2022 FIFA World Cup – Statistical Analysis

Mouaz Ali

Table of Contents

Introduction

Every four years, the FIFA World Cup takes center stage and brings people together in a symphony of passion, skill, and national pride. This extraordinary event captures the hearts of millions of people worldwide, cutting over boundaries, languages, and cultural divides. The World Cup, which has been regarded as the height of international soccer competition, is more than just a competition; it's a celebration of the beautiful game and evidence of the ability of sport to bring people together. As the drama plays out on the biggest platform, nations compete for glory, players rise to the status of heroes, and everyone on the planet holds their breath. The 2022 FIFA World Cup, hosted by Qatar, was no exception to this spectacle, featuring thrilling matches, stunning goals, and unforgettable moments that will be carved into soccer history. Against the backdrop of the desert sands and state-of-the-art stadiums, teams from across the globe battled it out for the coveted title of world champions.

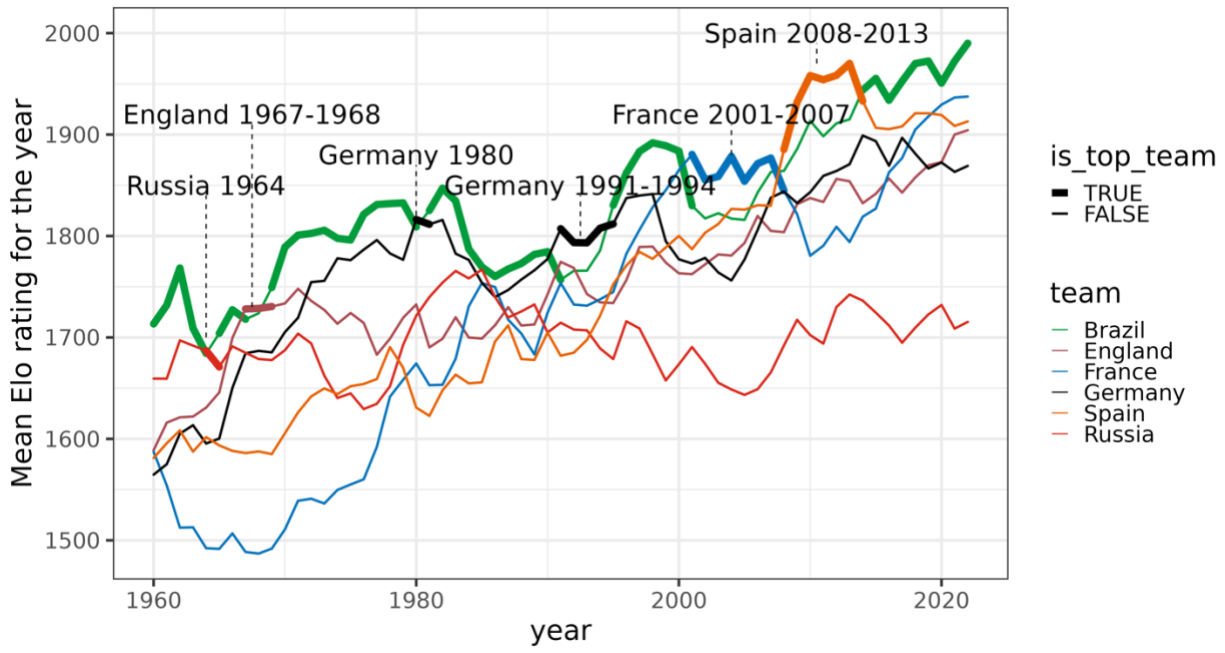
The FIFA World Cup, often simply called the World Cup, is an international association football, or soccer, competition contested by the senior men's national teams of the members of the Federation International de Football Association (FIFA), the sport's global governing body. Since the tournament was started in 1930, the World Cup is held every four years, except in 1942 and 1946 due to World War II. After witnessing the events unfold last year, the current champion is Argentina, which won its third title at the 2022 tournament in Qatar.

