Bundesliga Passing Accuracy Analysis Report

Author: Supratim Chakraborty

Date: June 13, 2024

Table of Contents

Introduction

This report aims to analyze the impact of passing accuracy on football matches. Understanding the factors that contribute to a team's success can help coaches and analysts improve team performance. The main purpose of this analysis is to check that whether higher passing accuracy is correlated with a greater likelihood of winning and passing accuracy difference between two competing team affects the outcome (win/loss or draw) or not .

As we delve into the intricate relationship between passing accuracy and match outcomes, it's impossible not to mention one of football's finest exemplars of this skill: Toni Kroos. Recently, Kroos announced his retirement from both club and international football (after the last dance in Euro Cup) , marking the end of an era for a player who has epitomized excellence in passing.

Throughout his career, Kroos has been a paragon of precision and efficiency on the field. His ability to control the tempo of the game, distribute the ball with pinpoint accuracy, and create scoring opportunities has been unmatched. This level of proficiency not only reflects his technical skills but also underscores the importance of passing accuracy in achieving footballing success.

Returning to the report preview, a comprehensive analysis of the problem at hand, methodologies applied, results, and conclusions drawn is provided. The Detailed Description of the Problem section outlines the dataset structures , pre processing and steps to tackle core questions raised. In the Methods section, details are given for the statistical and analytical techniques utilized to address these questions. The Evaluation section presents the step by step employment of the analysis specific to our research question and results . Summary section provides a concise recapitulation of the research and its implications , limitations and highlighting that winning teams generally exhibit higher passing rates than losing teams. Additionally, no significant difference in passing accuracy between games that end in a draw and those with a clear winner is found, indicating similar passing performances across these game outcomes. The Bibliography lists all the sources consulted and cited throughout the research. Finally, the Appendix includes supplementary material such as the link of full Python codebase used in our analysis.

Detailed Description of the Problem

Bundesliga dataset is containing match ID, passing accuracy percentage and win_indicator for the team playing. A unique game id signifies a single match where two entries present for each team and its passing accuracy and Yes/No Indicator to check whether respective team won or not . There are total 306 records in the dataset .

Couple of points regarding win_indicator :

1. If both entries for a single Match ID shows No, it signifies the game was drawn.
2. No such entries should not exist where both entries show Yes , cause only one team can be winner.

For Dataset , it was found there is missing values in column of passing accuracy and win_indicator for Match_ID = 139 . These 2 records are dropped from the dataset as part of Data Pre-processing and now there are 304 valid records at this stage .

Now coming back to 2 research questions mentioned ,

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

For both of the questions , the data is required to be at the each match level , because difference in passing rates between two team needs to be calculated which is possible only if data is at granularity of match ID level . So , as part of data transformation Data is self-joined at the level of match ID between two teams and 152 records are there at this point.

Coming to the first task where it is to determine if there is a statistically significant difference in passing accuracy between winner and loser for each match . Passing accuracy difference value is calculated for each match where the difference is (winner's passing accuracy – loser's passing accuracy) where positive value signifies winner having more accuracy than loser and negative value stands for loser having more accuracy than winner . An important point related to task is to eliminate the drawn games as analysis should be limited to non-draw games (114 records) based on the research question .

Coming to the second task where it is to determine if there is a statistically significant difference in passing accuracy between two team for two groups – drawn matches and non-draw matches. Passing accuracy difference value is calculated for each match where absolute value of the difference is taken in this case. An important point related to task is to eliminate the drawn games as analysis should be limited to as both non-draw and drawn games (152 records) based on the research question .

Additionally is_draw column is created to make data analysis easy as it would be used to check whether a certain match is draw (is_draw=True) or non-draw (is_draw=False)

Methods

For tasks such as data set loading, preprocessing, transformation, exploratory data analysis, and both statistical and inferential analyses, a Python notebook has been developed utilizing libraries such as numpy, pandas, math, matplotlib, scipy.stats, and seaborn. This codebase has supported the generation of various data, tables, and figures presented in this report. The github link to the codebase is available in the Appendix section of the report.

Post pre-processing and transformation of the data , for statistical analysis, several statistical methods will be employed. Descriptive statistics will provide an overview of the data, including measures of central tendency and dispersion. Histogram would be required to check whether distribution of passing accuracy difference is normal or not.

