POISSON DISTRIBUTION IN FOOTBALL BETTING PREDICTIONS

Mentor Shevroja

Department of Applied Mathematics, Faculty of Natural Science, University of Tirana, Albania

e-mail: mentor.shevroja@gmail.com

Abstract: Poisson distribution, coupled with historical data and numerical methods, can provide a method for calculating the likely number of goals that will be scored in a football match. The Poisson approach uses a calculated expected number of goals for a match as the mean in a Poisson distribution, which forms a probability distribution over the number of goals.
In this work we will treat a practical model based on Poisson distribution used by bookmakers for defining odds of football match and predicting the match winner. The model is based on defining Poisson distribution based on attack and defense strengths of previews matches for each team involved in match. Using Possion distribution of each team we will define model for correct score of the match and from there we can define other related odds of the match. To demonstrate the practical side of the above model we will demonstrate to predict the winner between Manchester United vs West Ham United on 5 December 2015. Studies of predicting odds models not only for football has a very special importance for betting industry. In the end, we compare the odds of our model with other betting companies, as well as look at some of the limitations of this approach. The Poisson distribution model for generating outcomes of the match for many sports is used in betting industry.

Key words: *Poisson distribution, odds, predict, attack and defense strength.*

Introduction

The goal of this paper is to create a prediction model which can assess the probabilities of the number of goals scored in a soccer match as well as the bookmakers. In order to create such a model and strategy, one must fully understand the mathematics behind sports betting odds and the mechanisms which influence them.

For the remainder of this report, the term betting industry refers to the market of bookmakers with focus on gambling on sporting events, while the term betting is sports gambling. The betting industry rests on the conflicting interest of bookmakers and customers wanting to earn money, respectively. Both sides will look for ways to improve their possibilities of achieving their goals, which raises a motivation of investigating how this could be done.

First part gives an introduction to sports betting, and introduces basic concepts of betting which are necessary for understanding the remainder of the report.

The second part is an introduction of Poisson distribution model for assessing number of goals scored in a football match. The statistics are examined and Chi-Squared test is done that supports that number of goals scored in a match by participating teams follows Poisson distribution.

And the last part of the report is shown how we can use past results to predict the next match between two teams in a tournaments. The proposed model is used to predict the match between Manchester United and West Ham United on December 5th, 2015.

Sports betting

A given sporting event or match has a finite number of outcomes. For a football match, for instance the number of possible outcomes is three, and can be one of *home*, *draw* or *away* if we want to predict the match winner. The are a lot of markets like match goals, double chance etc. For a given match with n possible outcomes,$outcome = 1, \dots, n$, the probability of the outcome i is $P(outcome = i)$. The outcomes are mentally exclusive, since a match cannot have two winners for example. So the following holds:

$$\sum_{i=1}^{n} P(outcome = i) = 1 \tag{1}$$

Bookmakers are basically companies trying to make money through sports wagers. Therefore they operate with a theoretical payback percentage when offering odds to their customers. The theoretical payback is set by the bookmaker and is the percentage of the turnover on a betting event which is expected to be paid back to the customers. The payback is less than 100%, normally around 90-95%. The higher the percentage, the lower the margin of theoretical profit for the bookmaker and the higher the odds. An odds for an outcome, can be calculated as:

$$Odds_i = tpb \ \frac{1}{P(outcome = i)} \tag{2}$$

where $Odds_i$, is the odds for outcome i, and tpb is the bookmaker's theoretical payback or payout. An odds calculated with payback of 1 (100%) is called a fair odds, since there is no theoretical advantage.

For football match with the outcomes home, draw and away, with a probability distribution of 60%, 25% and 15% respectively the odds can be calculated. Using the equation (2) with a theoretical payback of 100% the odds for respective outcomes would be 1.67, 4.00 and 6.67. If instead a payback of 92% is used in practice, the odds would be 1.53, 3.68 and 6.13.

People do not understand probabilities but they knows very well odds. In general, odds are likelihood that an event will happen. For example, the odds that today is your birthdays is 365/1. In sports betting, odds are generally refer to the price that you will be getting on your wagers. So for every $100 wagered on the odds 1.53 we will get 1.53 x $100 = $153 if the match ends in home win, and if the match ends in draw or away win you will lose the money of $100 placed.

Below we will try to create a model for the odds and explain why companies choose 1.53 instead of other higher or lower odds for a specific match.

