筑波大学

University of Tsukuba

Master's in Computer Sciences

Experiment Design in Computer Sciences

Experiment Report
Soccer Matches Results Analysis

May 2, 2022

Professor:
Claus Aranha

Felipe Nonato Cardoso Sobral Júnior

1 Introduction

This report aims to investigate the results of soccer games regarding the win-ratio of soccer teams in each place they play. That is, analyze the ratio of wins for the home team and compare different sets of positive results. A draw may or may not be a positive result, and for this reason experiments are done by considering all possible roles for it: draw as a negative result, positive result and not considering draw in the analysis.

It is expected by the end of this experiment to have ratios of positive results against negative results for the home team and determine with a certain level of confidence an interval of percentage of games that the home team obtains a positive result in different considerations for a draw.

2 Experiment Design

The experiment is a *Retrospective Experiment*, that is, historical data will be used in the analysis shown in the next sections. In this case, data from the *Brasileirão Assaí: Série A* - the Brazilian soccer championship - was used. This data is publicly available in the platform *Kaggle* [3]. The methodology applied is resumed in Figure 1.

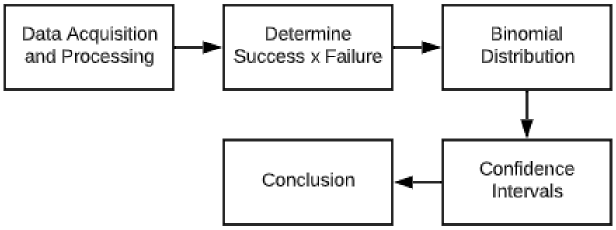


Fig. 1: Methodology flowchart.

In this report *SciPy*'s library *binom* [4] was used to obtain the results of each experiment. Nevertheless, the manual calculus was also included in the code for consultation purposes. The code is available in [2].

2.1 Data Acquisition and Processing

The data obtained [3] has other information that is not necessary for the analysis done in this report. For this reason, only the column named ”Res” was used. This column identifies the match result by using H for home team wins, A for away team wins and D for draws.

After separating the relevant columns, 3820 number of lines (matches) were identified. These lines were then checked for null values or any other incompatibility. Null values were deleted. After the deletion, data of 3819 matches were left behind.

2.2 Determine Success x Failure

The goal of this report is to analyze successes (positive results) against failures (negative results). These may have different meanings on different occasions of a soccer match. In this case, three possibilities were considered:

- Experiment 1 (E1):
  - **Success**: Home team win (H)
  - **Failure**: Away team win (A) + Draw (D)
  - **Sample (Total Matches):** Home team win (H) +Away team win (A) + Draw (D)
- Experiment 2 (E2):
  - **Success**: Home team win (H) + Draw (D)
  - **Failure**: Away team win (A)
  - **Sample (Total Matches):** Home team win (H) +Away team win (A) + Draw (D)
- Experiment 3 (E3):
  - **Success**: Home team win (H)
  - **Failure**: Away team win (A)
  - **Sample (Total Matches):** Home team win (H) +Away team win (A)

2.3 Binomial Distribution

The variable analyzed in this experiment - the match result - is a binary variable. That is, it can assume only two values: *Success* or *Failure*. The result of each match in this case is considered random and ruled by a probability. In this case, each team has a 50% chance of getting either a positive or negative result. For this reason the adequate distribution to fit the data of this experiment is the Binomial Distribution [1]. Because of this, some assumptions are taken into account:

- Let p represent the estimated proportion of successes obtained from a sample of n matches where X matches resulted in successes - it is important to remember that ”success” may vary depending on what experiment is being done (E1, E2 or E3). Equation 1 describes the relationship between this variables. With these, it is possible to plot the mass probability function of this distribution.

$$p = \frac{X}{n} \tag{1}$$

- The mean μ is described in Equation 2.

$$\mu = n \cdot p \tag{2}$$

- The variance σ^2 is described in Equation 3.

$$\sigma^2 = n \cdot p \cdot (1 - p) \tag{3}$$

2.4 Confidence Interval

The confidence interval is the statistical interval chosen to evaluate the data obtained by using Equations 1 to 3. These equations will enable the generation of an interval that has a calculated chance of describing the results of a general soccer match. The formula that describes the confidence interval is shown in Equation 4.

$$p - Z\alpha \cdot \sqrt{\frac{p \cdot (1 - p)}{n}} \leq p \leq p + Z\alpha \cdot \sqrt{\frac{p \cdot (1 - p)}{n}} \tag{4}$$

2.5 Experiments' Analysis and Results

In the Table 1 the general data obtained from the data-set after processing is presented.