A one sample one tailed t-test will be used to compare the mean passing accuracy between winning and losing teams to determine if the difference is statistically significant. One-tailed test is chosen cause the check is directional (whether winner's passing accuracy is greater than loser's passing accuracy) .

Mann Whitney U test or Sum rank Test will be used to compare absolute passing accuracy difference between two samples – a group of drawn games and another group of non-draw games.

Each statistical methods used, would be described in subsections below.

Histogram

A histogram is a fundamental tool in statistical analysis used to visually summarize and display the distribution of a dataset.

1. **Bins**: The data range is divided into intervals called bins. Each bin represents a range of values.
2. **Frequency**: The height of each bar represents the frequency of data points within each bin. Taller bars indicate more data points in that range.
3. **Shape**: The overall shape of the histogram provides insights into the distribution:
    - **Normal Distribution**: Symmetrical, bell-shaped curve centered around the mean.
    - **Skewed Distribution**: Asymmetrical, with a longer tail on one side.
        - **Positively Skewed**: Tail extends to the right.
        - **Negatively Skewed**: Tail extends to the left.
    - **Uniform Distribution**: All bins have roughly the same height.

- **Bimodal/Multimodal Distribution**: Two or more peaks in the histogram, indicating multiple modes.

4. **Outliers**: Bars that are isolated from the rest of the data indicate outliers.
5. **Spread**: The width of the histogram shows the range of the data, indicating how spread out the values are.

Histograms are useful for visualizing the central tendency, variability, and the shape of the data distribution, helping to identify patterns and potential outliers.

Hypothesis Set up

A hypothesis test is a statistical method that uses sample data to evaluate a hypothesis about a population[4].

Notation for the Hypotheses [1,4] :

$H1$: The **alternative hypothesis** is the claim we wish to establish.

$H0$: The **null hypothesis** is the negation of the claim.

One Sample One tailed t-test

Steps to Perform a One Sample One tailed t-test:

1. **State the Hypotheses:** Define the null and alternative hypotheses[1,2].
2. **Select Significance Level (α):** Common choices are 0.05 or 0.01. This is the probability of rejecting the null hypothesis when it is actually true [4] .
3. **Calculate the sample mean ,sample standard deviation, sample size.**
4. **Calculate the t-statistic using the formula**: mentioned below.
5. **Determine the Critical t-value ($t\alpha$):** Use a t-distribution table or software to find the critical t-value for α and (n−1) degrees of freedom [6].
6. **Calculate the p-value:** The p-value can be determined using the t-distribution to find the probability of observing a test statistic as extreme as, or more extreme than, the observed value, under the null hypothesis.
7. **Make a Decision:** If $|t|>t\alpha$, or if the p-value is less than α, reject the null hypothesis. If $|t|\leq t\alpha$, do not reject the null hypothesis [2].

Hypotheses:

◎ **Null Hypothesis(H0):** The mean of the sample is equal to the known or hypothesized population mean($\mu 0$).

$\mu=\mu 0$

◎ **Alternative Hypothesis (H1):** The mean of the sample is greater or less than population mean.

$\mu < \mu 0$ or $\mu>\mu 0$

Assumptions:

- **Normality:** The sample data should be approximately normally distributed, which is crucial for small sample sizes.
- **Independence of Observations:** Each observation must be independent of the others, typically ensured by random selection.
- **Scale of Measurement:** Data should be on an interval or ratio scale to meaningfully compute statistical measures like mean and standard deviation.
- **No Outliers:** The data should be free from significant outliers, as these can skew results.

Test Statistic:

The test statistic for a one-sample t-test is calculated using the formula below [2]:

$$\boldsymbol{T} = \frac{(\bar{\mathbf{X}} - \boldsymbol{\mu})}{\boldsymbol{S}/\sqrt{\boldsymbol{n}}}$$

Where

$\bar{X}$= Sample Mean	μ = Population Mean
S = Sample Standard Deviation	n = Sample Size

Determine the Critical t-value ($t\alpha$):

Use a t-distribution table or software to find the critical t-value for α and (n−1) degrees of freedom.