Poisson distribution

The Poisson distribution is a discrete probability distribution, known from probability theory and statistics. It expresses the probability of a number events occurring within a fixed period of time.

The probability mass function of Poisson distribution is given by:

$$f(X = k;\ \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}, k = 0, 1, 2, \dots, \lambda > 0 \quad (3)$$

Where e is Euler's number ($e = 2.71828$) and $k!$ is the factorial of k.

The positive real number λ is equal to the expected value of X and also to its variance $\lambda = E(X) = var(X)$.

Poisson distribution is a mathematical concept for translating mean averages into a probability for variable outcomes. For example, Chelsea might average 1.7 goals per game. Entering this information into a Poisson formula would show that this average equates to Chelsea scoring 0 goals 18.3% of the time, 1 goal 31% of the time, 2 goals 26.4% of the time and 3 goals 15% of the time.

The Poisson approach is a naive prediction model, which also uses past results to assess the probability of the number of goals in a match. This model assumes, that the number of goals in a soccer match follows a Poisson distribution. For a given match, the two participating teams taken into account, the average number of goals, Avr, in prior matches are calculated. A prediction of the probability distribution of the number of goals is then made, based on a Poisson distribution with mean value Avr.

Dixon and Coles [DC97] proposes a model for assessing probabilities of soccer match base only on the number of goals scored in previews matches by two participating teams with average number of scored by to team calculated from the formula:

$$\lambda_h = \alpha_h \cdot \beta_a \cdot \bar{G}_h,\ \lambda_a = \alpha_a \cdot \beta_h \cdot \bar{G}_a \tag{4}$$

where λ_h and$\lambda_a$are excepted number of goals for home team and away team, α_h and α_a are home and away offensive/attacking strengths on previews matches, β_h andβ_a are home and away defensive strengths on previews matches, $\bar{G}_h$ and$\bar{G}_a$ are average goals scored by home and away team in previews matches

In Poisson model it's assumed that X and Y are independent which gives us, that the probability of a match result is given by the product of the probability of the home team goals and the away team goals:

$$P(X = x,\ \mathrm{Y} = \mathrm{y}) = P(X = x; \lambda_h) \cdot P(Y = x; \lambda_a) = \frac{{\lambda_h}^x e^{-\lambda_h}}{x!} \frac{{\lambda_a}^y e^{-\lambda_a}}{y!}\ \ \mathrm{x}, \mathrm{y} \geq 0 \tag{5}$$

Using the formula (5) we can calculate probability of any possible score which is called correct score matrix probabilities of the match. If we have the probability matrix of any possible final score we can predict the probabilities of the match winner. Before we can predict a match we have to make sure that number of goals scored in a match follows Poisson distribution.

Statistics

To establish the plausibility of using a Poisson distribution of the expected number of goals, the result data set is examined. If this approach is plausible, then for a set of soccer matches, the distribution of the matches over the number of goals must follow the Poisson distribution with a mean value equal to the average number of goals scored per match in the set of matches. The result data set is chosen for 5 European Tournaments in last season 2014/2015 (5ET 2014/15) which are English Premier League, German Bundesliga, Italian Serie A, France League 1 and Spanish Primera Division. Only regular time result is considered in the data, overtime is not included. Result data set consist of 1752 matches played and 4552 goals scored in 5ET 2014/15. The results are shown in table 1 as below:

Scored Goals	0	1	2	3	4	5	6	7+
No. Matches	151	322	437	364	260	133	55	30
Actual prob (%)	8.62	18.38	24.94	20.78	14.84	7.59	3.14	1.71
Poisson prob (%)	7.43	19.31	25.10	21.76	14.14	7.36	3.19	1.72

Table 1: Actual probabilities and Poisson probabilities for 5ET 2014/2015.

We can see that 151 matches ended with 0 goal, and 322 matches ended with 1 goal and so on up to 7+ goals. If we convert the occurrences to probability by dividing by total number of matches 1752 we get 8.62% chance match finish with 0 goal and 18.38% of the matches ended with one goal and so on.

