technische universität
dortmund

Passing Rate Analysis: Unraveling the Impact on Bundesliga Match Outcomes

Department of Statistics

TU Dortmund University

Amirreza Alasti

Date: 03/02/2024

Table of Contents

1. Introduction

Soccer, as one of the most popular sports globally, attracts significant attention from fans, analysts, and coaches alike. Understanding the factors that contribute to a team's success is of paramount importance for both strategic planning and performance evaluation. In this report, we aim to investigate whether a good passing rate has a positive influence on the chance of winning a soccer game in the 1st soccer division in Germany (1. Bundesliga). The passing rate, defined as the ratio between passes played by a team and passes received by players of the same team, is analyzed in relation to game outcomes. Specifically, we seek to answer two questions:

1. Does the winner of a match have a higher passing rate than the loser?
2. Is the difference in the passing rate in games with a winner higher than the difference in games that ended in a draw?

2. Detailed Description of the Problem

The dataset comprises information for each game of the first half of the season, identified by a unique game ID. For each game, the data includes the passing rate (in percent) for both teams and whether the respective team won the game or not. If a game ended in a draw, neither team won it, so both teams' winning status will be “No”. In this report, our task is to analyze this data ("passes.csv") to determine the relationship between passing rates and game outcomes. A sample of the dataset is presented in Table 1, and a histogram of our data is shown in Figure 1.

game_id	passing_quote	winner
11	72.0	No
11	91.0	Yes

Table 1

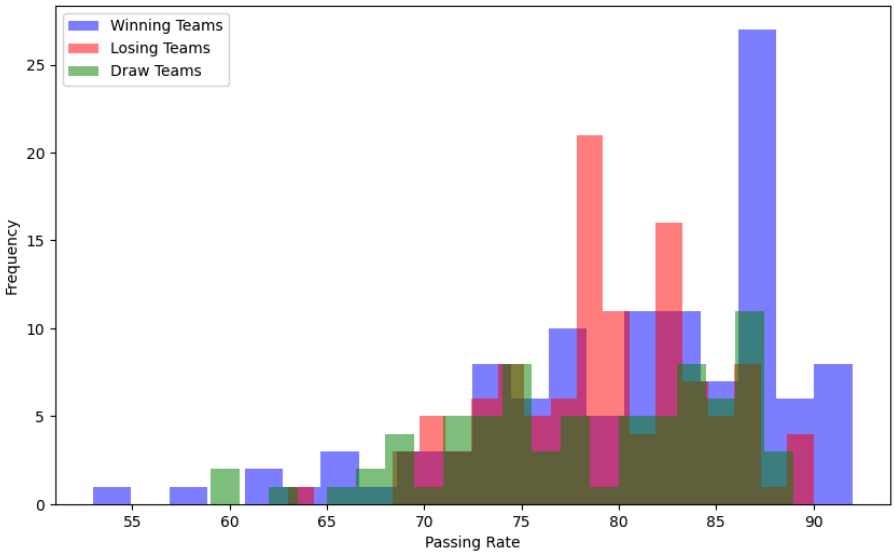

Figure 1.Distribution of Passing Rates

3. Methods

3.1. Hypothesis testing

Hypothesis testing: one of the reasons we will use this method is because we want to know whether there is a high difference in passing rates between winning and losing teams, as well as between drawn games and games with a winner and for this problem this method allows us to make data-driven conclusions about the connection between variables, providing insights into the significance of observed differences in passing rates and game outcomes. Generally this method is to decide whether the data sufficiently support a particular hypothesis.

First of all, we need to determine whether our data is normally distributed or not, and for that, we can use different methods. In this report, we are going to use two methods: the first one is a Box plot, a visual method, and the second one is the Shapiro-Wilk Test. We will use these methods to determine whether to use the Mann-Whitney U test as a non-parametric alternative to the t-test or not.

3.1.1. Shapiro-Wilk Test

The Shapiro-Wilk Test measures how well our data fits a normal distribution. This method is usually useful when we are dealing with relatively small sample sizes, like our dataset, which only has 306 rows.

Mathematically, the Shapiro-Wilk test statistic **W** is calculated as follows:
Let X_1,X_2,...,X_n be there ordered sample data.