Calculate the p-value: The p-value can be determined using the t-distribution to find the probability of observing a test statistic as extreme as, or more extreme than, the observed value, under the null hypothesis.

Make a Decision: If $|t|>t\alpha$, or if the p-value is less than α, reject the null hypothesis. If $|t|\leq t\alpha$, do not reject the null hypothesis.

Mann-Whitney U Test or Rank Sum Test

Purpose:

- To compare the distributions of two independent groups to determine if there is a significant difference between them.

Steps to Perform a Two-Tailed Mann-Whitney U Test [3,6]:

1. **State the Hypotheses**: Define the null and alternative hypotheses [1,3].
2. **Select Significance Level (α) and calculate $Z\alpha$ value**: Common choices are 0.05 or 0.01. Z_α value can be evaluated using z-score table or Statistical software based on the Significance Level (α) [4,6].
3. **Combine the Samples**: Combine the data from both groups and rank them together.
4. **Assign Ranks**: Assign ranks to the combined data. If there are ties, assign the average rank to the tied values.
5. **Sum of Ranks**: Calculate the sum of the ranks for each group ($R1$ and $R2$).
6. **Calculate the Test Statistic , Mean and Standard Deviation**: Use the formula mentioned below step by step.
7. **Calculate the Z value**: Using the value above and formula mentioned below, Z value is evaluated.
8. **Make a Decision**: Reject the null hypothesis if $Z < -Z_\alpha$ or $Z > Z_\alpha$.

Hypotheses:

- **Null Hypothesis (H_0)**: The distributions of the two groups are equal.
- **Alternative Hypothesis (H_1)**: The distributions of the two groups are not equal (two-tailed).

For a two-tailed test:

- **H_0**: The distributions of group A and group B are equal.
- **H_1**: The distributions of group A and group B are different.

Assumptions:

- The samples are independent.
- The data are ordinal, interval, or ratio.
- The shapes of the distributions in both groups are similar (though the test is robust to some differences in shape).

Critical Z-values for a Two-Tailed Test:

1. **Significance Level (α):** Preferred Choices are 0.05 or 0.01

2. **Critical Values (Z_α) :** Values will be evaluated from z score table or any statistical software.

Test Statistic:

The Mann-Whitney U statistic is calculated based on the ranks of the data rather than their actual values. Here are the steps [3,5]:

1. **Combine the Samples**: Combine the data from both groups and rank them together.
2. **Assign Ranks**: Assign ranks to the combined data. If there are ties, assign the average rank to the tied values.
3. **Sum of Ranks**: Calculate the sum of the ranks for each group ($R1$ and $R2$).

The test statistics U_1 and U_2 are calculated as:

$$\boldsymbol{U1 = n1.n2 + \frac{n1\,(n1+1)}{2} - R1}$$

$$\boldsymbol{U2 = n1.n2 + \frac{n2\,(n2+1)}{2} - R2}$$

Where:

- n1 and $n2$ are the sample sizes of the two groups.
- $R1$ and $R2$ are the sums of the ranks for the two groups.

The test statistic U is the smaller of U_1 and U_2.

$$\mathbf{U = min(U_1,U_2)}$$

Mean (μ) and standard deviation (σ) of U statistic

$$\mu = \frac{\boldsymbol{n1.n2}}{\boldsymbol{2}}$$

$$\boldsymbol{\sigma = \sqrt{\frac{n1.n2\,(n1+n2+1)}{12}}}$$

Z-value can be evaluated from the U statistic value , mean and variance .

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

Final Decision Rule

- **Reject the null hypothesis if**: $Z<-Z_\alpha$ or $z>Z_\alpha$
- **Fail to reject the null hypothesis if** : z-value falls inside the range $[-Z_\alpha, Z_\alpha]$.

Evaluation

The evaluation will consist of two main parts: descriptive analysis and inferential analysis.

Descriptive Analysis

Table 1. Sample Size, central tendency, range of passing accuracy in the original data set (minus error record removed).

Measure	Value
Count	304
Mean	79.68092
Standard Deviation	6.960058
Min	53
25%	75
(Median) 50%	80
75%	85
Max	92

The dataset containing passing rate information consists of 304 observations. The mean passing rate across these observations is approximately 79.68%. The standard deviation, which measures the variation in passing rates, is around 6.96%. This indicates a moderate spread of data around the mean.