The dataset, which was put together through web scraping from FIFA.com, contains numerical information of all matches of the 2022 FIFA World Cup. Having 88 columns, the data can be used for multiple purposes such as research, application, or simply learning. After being cleaned and discarding unnecessary information, the dataset includes various statistical data for each match played during the tournament, such as: score, assists, possession, crosses, passes,

fouls, number of cards given out, offsides, and what part of the pitch certain events took place. These small pieces of information and statistics have played a major role in each one of the matches played at the tournament.

Statistical Predictions

Leading up to the 2022 FIFA World Cup, the predicted winners were Brazil based on statistical analysis. By examining all international soccer matches, including friendly games, for teams with at least 200 games, a model was created. To keep it simple, only wins, draws, and losses were taken into account. A win would give a team one point, while a draw gives half a point, and a loss will result in zero points. For example, in the 2014 World Cup final, Germany beat Argentina to become world champions with a score of 1-0, which gives Germany one point in the model and Argentina zero. Additionally, advanced statistical models such as Elo ratings were employed to forecast the outcomes of matches. Elo ratings, named after physicist Arpad Elo, are a method to quantify the relative skill levels of players in zero-sum games such as chess. This approach was adapted to soccer and proved to be remarkably accurate in predicting match outcomes. By analyzing historical data and accounting for factors such as team performance, player form, and match conditions, these models provide valuable insights into the likely results of matches, helping fans and analysts alike to anticipate the twists and turns of the tournament.



Performance Analysis

Analyzing the 2022 FIFA World Cup statistics offers some intriguing insights into the teams' performances. Possession is an interesting statistic to examine. While Argentina and England concentrated on fast counterattacks, teams like Brazil and Spain preferred to enjoy more possession and showcase their quick passing technique. Goal difference, which indicates how many more goals a team scored than they gave up, is another important statistic. Players like Neymar and Gabriel Jesus helped Brazil score a lot of goals, but Italy's defense was solid under the leadership of Giorgio Chiellini and Leonardo Bonucci. It's also thrilling to look at individual player statistics, as stars like Lionel Messi and Kylian Mbappe had been demonstrating their amazing abilities all tournament long, especially during the final.

Furthermore, examining performance indicators like successful tackles, shots on target, and pass completion percentage can reveal important details about individual effectiveness as well as team tactics. Spain and Germany, two teams with strong pass completion rates, were known for managing the game's speed and patiently setting up scoring opportunities. Teams like

Portugal and Belgium, on the other hand, who had a lot of shots on goal, relied on their offensive strength to penetrate rival defenses. During pivotal stages of the tournament, teams such as England and France showcased their defensive strength, showcasing their ability to recover possession and play effective defense with successful tackles.

Elo Ratings

Elo ratings were developed to quantify chess players' skills and deter elite players from merely playing against weaker opponents in order to rack up a lot of points. Elo ratings are named after physicist Arpad Elo and are not to be confused with the British music band ELO from the 1970s. Players with comparable Elo ratings have an equal chance of winning since Elo ratings may also be converted into probabilities.

Well, how does Elo function? The calculations are pretty straightforward actually. The winner receives points from the loser, with the amount of points awarded based on the disparity in ranks. A player with a higher ranking receives few points when they defeat a player with a lower ranking, but they receive many points when they defeat a player with a higher ranking. In this case, we know that England lost to Croatia 1-2 in the 2018 FIFA World Cup. Before the game had began, England was favored to win with a higher Elo score. Surprisingly, Croatia was able to defeat England in extra time, gaining 12 ELO points. On the other hand, England lost 12 ELO points. Generally, ELO changes function both ways. Switzerland's victory over Spain in the 2010 World Cup group stage was a major upset. Spain, the eventual World Cup victor, had the greatest Elo rating at the time and, again excluding draws, was expected to win by 84%. As we will see in a moment, they fell short by 17 Elo points. Another example of this in the previous World Cup is Argentina being defeated by Saudi Arabia. No one had expected Argentina to lose their first game against Saudi, yet they eventually went on to win the tournament.