1. Calculate the sample Mean ($\overline{X}$) and sample variance (s^2) of the data.
2. Calculate the coefficients (a_i) for the Shapiro-Wilk test. These coefficients depend on the sample size (n) and are pre-computed and tabulated.
3. Calculate the test statistic **W** using the formula (Shapiro, S.S. and Wilk, M.B, 1965):

$$\mathbf{W} = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n} (X_i - X)^2}$$

   where X_i represents the ith order statistic (i.e., the ith smallest value) from the ordered sample data.

4. If **W** is greater than 0.05 (the common significance level of alpha), you fail to reject the null hypothesis that the data is normally distributed

After determining whether our data is normally distributed or not, we are going to use our main hypothesis testing methods, which are the Two-Sample T-Test based on the result of the last phase to answer our first question, and the F-test to answer our second question.

3.1.2. Two-Sample T-Test

The two-sample t-test is a statistical test for comparing two groups' means. In pain research, it is one of the most commonly utilized statistical hypothesis tests (Yim et al, 2010)

$$t = \frac{\bar{x} - \bar{y}}{s_p \sqrt{\frac{1}{n1} + \frac{1}{n2}}}$$

First formula is for computing the T-statistic which the parameters definitions are as mentioned below:

- X and Y: two sets of data
- n1 and n2: sample sizes
- $\bar{x}$ and $\bar{y}$: sample means
- s_p: pooled standard deviation which its formula is: (s_x and s_y are sample standard deviations)

$$s_p = \sqrt{\frac{(n1-1)s_x^2 + (n2-1)s_y^2}{n1+n2-2}}$$

1. Now Compute the degrees of freedom (C.-H. Chang and N. Pal, 2008)::

$$df = n1 + n2 - 2$$

2. Determine Critical Value: Look up the critical value $t_{\alpha/2,df}$ in the t-distribution table for the desired significance level α and degrees of freedom df. Level of alpha represents the significance level chosen for hypothesis testing. It denotes the probability of committing a Type 1 error, i.e.m rejecting a true null hypothesis. It usually has small values[0.01,0.05],

which in this case an alpha level of 0.05(most common) is selected; this means that there is a 5% chance of incorrectly rejecting the null hypothesis when it is actually true.

3. Decision Rule:
    - If $|t| > t_{\alpha/2,df}$, reject the null hypothesis H0 in favor of the alternative hypothesis H1.
    - If $|t| \leq t_{\alpha/2,df}$, fail to reject the null hypothesis.

3.1.3. Mann-Whitney U test

The Mann-Whitney U test, also known as the Mann-Whitney-Wilcoxon test, is a non-parametric statistical test used to determine whether there is a significant difference between the distributions of two independent samples. It is often used when the assumptions of the parametric t-test are not met, such as when the data is not normally distributed or when the sample sizes are small(Nadim Nachar, 2008).

Here's how the Mann-Whitney U test works mathematically:
Let's assume that there are two independent samples, X and Y, with sample sizes n1 and n2, respectively.

1. First we are going to combine the data from both samples and rank all the values from smallest to largest, ignoring ties. Assign ranks R_i to each observation, where i = 1,2,…,N, and N = n1 + n2. The formula to assign ranks (R_i) to each observation:

$$R_i = \frac{1}{T}\sum_{j=1}^{T} r_j$$

Where:

- T is the number of tied observations
- r_j is the rank of the tied observation j

2. Calculate the sum of ranks for each sample U_1 and U_2

3. Determine the test statistic, U, as the smaller of U_1 and U_2:

$$U = min(U_1, U_2)$$

4. Compare the computed U statistic to a critical value from the Mann-Whitney U distribution to deter,ome statistical significance.
5. If U is greater than the critical value, reject the null hypothesis, suggesting a significant difference between the distributions of the two samples. Otherwise, fail to reject the null hypothesis

3.1.4. F-test

The F-test is a statistical test used to compare the variances of two populations or more. It is often employed to determine whether the variances of two samples are equal or if they differ significantly.(Sara Tomek and Shumacker Randall, 2013)

Mathematically, the F-test compares the ratio of variances between two samples. Specifically, it compares the ratio of the variances of the two samples to a theoretical F-distribution. The null hypothesis for the F-test is typically that the variances of the two populations are equal.

$$F = \frac{s_1^2}{s_2^2}$$

Where:

- s_1^2 is the variance of sample 1
- s_2^2 is the variance of sample 2

Once the F-value is computed, it is compared to a critical value from the F-distribution to determine statistical significance. If the computed F-value exceeds the critical value, it suggests that there is a significant difference in variances between the two populations.

4. Evaluation

Now, based on the methods described in section 3, we are going to answer the questions mentioned in the report.