The minimum passing rate recorded in the dataset is 53%, while the maximum is 92%, showing a wide range of outcomes. The median passing rate is 80. These statistics provide a comprehensive overview of how passing rates are distributed across the dataset, helping to understand the central tendency and variability within the data.

Table 2. Sample Size, central tendency, range of the passing accuracy difference of non-draw games in pre-processed data set.

Measure	Value
Count	114
Mean	1.816
Standard Deviation	9.888
Min	-29
25%	-5
(Median) 50%	3
75%	8
Max	25

The dataset on passing accuracy difference for non-draw games, calculated as winner’s accuracy minus loser’s accuracy, consists of 114 observations with an average difference of 1.816. This suggests a slight tendency for winners to have higher passing accuracy. The data shows significant variability, as indicated by a standard deviation of 9.888, with differences ranging from -29 to 25. The negative values indicates that the losing team surpassed the winner in accuracy for those matches. The median difference of 3 indicates a slight edge in passing accuracy for winners in the majority of games.

Histogram will help to illustrate the central tendency, dispersion, and shape of the data's distribution. In the context of analyzing differences such as passing accuracy between winners and losers, it will allow to observe the frequency of data occurrences within specific intervals, making it easier to identify patterns such as normality, skewness in the data.

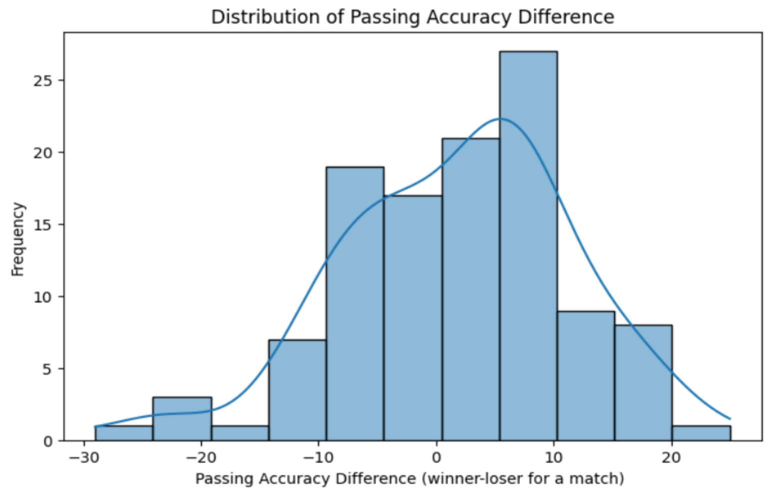

Figure 1. Histogram of the difference in passing accuracy. The x-axis represents the passing accuracy difference, calculated as the winner's passing accuracy minus the loser's passing accuracy for each match. The y-axis shows the frequency of these differences. The bell shaped curve suggests that the data approximates a normal distribution, centering around a positive difference. This indicates that winners tend to have higher passing accuracy than losers, on average.

From the histogram, it is evident that the distribution of passing accuracy differences closely approximates a bell-shaped curve, which suggests that the data are approximately normally distributed. This normal distribution indicates that while there are variations, most differences cluster around a positive mean difference. This central clustering around a positive value implies that, on average, winning teams tend to exhibit higher passing accuracy compared to their losing counterparts. The presence of data points across a range from negative to positive differences also highlights variability in the data, showing that there are instances where losing teams have outperformed winners in terms of passing accuracy.

Inferential analysis

1st Question:

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

To test whether the winner of a match has a higher passing rate than the loser, we can set up the hypothesis test as follows :

Null Hypothesis (H0): There would not be any bias between passing rate difference and (outcome) win or loss in the match . This means the mean difference in passing rates (winner passing rate - loser passing rate) would be zero.

H0: $\mu = 0$

Alternative Hypothesis (H1): Winners have a higher passing rate than losers, which means the mean difference in passing rates is greater than zero.

H1: $\mu > 0$

Given that the differences appear to be normally distributed and hints at winner having higher passing accuracy than loser [Refer to Figure 1], we can proceed with a one-sample t-test to test our hypothesis.