Team	Current Elo rating
Brazil	2000
Argentina	1944
Spain	1915
France	1906
Belgium	1897

In the chart above are the ELO ratings leading up to the 2022 FIFA World Cup. As you can see, Brazil has been on top of the ratings since 2013. ELO ratings offer a compelling framework for assessing team strengths and predicting match outcomes. In the context of the 2022 FIFA World Cup, teams entered the tournament with varying ELO ratings, reflecting their recent performances in international competitions. For instance, Brazil, historically a dominant force in world soccer, consistently maintained a high ELO rating leading up to the tournament. This high rating was reflective of their strong performance in qualifiers and international friendlies in the preceding years. On the other hand, teams like Croatia and Belgium had ELO ratings slightly lower than traditional powerhouses like Brazil and Germany. However, as evidenced by past tournaments, ELO ratings provide a dynamic view of team strengths, with upsets and surprising victories often reshuffling the rankings. Therefore, while ELO ratings offer valuable insights,

they are not foolproof predictors of match outcomes, as demonstrated by the unpredictable nature of soccer.

Now, let's use the data set to answer some questions.

Problem 1

Let's consider Group F for the group stage matches in the 2022 FIFA World Cup involving Belgium (BEL), Canada (CAN), Morocco (MAR), and Croatia (CRO). We'll consider each team playing two matches, and each match can result in either a win for one team or a draw.

a. Define the sample space S using set notation:

S={(BEL,BEL),(BEL,CAN),(BEL,MAR),(BEL,CRO),(CAN,BEL),(CAN,CAN),(CAN,MAR),(CAN,CRO),(MAR,BEL),(MAR,CAN),(MAR,MAR),(MAR,CRO),(CRO,BEL),(CRO,CAN),(CRO,MAR),(CRO,CRO)}

b. Determine the probability of Belgium (BEL) winning at least one match.

$$P(BEL) = \frac{\#\ of\ outcomes\ where\ Bel\ wins\ at\ least\ 1\ match}{Total\ \#\ of\ outcomes\ in\ S}$$

Assuming Belgium wins both matches, there are two outcomes where Belgium wins at least one match:

$$P(BEL) = \frac{2}{16} = \frac{1}{8}$$

c. Find the probability of either Canada (CAN) or Morocco (MAR) winning their respective matches.

$$P(CAN \cup MAR) = \frac{\#\ outcomes\ where\ CAN\ or\ MAR\ win}{Total\ \#\ of\ outcomes}$$

$$P(CAN \cup MAR) = \frac{4}{16} = \frac{1}{4}$$

d. Calculate the probability of no team winning a match (all matches ending in a draw)

$$P(\overline{BEL} \cap \overline{CAN} \cap \overline{MAR} \cap \overline{CRO}) = \frac{1}{16}$$

e. Determine the probability of Croatia (CRO) not winning any match

$$P(\overline{CRO}) = \frac{6}{16} = \frac{3}{8}$$

Problem 2

Consider the knockout stage of the 2022 FIFA World Cup, where there are two remaining teams: Argentina (ARG) and Croatia (CRO). The probability of Argentina reaching the final is P(ARG)=0.6, and the probability of Croatia reaching the final is P(CRO)=0.4.

Defining the following events:

A: Argentina wins the semifinal match

B: Croatia wins the semifinal match.

a. Write the expression for P(ARG|A) in terms of conditional probability.

$$P(ARG|A) = \frac{P(ARG \cap A)}{P(A)}$$

b. Calculate the conditional probability P(ARG|A) using the information provided.

$$P(ARG|A) = \frac{P(ARG \cap A)}{P(A)} = 0.6$$

c. Calculate the conditional probability P(CRO|B) using the information provided.

$$P(CRO|B) = \frac{P(CRO \cap B)}{P(B)} = 0.4$$

Problem 3

Are the probabilities of Argentina winning a match and the match being played in the evening independent?