4.1.1. Does the winner of a match have a higher passing rate than the loser?

First, we need to do some preprocessing on our data and add a status column to our dataset, which has three states: "win," "loss," and "draw." Then, based on the status of each team, we are going to separate the games that resulted in a draw from others for further analysis. The results are shown in Table 2.

game_id	passing_quote	winner	status
11	72.0	No	loss
11	91.0	Yes	win
15	85.0	No	draw
15	77.0	No	draw

Table 2

Now, for better understanding of our separate data, we are going to visualize it with box plots in Figure 4 to have a better understanding of the distribution of the Passing Rate between winning teams and losing teams.

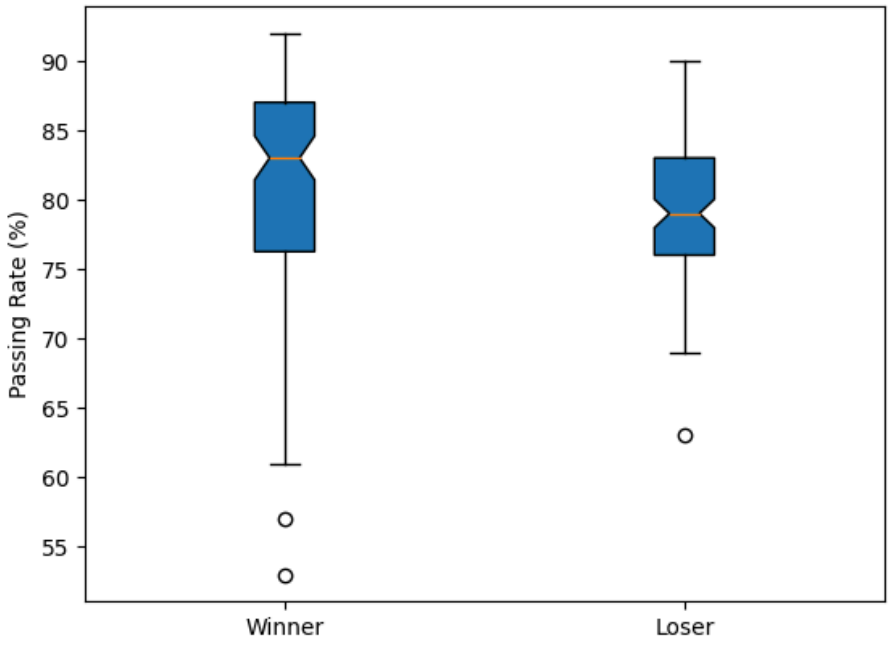

Figure 2.Passing Rate by Winner

After this, to analyze the location (central tendency) and dispersion (spread) of the data, we can compute summary statistics such as mean, median, standard deviation, and interquartile range (IQR). In Table 3, we are computing these statistics for the three groups (win, draw, loss).

Status/Statistic	Mean	Median	Standard Deviation	Interquartile Range (IQR)
Win	81.07	83.0	8.06	10.75
Draw	78.21	79.5	7.25	12.0
Loss	79.26	79.0	5.13	7.0

Table 3

First, by comparing the means and medians of passing rates between winner teams and loser teams, we can see that the passing rates of winner teams are slightly higher than those of

loser teams. However, when comparing their standard deviations, we can conclude that winner teams exhibit higher variability in passing rates compared to loser teams. The interquartile range (IQR) also indicates that winner teams have a wider spread of passing rates compared to loser teams. So, based on the statistical data, can we conclude that the winner of a match has a higher passing rate? To answer this, it is a little premature; first, we need to check if our data follows a normal distribution.

We will use a Q-Q plot (quantile-quantile plot) to compare the probability distributions of winning teams and losing teams together in Figure 3.