<u>Why One sample one tailed T-test?</u>

A one tailed t-test is chosen in this case because our alternative hypothesis is directional, specifically stating that the winners have a higher passing rate than the losers (**H1:** $\mu > 0$)

Common choice for significance level of the test is 0.05 [4]. Here are the underlying values for the t-test formula:

- **Sample Mean ($\bar{X}$):** 1.816 [Refer to Table 2]
- **Population Mean (μ):** 0 (Hypothesized as 0)
- **Sample Standard Deviation (S):** 9.888 [Refer to Table 2]
- **Sample Size (n):** 114 [Refer to Table 2]
- **Degree of Freedom :** 113 (Sample Size -1)

These values are used in the formula

$$T = \frac{(\bar{X} - \mu)}{S/\sqrt{n}}$$

$$T = \frac{(1.816 - 0)}{9.888/\sqrt{114}}$$

Evaluating the values , **T-statistic (T) =** 1.961

Table 3. t-distribution Table for one-tailed test [6].

Degrees of Freedom	α = 0.1	α = 0.05	α = 0.025	α = 0.01	α = 0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.92	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.44	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.86	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.25
10	1.372	1.812	2.228	2.764	3.169
50	1.299	1.676	2.009	2.403	2.678
100	1.29	1.66	1.984	2.364	2.626
113	1.289	1.658	1.981	2.36	2.62
1000	1.282	1.646	1.962	2.33	2.581
infinite	1.282	1.645	1.96	2.326	2.576

The critical value for a one-sample t-test at a 0.05 significance level with 113 degrees of freedom is approximately 1.658 [Refer to Table 3] . This means that if the calculated t-statistic is greater than 1.658, we would reject the null hypothesis in favor of the alternative hypothesis.

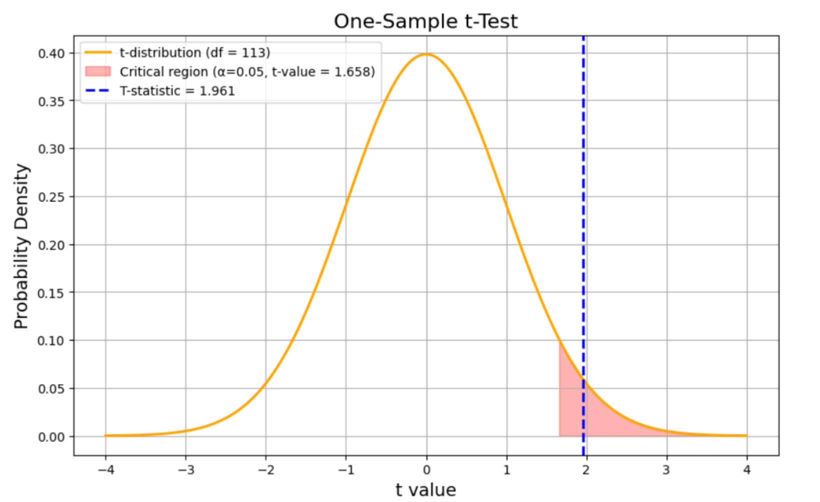

Figure 2. Probability density function of the one-sample t-test. This graph illustrates the t-distribution with 113 degrees of freedom. The red vertical line at t = 1.658 denotes the critical value for a one-sided test at the 0.05 significance level. The blue dashed line represents the calculated t-statistic of 1.961. The shaded area under the curve to the right of the critical value indicates the rejection region where the null hypothesis would be rejected in favor of the alternative hypothesis. The observed t-statistic falls within this critical region, suggesting that the sample provides sufficient evidence, at the 5% significance level, to reject the null hypothesis in favor of the alternative.

Conclusion

We reject the null hypothesis that there is no difference in passing rates between the winning and losing teams. The data provides evidence that winners have a higher passing rate than losers. This conclusion is based on the t-statistic of 1.961, which exceeds the critical value of 1.658 at the 0.05 significance level. The p-value of approximately 0.026 is less than the significance level of 0.05, further supporting the rejection of the null hypothesis. This analysis supports the alternative hypothesis that winning teams exhibit a higher passing rate. The conclusion is made with a confidence level of 95%. This means that if the experiment were repeated under the same conditions, we would expect the conclusion that winners have a higher passing rate than losers to hold true in 95% of those repetitions.