Match Result	Number of Matches
H (Home Team Win)	1874
A (Away Team Win)	919
D (Draw)	1026
Total	3819

Table 1: Data-set description.

It is noticeable in Table 1 that the most frequent result of a match is the victory of the home team, followed by a draw and the victory of the away team (or loss of the home team).

In Table 2 is shown the result value for successes and failures used for each experiment as explained in Section 2.2. It is important to note that the sample size (n) for Experiment 3 is smaller due the fact that draws are not considered.

	Experiment 1	Experiment 2	Experiment 3
Sucesses (X)	1874	2900	1874
Failures	1945	919	919
Sample Size (n)	3819	3819	2793

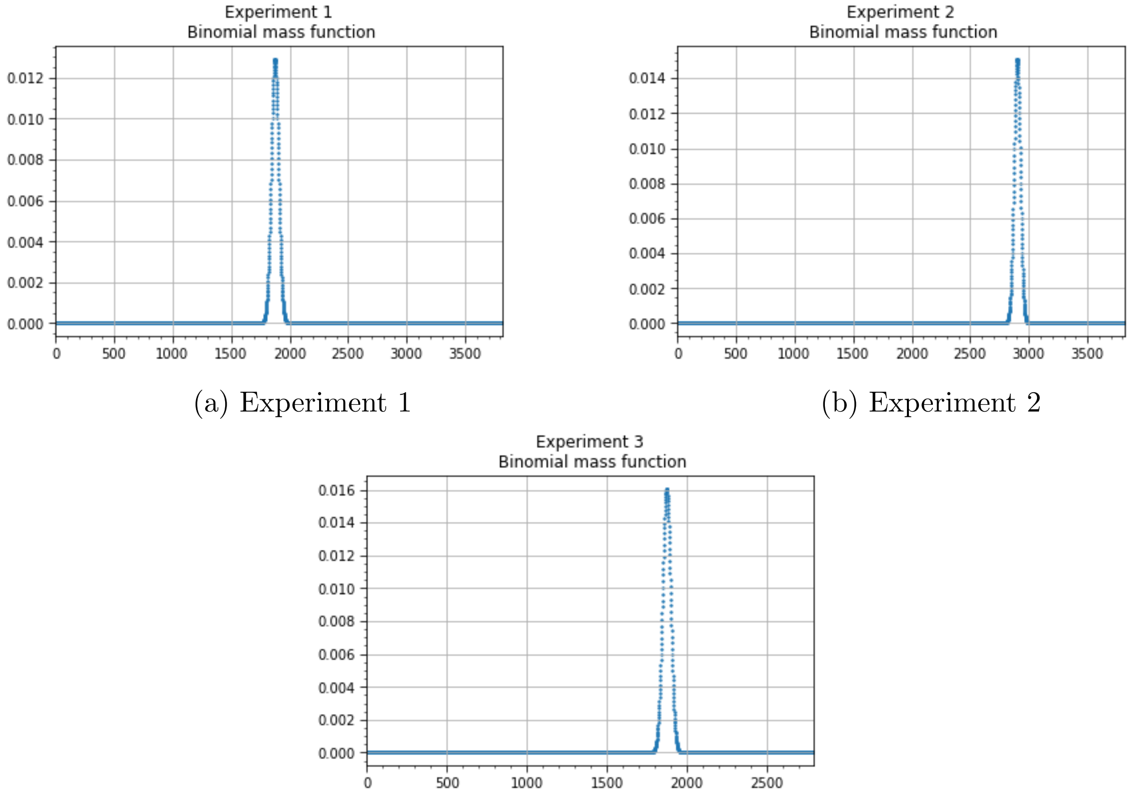
Table 2: Values of successes, failures and sample size (total games considered) for each experiment.