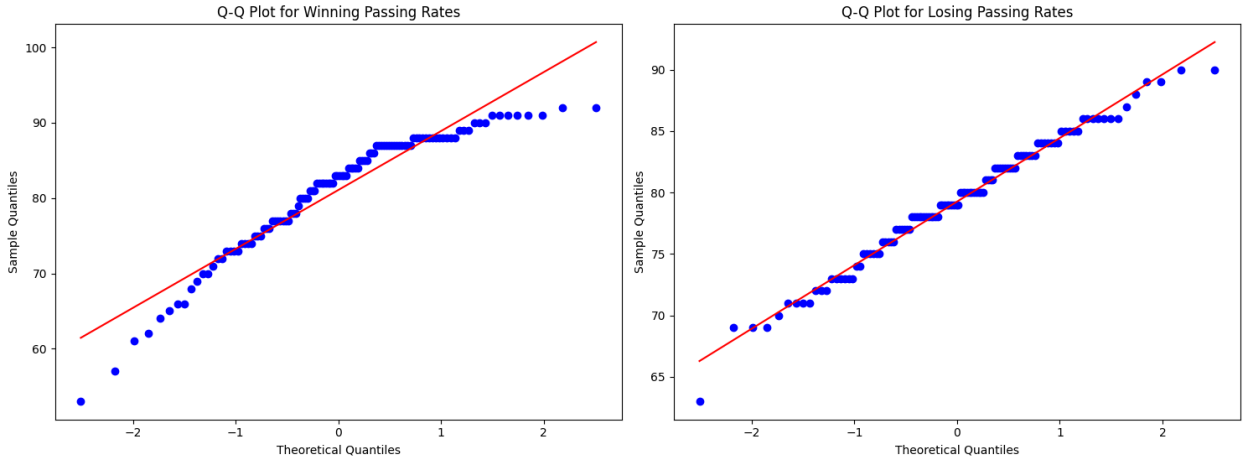

Figure 3.Q-Q Plots for Winning and Losing Passing Rates

As we can see, our data is not normally distributed. Upon employing the Shapiro-Wilk Test method, the p-value for winning teams is approximately 2.0198649508529343e-06, indicating strong evidence against the null hypothesis (the hypothesis that the data comes from a normal distribution). This suggests a significant deviation from normality. However, for losing teams, the p-value is 0.3103, indicating that the data is normally distributed.

Based on these results, the Mann-Whitney U test would likely provide more accurate results than the Two-Sample T-Test.

First, by applying the Mann-Whitney U test, we find that the p-value (0.001787) is less than 0.05, which is a commonly used significance level. This indicates that we reject the null hypothesis. Put simply, the observed difference in passing rates between winning and losing teams is unlikely to be due to chance, providing evidence that the passing rate differs between the two groups.

However, the Mann-Whitney U test only informs us that the medians of the two groups are likely different, without indicating which group has the higher median. To determine this, we would need to:

1. Directly compare the medians of the two groups. Looking at statistics in table three supports this.
2. Conduct additional statistical tests to ascertain the direction of the difference, such as applying the Two-Sample T-Test to our data.

Let's delve into the Two-Sample T-Test. Our null hypothesis is: “There is no difference in passing rate between winning and losing teams.” With a p-value of 0.0439, lower than the commonly chosen significance level of 0.05, we reject the null hypothesis. This implies that the passing rate of winning teams differs from that of losing teams.

Moreover, the positive t-statistic (2.0279) suggests that the average passing rate for winning teams is likely higher than for losing teams. Nonetheless, we cannot definitively determine the specific direction of the difference or the strength of the evidence from a single p-value alone.

4.1.2. Is the difference in the passing rate in games with a winner higher than the difference in games that ended in a draw?

To answer this question, first let's compare the passing rates of games that ended in a draw with those that didn't, as shown in Figure 4.