2nd Question:

Is there statistically significant difference in passing accuracy between two team for two groups – drawn matches and non-draw matches ?

To test whether the passing accuracy between two team for two groups – drawn matches and non-draw matches differ , we can set up the hypothesis test as follows:

Hypotheses

1. **Null Hypothesis (H0):** There is no difference in the expected difference in passing rate in games that end in a draw from the expected difference in passing rate in games with a winner.
2. **Alternative Hypothesis (H1):** The expected difference in passing rate in games that end in a draw is different from the expected difference in passing rate in games with a winner.

Critical Z-values for a Two-Tailed Test:

3. **Significance Level (α):** Standard Significance level of 0.05 is chosen
4. **Critical Values:**
   - The cumulative probability for the lower tail (left side) is 0.025.
   - The cumulative probability for the upper tail (right side) is 1 - 0.025 = 0.975.

Using the standard normal distribution table or Z-table [Refer to Table 4] :

- **Lower Critical Value($-Z_\alpha$):** Z-value corresponding to a cumulative probability of 0.025 is approximately -1.96
- **Upper Critical Value(Z_α):** Z-value corresponding to a cumulative probability of 0.975 is approximately 1.96

Table 4. Positive and Negative z-Score Table [6].

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931

The groups mentioned are

1. Drawn games (category I)
2. Non-draw games (category II)

There are 38 drawn games ($\boldsymbol{n1} = \mathbf{38}$) and 114 non-draw games ($\boldsymbol{n2} = \mathbf{114}$)

After Combining the Samples, ranks are assigned to the combined data.

If there are ties, assign the average rank to the tied values.

Category	II	I	II	II	II	I	……
Rank	1.5	1.5	8	8	8	8	……
Value (abs_diff_passing_rate)	0	0	1	1	1	1	……

$\boldsymbol{R1} = 2831$ (for Category I) $\boldsymbol{R2} = 8797$ (for Category II)

The test statistics U_1 and U_2 are calculated as:

$$\boldsymbol{U1 = 38 * 114 + \frac{38\,(38+1)}{2} - 2831}$$

$$\boldsymbol{U1 = 2242}$$

$$\boldsymbol{U2 = 38 * 114 + \frac{114\,(114+1)}{2} - 8797}$$

$$\boldsymbol{U2 = 2090}$$

The test statistic U is the smaller of U_1 and U_2.

U = min(U_1,U_2)

$$\boldsymbol{U = 2090}$$

Mean (μ) and standard deviation ($\boldsymbol{\sigma}$) of $\boldsymbol{U}$ statistic

$$\mu = \frac{\boldsymbol{n1.n2}}{\boldsymbol{2}}$$

$$\mu = \frac{\boldsymbol{38 * 114}}{\boldsymbol{2}}$$

$$\mu = \boldsymbol{2166}$$

$$\boldsymbol{\sigma = \sqrt{\frac{n1.n2\,(n1+n2+1)}{12}}}$$

$$\boldsymbol{\sigma = \sqrt{\frac{38 * 114\,(38+114+1)}{12}}}$$

$$\boldsymbol{\sigma = 235.02}$$

Z-value can be evaluated from the U statistic value , mean and standard deviation .

$$\boldsymbol{Z = \frac{U - \mu}{\sigma}}$$

$$Z = \frac{\mathbf{2090 - 2166}}{\mathbf{235.02}}$$

$$\boldsymbol{Z = -0.323}$$

Decision Rule

- **Reject the null hypothesis if**: $z<-1.96$ or $z>1.96$
- **Fail to reject the null hypothesis if** : z-value falls inside the range $[-1.96, 1.96]$.

Test Result

- **Calculated Z-value**: -0.323
- **Decision**: Since the calculated Z-value of -0.323 falls within the range $[-1.96, 1.96]$, we fail to reject the null hypothesis [Refer to Figure 3].