Given:

- P(A) = 0.65 (probability of Argentina winning a match)
- P(B) = 0.40 (probability of a match being played in the evening)
- $P(A \cap B) = 0.30$ (probability of Argentina winning a match played in the evening)

Using the formula for independent events to check,

$$P(A \cap B) = P(A) \times P(B)$$

$$0.30 = 0.65(0.40)$$

$$0.30 \neq 0.26$$

Since the equation is not true, it can be concluded that the events "Argentina winning a match" and "match being played in the evening" are not independent.

Problem 4

Consider the group stage of the 2022 FIFA World Cup, where four teams (England, Iran, USA, Wales) compete. The probability of each team advancing to the knockout stage is as follows:

P(England)=0.3, P(Iran)=0.2, P(USA)=0.4, P(Wales)=0.1

Additionally, the probability of a team winning a match in the knockout stage is given as follows:

P(Win|England)=0.5

P(Win|Iran)=0.6

P(Win|USA)=0.7

P(Win|Wales)=0.4

Write the expression for the law of total probability for the event W in terms of the events K (advancing to the knockout stage) and the conditional probabilities of winning given reaching the knockout stage.

$$P(W) = P(Win|England) \times P(England|K) + P(Win|Iran) \times P(Iran|K)$$
$$+ P(Win|USA) \times P(USA|K) + P(Win|Wales) \times P(Wales|K)$$

$$P(W) = (0.5 \times 0.3) + (0.6 \times 0.2) + (0.7 \times 0.4) + (0.4 \times 0.1)$$

$$P(W) = 0.59$$

Problem 5

Now, suppose a team has won their knockout stage match. What is the probability that this team is England? Formulate the problem using Bayes' Rule and then solve it.

A: England wins the knockout stage match.

B: A team wins their knockout stage match.

Bayes' Rule:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

Therefore,

$$P(England|Win) = \frac{P(Win|England) \times P(England)}{P(Win)}$$

To calculate probability,

$$P(Win) = (0.5 \times 0.3) + (0.6 \times 0.2) + (0.7 \times 0.4) + (0.4 \times 0.1)$$

$$P(Win) = 0.15 + 0.12 + 0.28_{0.04} = 0.59$$

The probability that England wins their knockout stage match given that a team wins a knockout stage match is 25.4%.

Problem 6

In the 2022 FIFA World Cup, Portugal played a total of 7 matches in the group stage. The probability of Portugal winning any individual match is 0.6.

What is the probability that Portugal wins at least 5 matches in the group stage?

- n = 7 (# of matches played in group stage)
- p = 0.6 (probability of Portugal winning a match)
- x = # of matches won by Portugal

Using the binomial distribution formula,

$$P(X \geq 5) = P(X = 5) + P(X = 6) + P(X = 7)$$

$$P(Y = 5) = \binom{7}{5}(0.6^5)(1 - 0.6)^{7-5} = 0.20$$

$$P(Y = 6) = \binom{7}{6}(0.6^6)(1 - 0.6)^{7-6} = 0.13$$

$$P(Y = 7) = \binom{7}{7}(0.6^7)(1 - 0.6)^{7-7} = 0.02$$

$$P(X \geq 5) = 0.20 + 0.13 + 0.02 = 0.35$$

Problem 7

In the 2022 FIFA World Cup, there are 32 teams in total, including Portugal (P), Argentina (A), and France (F). Let's assume that a draw is conducted, randomly selecting 8 teams to form Group A. The hypergeometric distribution can be used to model the probability of specific outcomes based on this draw.

Define the following events:

K: The number of selected teams in Group A that are from Portugal.

N: The total number of teams in the draw (32).

n: The number of teams selected in the draw to form Group A (8).

k: The specific number of teams from Portugal in Group A.

a. Write the hypergeometric probability mass function (PMF) for K in terms of N, n, and k.

$$P(K = k) = \frac{\binom{K}{k} \times \binom{N - K}{n - k}}{\binom{N}{n}}$$

b. Calculate the probability that exactly 3 teams from Portugal are selected in Group A

$$P(K = 3) = \frac{\binom{3}{3} \times \binom{29}{5}}{\binom{32}{8}} = 0.011$$

Problem 8

In the 2022 FIFA World Cup, consider a scenario where a fan wants to create a personalized playlist to represent the group stage matches. The fan selects 5 specific teams: Portugal (P), Argentina (A), France (F), Brazil (B), and Germany (G).