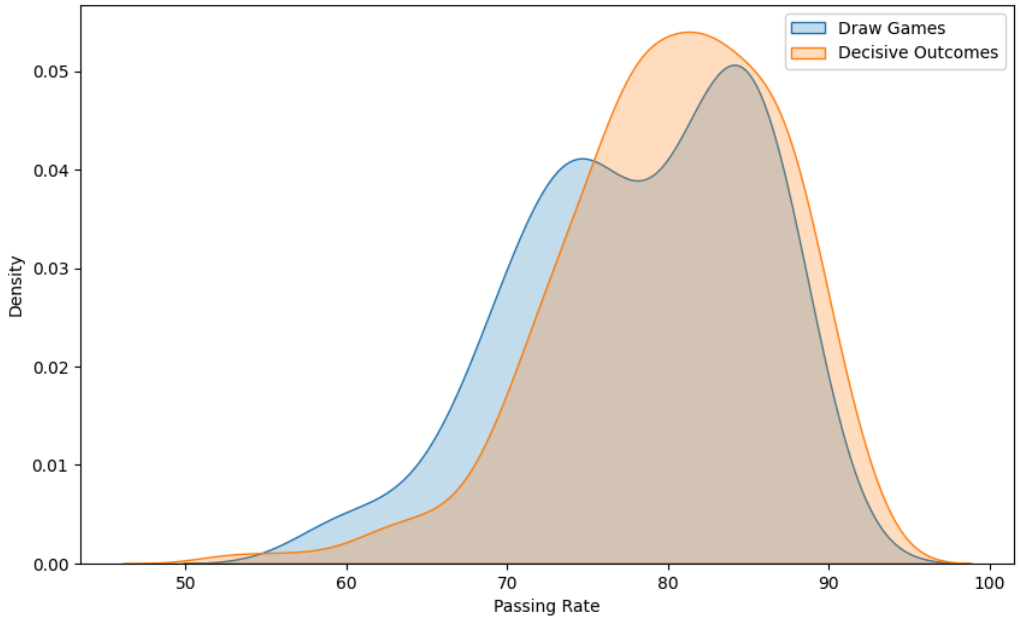

Figure 4. Passing Rates for Non-Draw Games and Decisive Outcomes

Based on our data, the best approach for answering our question is to utilize the F-test. The F-test was employed to compare the variances of passing rates between games with a winner and drawn games. This test was chosen because it directly focuses on variance comparisons, is robust to potential non-normality in the data, and permits hypothesis testing. By analyzing the F-statistic and p-value, we can investigate whether the spread of passing rates, potentially reflecting consistency and the influence of other factors, differs significantly between these game outcomes.

The null hypothesis tested whether the difference in passing rates between these two groups of games is statistically significant.

Based on the results of the F-test:

- The test statistic is 6.247.
- The associated p-value is approximately 0.0133.

With a significance level of 0.05 (alpha), since the p-value (0.0133) is less than the chosen significance level, we reject the null hypothesis. This indicates that the difference in passing rates between games with a winner and games that ended in a draw is statistically significant.

Therefore, we can conclude that there is a significant difference in the passing rates between games with a winner and games that ended in a draw in the German 1st soccer division. This can be supported by the statistical report in Table 3 as well.

5. Summary

This research project aimed to investigate the influence of passing rates on game outcomes in the German 1st soccer division (1. Bundesliga). Specifically, it sought to answer two key questions: whether winning teams have higher passing rates than losing teams, and if the difference in passing rates between games with winners and those ending in a draw is significant.

The analysis revealed compelling insights. Firstly, it was found that winning teams indeed tend to have higher passing rates compared to losing teams. Statistical tests confirmed this difference to be significant, suggesting that passing efficiency plays a crucial role in determining game outcomes. Moreover, the investigation into games ending in a draw demonstrated a notable variance in passing rates compared to games with winners.

These findings have significant implications in the real world context of soccer. Coaches, analysts, and teams can utilize this information for strategic planning and performance evaluation. Emphasizing passing efficiency in training and gameplay could potentially enhance a

team's chances of success. Furthermore, understanding the patterns of passing rates in different game outcomes can inform tactical adjustments during matches.

Moving forward, there are several open questions and further topics for analysis. Exploring additional factors that may influence passing rates and game outcomes, such as player positions, playing styles, or match conditions, could provide deeper insights. Additionally, investigating the longitudinal trends of passing rates across multiple seasons or comparing them across different soccer leagues could offer broader perspectives on the topic. Overall, this research opens avenues for continued exploration into the intricate dynamics of passing rates and their impact on soccer performance.

6. Bibliography

1. Shapiro, S.S. and Wilk, M.B. (1965). An analysis of variance test for normality (complete samples), Biometrika, 52, 591–611.
2. K. H. Yim, F. S. Nahm, K. A. Han and S. Y. Park, Analysis of statistical methods and errors in the articles published in the Korean Journal of Pain, The Korean Journal of Pain 23(1) (2010), 35-41.
3. C.-H. Chang and N. Pal, A revisit to the Behrens-Fisher problem: comparison of five test methods, Comm. Statist. Simulation Comput. 37(6) (2008), 1064-1085
4. Nadim Nachar. (2008). The Mann-Whitney U: A Test for Assessing Whether Two Independent Samples Come from the Same Distribution
5. Schumacker, Randall & Tomek, Sara. (2013). F-Test. 10.1007/978-1-4614-6227-9_11.
6. Wikipedia contributors. "F-test." Wikipedia, The Free Encyclopedia. Last modified January 5, 2024. https://en.wikipedia.org/wiki/F-test. Accessed March 2, 2024.
7. Python (matplotlib, pandas, scipy, seaborn)