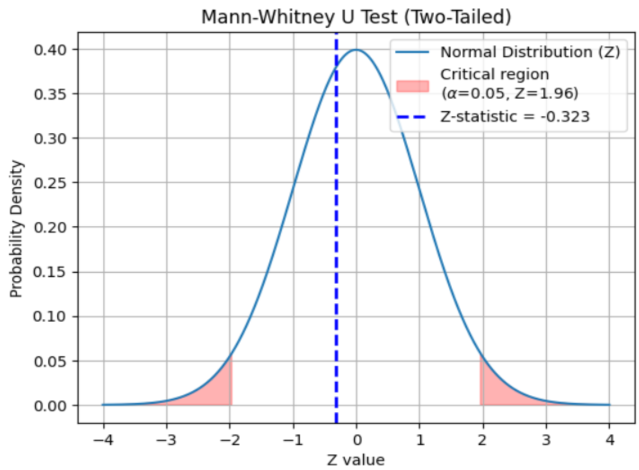

Figure 3. Probability density function of the Mann-Whitney U test (Two-Tailed). This graph represents the standard normal distribution (Z). The critical region for a two-tailed test at a significance level of α=0.05 is indicated by the red shaded areas, corresponding to Z-values of ±1.96. The blue dashed line denotes the observed Z-statistic of −0.323. Since this Z-statistic does not fall into the critical regions, it lies within the range of common values under the null hypothesis, suggesting that there is insufficient evidence to reject the null hypothesis. Therefore, the data does not show a statistically significant difference in passing accuracy between games that end in a draw and those with a winner, implying similar passing performance across these types of games

Conclusion

Based on the Mann-Whitney U test, the Z-statistic is -0.323, which does not fall within the critical region defined by Z=±1.96 for a significance level of α=0.05. Based on this test, it is indicated that we do not have sufficient evidence to state that there is a statistically significant difference in the passing accuracy difference between games that end in a draw and games with a winner. Therefore, it can be concluded that the expected difference in passing rate is similar for both types of games.

Summary

This report analyzed the impact of passing accuracy on football matches. The analysis showed that there is a significant difference in passing accuracy between winning and losing teams, with winners demonstrating higher passing accuracy. These findings suggest that improving passing accuracy could be a key strategy for enhancing team performance. While winners generally exhibit higher passing accuracy compared to losers, the difference in passing rates between drawn and non-drawn matches is not significant. This indicates that while superior passing can contribute to a team's likelihood of winning, it does not necessarily prevent matches from ending in a draw. Further research could explore additional factors that contribute to match outcomes.

Added Insights and Limitations:

- **Contextual Limitations:** The dataset is limited to 152 matches, which may not fully represent broader trends across different leagues or seasons. The findings might be influenced by specific team strategies or league styles that prioritize passing differently.
- **Data Limitations:** Since the dataset only provides passing accuracy without considering other tactical aspects such as possession time, shots on target, chance created or defensive strategies, the analysis is somewhat restricted. Passing accuracy alone may not fully capture the complexity of match outcomes.
- **Implications for Teams:** Coaches and analysts might focus on enhancing passing accuracy as part of their strategy for winning, but should also consider other factors that contribute to outcome of the match, such as defensive solidity or effectiveness in finishing.

This nuanced understanding underscores the multifaceted nature of football, where multiple tactical elements interact to determine the outcome of a match, beyond just the accuracy of passes.

Bibliography

1. A. Johnson, Richard. *Miller & Freund's Probability and Statistics for Engineers*, 2017, Ninth Edition, 242-247
2. S. Moore, David; P. McCabe, George; A. Craig, Bruce. *Introduction to the Practice of Statistics* , 2009, Sixth Edition, 422-427
3. DATAtab Team (2024). *DATAtab: Online Statistics Calculator. DATAtab e.U. Graz, Austria*. URL https://datatab.net/tutorial/mann-whitney-u-test
4. J. Gravetter, Frederick; B. Wallnau, Larry. *Statistics for the Behavioral Sciences*, 2015, Tenth Edition, 225-233
5. Newbold, Paul; L. Carlson, William; Thorne, Betty . *Statistics for Business and Economics* , 2013, Eighth Edition, 628-632
6. Navidi, William . *Statistics for Engineers and Scientists* ,2023, Sixth Edition, 824-826

Appendix

Bundesliga Passing Accuracy Report Python Codebase