How many different ways can the fan arrange these 5 teams in a sequence to represent the order of matches?

$$P(5) = 5! = 120$$

Problem 9

In the 2022 FIFA World Cup, a group of friends plans to attend three matches together. The matches involve five teams: Portugal (P), Argentina (A), France (F), Brazil (B), and Germany (G). The friends want to select two teams to support in each match.

How many different ways can the group of friends select two teams out of the five (P, A, F, B, G) to support in the first match?

$$C(5,2) = \frac{5!}{2!\,(5-2)!} = 10$$

Problem 10

Suppose we take 10 matches from the dataset. Let's define a random variable X as the number of matches where Brazil scored 3 or more goals. Based on historical data, the probability distribution for X is as follows:

$$P(0) = P(X = 0) = 0.3$$

$$P(1) = P(X = 1) = 0.4$$

$$P(2) = P(X = 2) = 0.2$$

$$P(3) = P(X = 3) = 0.1$$

Problem 11

Leading up to the 2022 FIFA World Cup final, Kylian Mbappe, the star striker for France, had score 5 goals in 6 matches. Mbappe went on to score a hattrick in a historical final.

What was the probability that Mbappe scored exactly 3 goals in the final match?

To solve this problem using the Poisson distribution, we'll use the formula:

$$P(X = x) = \frac{e^{-\lambda} \cdot \lambda^{k}}{k!}$$

where:

- P(X=x) is the probability of scoring x goals
- ↘ is the average number of goals scored per match

Given that Mbappe scores an average of 0.83 goals per match ($\lambda = 0.83$), and we want to find the probability of him scoring exactly 3 goals ($k = 3$), these values can be plugged into the formula:

$$P(X = 3) = \frac{e^{-0.83} \cdot 0.83^3}{3!} \approx 0.04$$

As we all witnessed, Mbappe shocked the world with his phenomenal performance in the final against Argentina. Just when we thought the game was pretty much over, with Argentina winning 2-0, Mbappe put up one of the greatest comebacks in world cup history. In fact, he ended up forcing the game to extra time, where he completed his hat trick. Although France ended up losing to Argentina in a penalty shootout, no one expected Mbappe to almost flip the game. When you look at it from a statistical perspective, this becomes even crazier, as Mbappe only had a 4% chance of scoring a hattrick in the final match.

Question 12

In the group stage of the 2022 FIFA World Cup, Japan scored an average of 1.8 goals per match with a standard deviation of 0.9 goals.

Using Chebyshev's inequality, determine the minimum percentage of matches in which the number of goals scored by either Japan deviates from their respective means by at least 1 goal.

$$k_b = \frac{d}{\sigma} = \frac{1}{9} \approx 1.111$$

Percentage: $1 - \frac{1}{k^2} = 1 - \frac{1}{2^2} = 0.75$

Therefore, the minimum percentage of matches in which the number of goals scored by Japan deviates from their mean by at least 1 goal is 75%.

Question 13

In the 2022 FIFA World Cup, the average number of shots on target by Croatia per match was 10, with a minimum of 7 shots and a maximum of 15 shots.

Using the uniform distribution, calculate the probability of Croatia having exactly 9 shots on target in a match.

- Min (a) = 7 shots
- Max (b) = 15 shots
- Total interval (c to d) = 7 to 15 shots

The probability density function (PDF) for a uniform distribution is given by:

$$f(x) = \frac{1}{b-a}, for\ a \leq x \leq b$$

Therefore, the probability of Croatia having exactly 9 shots on target is:

$$P(9) = \frac{1}{15-7} = \frac{1}{8} = 0.125$$

This result shows that the probability of Croatia having exactly 9 shots on target in a match is 12.5%.