	Experiment 1	Experiment 2	Experiment 3
p	0.4907	0.7594	0.6710
n	3819	3819	2793
Mean	1874	2900	1874
Variance	954.42	697.8528	616.6152

Table 3: Results of experiments.

Next, with the values in Table 2 it is possible to determine the proportion p (proportion of success) of each experiment. And with the value of p and n, calculate the other values shown in Section 2.3. These calculated values are presented in Table 3.

A form to see the probabilities of a Binomial Distribution is the probability mass function. This graph can be obtained with the values of p and n. In Figures 2a to 2c are displayed these mass functions. The X-axis of these graphs represents the number of games and the Y-axis the probability of successes of this specific quantities.



(a) Experiment 1

(b) Experiment 2

(c) Experiment 3

Fig. 2: Probability mass functions of each experiment.

It is possible to see that the curve of E2 is dislocated when compared to E1 (they both have the same sample sizes). This happens due the shift in the ”meaning” of success in these two experiments. Where in E1 a draw was considered a negative result, in E2 it was considered a positive result. This change can be also seen in Table 3, where the value of p and the *mean* are higher in E2.

On the other hand, in E3 it is possible to identify that the p value although higher than E1 is lower than E2. The mean is also equal to E1's, although the curve is dislocated from the center, as the p value is more than 0.5 (there are more victories of the home team than the away team.)

Finally, the confidence intervals with 95% of confidence are in exhibited in Table 4

	Confidence Interval (95% confidence)
Experiment 1	$0.47473 \leq p \leq 0.5068$
Experiment 2	$0.74574 \leq p \leq 0.77298$
Experiment 3	$0.65342 \leq p \leq 0.68851$

Table 4: Confidence intervals of each experiment.

As shown in Table 4, for E1 it is possible to say with 95% confidence that between 47.47% and 50.68% of the matches result in a positive outcome for the home team. For E2, this interval goes up to between 75.57% and 77.3% of matches. And for E3 between 65.34% and 68.85% of the matches results are positive for the home team.

3 Conclusion

In the Brazilian Championship analyzed in this report, each victory awards a team with three points, a draw awards one point for both teams and the losing team does not take any points. For this reason, a draw may be seem as a positive result, as the team "avoided" not earning any points at from said match (E2). But contrary to this belief supporters generally include a draw in the negative results batch (E1), as since the home team is playing close to its supporters, the moral boost would make the team win the game or the supporters' pressure would cause the visiting team to make more mistakes.

In this report it is shown with 95% of confidence in the calculus for all three experiments, that when following most of the supporters perception of a draw (negative result), the home team has approximately 47.5% to 50.7% chance to get a positive result when it plays in its own stadium (See Table 4 for exact estimates). On the other hand, sometimes, specially when playing against a stronger team, a draw can be considered a positive result. Another way to examine E2's proposal is to consider that a positive result is when the home team does not lose a game when playing at its origin city. In this case, chances goes up by a fair amount: between 74.5% and 77.2% of matches wield a positive result.

Finally, E3 brings the result considering only matches that did not end in a draw. If this was the case in real soccer matches (if draws did not exist), the perception of most supporters would be justified, as in E3 between 65.3% and 68.8% of the matches had a positive result for the home team (ended in a home team victory).

References

[1] Susan Dean Alexander Holmes, Barbara Illowsky. *Introductory Business Statistics*, 11 2017. 2

[2] Felipe Sobral Junior. *Report 1 of Experiment Design in Computer Sciences - Soccer Matches Results Analysis*, 05 2022. Available at `https://www.kaggle.com/code/felipesobraljr/firstreport-experiment-design`. 1

[3] Gabriel Meireles. *Brazilian Football Championship - Results of the first division of the Brazilian championship since 2013*, 04 2022. Available at `https://www.kaggle.com/datasets/gabrielmeireles/brazilian-football-championship`. 1

[4] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. 1