However, it's important to acknowledge that the Shapiro-Wilk test indicated potential non-normality in the data.
  - **Mann-Whitney U Test (Non-parametric)** – Due to the potential non-normality, a non-parametric Mann-Whitney U test was performed as a confirmatory test.
  - **Effect Size** – Cohen's d was calculated to assess the magnitude of the effect size.
  - **Median and boxplot** – Using median and boxplot of these groups, need to check if winners have a higher passing rate or not.
- Result –
  - <u>Two Sample T-test</u> –
    - Test Statistic (t) – 2.0279
    - p-value – 0.0437
  - <u>Mann-Whitney U Test</u> –
    - U Statistic – 8052.0000
    - p-value – 0.0018
  - <u>Cohen's D</u> – 0.26736271
  - <u>Median</u> –
    - Winners – 83.0
    - Losers – 79.0
  - <u>Box-plot</u> –

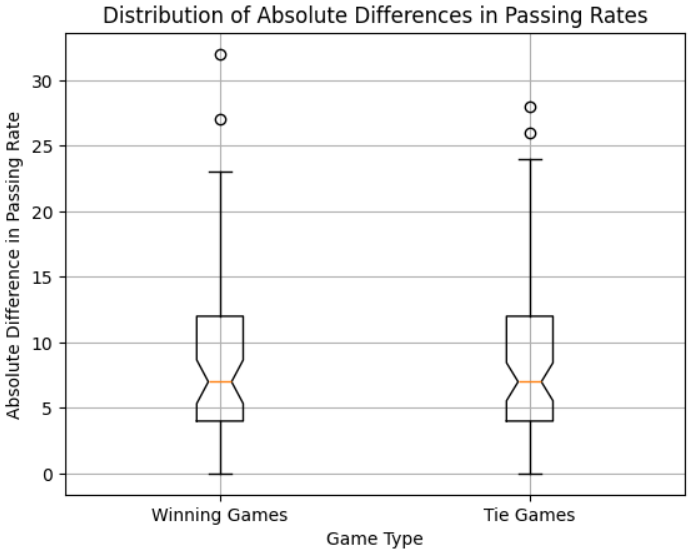



**4.6 <u>Task 6</u> – Research Question 2 – Is there a larger difference in passing rates within winning games compared to tie games?**

- Method –
  - **Null Hypothesis ($H_0$)** – There is no statistically significant difference in the average passing rate between winning and tie games.
  - **Alternate Hypothesis ($H_1$)** – There is a statistically significant difference in the average passing rate between winning and tie games.
  - **Two-Sample T-Test (Parametric)** – A two-sample t-test was conducted to compare the mean passing rates, but again, we acknowledge potential non-normality.
  - **Mann-Whitney U-Test (Non-parametric)** – The Mann-Whitney U test was performed for confirmation.
  - **Effect Size** – Cohen's d was calculated to assess the magnitude of the effect size.
  - **Median and boxplot** – To check and support the interpretation of spread.
- Result –
  - <u>Two Sample T-test</u> –
    - Test Statistic (t) – 0.4464
    - p-value – 0.6560
  - <u>Mann-Whitney U Test</u> –
    - U Statistic – 2191.5
    - p-value – 0.8055
  - <u>Cohen's D</u> – 0.0789885704911432
  - <u>Median</u> –
    - Winners – 7

- Losers – 7

- Box-plot –



# Summary

This study examined the impact of passing rate on game outcomes from 1st soccer division in Germany.

We investigated two key questions:

1. **Do winning teams have a statistically different passing rate compared to losing teams?**
   Key Result – Yes, there's a significant difference (p-value < 0.05) in passing rates. Winning teams boast a higher median passing rate, suggesting a connection between successful passing and winning. However, the effect size was small, indicating other factors likely influence winning besides passing.

2. **Is the difference in passing rate larger within winning games compared to ties?**
   Key Result – No significant difference (p > 0.05) was found in passing rates between winning and tie games. This suggests passing rate may not be a major factor distinguishing wins from ties. In other words, teams with both high and lower passing rates can emerge victorious in games that end in a tie. This finding highlights the complex nature of game outcomes, where various factors beyond passing success can contribute to a tied result.

Overall, the findings suggest that passing rate is a factor influencing game outcomes, with higher passing rates associated with winning teams. However, the difference in passing rates between winning and tie games seems negligible.

Future scope of this report could include further research exploring other factors potentially influencing game outcomes alongside passing rate.

# Bibliography

1 – An Analysis of Variance Test for Normality (pg. 591-611), Biometrika (1965)- S. S. Shapiro; M. B. Wilk Dec.

2 – Shapiro-Wilk test – Wikipedia

3 – Shapiro-Wilk test – Statistics Kingdom

4 – Chapter 13 – "Two-sample T-test" - APS 240: Data Analysis and Statistics with R by Dylan Z. Childs, Bethan J. Hindle and Philip H. Warren (14-1-2021)

5 – Independent two-sample t-test – Wikipedia

6 – Chapter 18.6 – "The Mann-Whitney U Test" (pg. 612), Statistics for the Behavioral Sciences by Gregory J. Privitera

7 - Effect size - Wikipedia

8 – Statistical Power Analysis for the Behavioral Sciences by Jacob Cohen

9 – Chapter 8.7 – "Measuring the Size of an Effect: Cohen's d" (pg. 248), Statistics for the Behavioral Sciences by Gregory J. Privitera