Question 14

Let's consider a total of n group stage matches played by each team in Group B: England, USA, Iran, and Wales.

a. $P(X = 1, Y = 2)$: Probability that England wins 1 match and USA wins 2 matches

- Other teams win the remaining matches $((4-1-2) \times n$ ways
- Total # of outcomes: $n \times n \times (4-1-2) \times n = n^3$

b. $P(X=1, Y=2)$: Probability that England wins 0 matches and USA wins 3 matches

- Other teams win the remaining matches $((4-0-3) \times n$ ways
- Total # of outcomes: $n \times n \times (4-0-2) \times n = n^3$

c. $P(X=1, Y=2)$: Probability that England wins 2 matches and USA wins 1 match

- Other teams win the remaining matches $((4-2-1) \times n$ ways
- Total # of outcomes: $n \times n \times (4-2-1) \times n = n^3$

d. $P(X=1, Y=2)$: Probability that England wins 3 matches and USA wins 0 matches

- Other teams win the remaining matches $((4-3-0) \times n$ ways
- Total # of outcomes: $n \times n \times (4-3-0) \times n = n^3$

Total # of possible outcomes = : $n^3 + n^3 + n^3 + n^3 = 4n^3$

Therefore, the probabilities are:

a. $P(X=1, Y=2) = \frac{n^3}{4^3} = \frac{1}{4}$

b. $P(X=0, Y=3) = \frac{n^3}{4^3} = \frac{1}{4}$

c. $P(X=2, Y=1) = \frac{n^3}{4^3} = \frac{1}{4}$

d. $P(X=3, Y=0) = \frac{n^3}{4^3} = \frac{1}{4}$

So, the joint probability function for X and Y is:

$$P(X,Y) = \begin{cases} \frac{1}{4} \; if \; (X,Y) is \; one \; of \; the \; scenarios \\ 0 \; otherwise \end{cases}$$

Conclusion

Utilizing an extensive dataset that includes every statistical aspect of the 2022 FIFA World Cup matches has been the main goal of this paper. The objective was to create a set of statistics problems that provide useful insights into the tournament dynamics while also testing theoretical ideas through a thorough analysis of this dataset. The information in the dataset was used to develop a wide range of questions involving different statistical approaches and methodologies. Each challenge, which ranged from probability distributions to combinations and poison distribution, was designed to highlight the variety statistical applications in relation to one of the most renowned athletic events in the world, while covering all chapters that were reviewed in class.

Advanced models like Elo ratings provided useful insights into team strengths and anticipated results through statistical projections. Even though these models provide a foundation for comprehending the dynamics of the competition, surprises were always possible in soccer due to its unpredictable character. Statistics could never fully capture the charm that makes soccer the most popular sport in the world, as witnessed by the incredible comebacks and exciting upsets that have occurred on the global scene. In-depth team strategies, player performances, and important statistical indicators that influenced the outcome of games were all examined in performance analysis. Data points ranging from possession statistics to goal differentials provided valuable insights into the tactics used by teams and the pivotal moments that turned games in their favor.

Essentially, the goal of this paper was to show how statistics can be used to solve puzzles related to sports and capture the spirit of the beautiful game, as shown by the 2022 FIFA World Cup. Each squad brought its unique style to the tournament, captivating audiences and leaving an

incredible mark on history. As supporters reflect on the moments made during those incredible weeks in Qatar, they can be proud that statistical data was a key factor in solving the mysteries surrounding their favorite matches from the tournament. By working through these problems, readers are encouraged to learn and discover, where statistics serve as a window into the thrilling drama, strategic brilliance, and utter excitement of international soccer on the biggest stage.

Work Cited

Iron486. (2022, December 18). *FIFA World Cup 2022: Complete dataset*. Kaggle. https://www.kaggle.com/datasets/die9origephit/fifa-world-cup-2022-complete-dataset

FOX Sports. (n.d.). *FIFA Men's World Cup Stats & Leaders*. FIFA Men's World Cup Stats & Leaders | FOX Sports. https://www.foxsports.com/soccer/fifa-world-cup-men/stats