In examined data there are 4552 goals scored which gives 4552/1752 = 2.598 goals per match. Using 2.60 goals per match we get the values of Poisson distribution in table 1. The results of table 1 is plotted in figure 1 below:

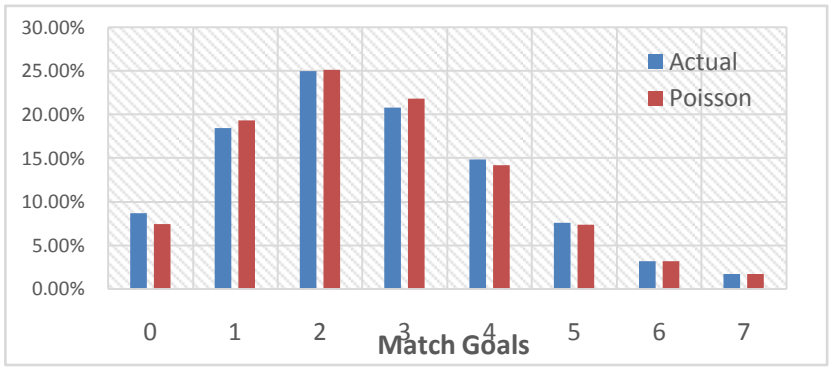

Figure 1: Plot of histogram of actual match goals distribution vs Poisson distribution probabilities for 5ET 2014/15.

Figure 1 shows actual distribution vs Poisson distribution probabilities of the table 1. The x-axis is the number of goals scored in 1752 matches, while the y-axis is the probability of matches over the number of goals for the T5ET 2014/15 season. The red columns show the Poisson distribution with mean value 2.80, which is the average number of goals scored per match in the 2006/07 season.

In the figure 2 are shown histograms for expected vs actual probabilities of the 5ET 2014/15 for home team and away team separated:

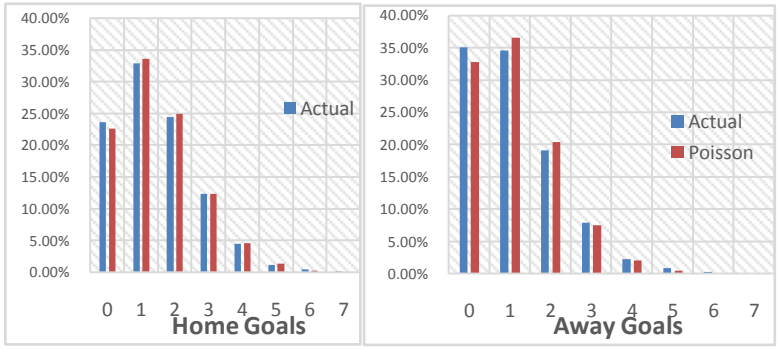

Figure 2. Plot of histogram of actual match goals distribution vs Poisson distribution probabilities for 5ET 2014/15 for home team (left) and away team (right).

We can see from the figure 1 and 2 values from expected probabilities are very closed to actual probabilities, the difference is small. Visualization is not enough about the assumption of scored goals in a match follows Poisson distribution.

Chi-square χ^2 goodness of fit test is applied to determine whether there is a significant difference between the expected frequencies and the observed frequencies.

At default significance level, $\alpha = 0.05$, degrees of freedom $df = 6$, critical region is $\chi^2 > 12.592$. Value 10.2854 does not lie in the critical region. There is no evidence at the 5% significance level, to suggest that the number of goals scored in football match does not follow Poisson distribution.

```
>>bins = 0:7;
>>obsCounts = [151, 322, 437, 364, 260, 133, 55, 30];
>>n = sum(obsCounts);
>>pd = fitdist(bins','Poisson','Frequency',obsCounts');
>>expCounts = n * pdf(pd,bins);
>>[h,p, st] = chi2gof(bins,'Ctrs',bins, 'Frequency',obsCounts,
'Expected',expCounts, 'NParams', 1)
h =
     0
p =
    0.1131
st =
    chi2stat: 10.2854
          df: 6
       edges: [-0.5000 0.5000 1.5000 2.5000 3.5000 4.5000 5.5000
6.5000 7.5000]
           O: [151 322 437 364 260 133 55 30]
           E:  [131.7867  340.9755  441.1078  380.4303  246.0746
127.3352 54.9097 20.2956]
```

Matlab code for testing hypothesis that number of goals for scored in a match follows Poisson distribution. The returned value *h=0* indicates that *chi2got* does not reject the null hypothesis at the default 5% significance level. The vector *E* contains the expected counts for each bin under the null hypothesis, and *O* contains the observed counts for each bin.

Visualization and analysis says that there is a strong resemblance of the two distributions (expected and Poisson distribution) for the 5ET 2014/15. The use of the expected number of goals as the mean of a Poisson distribution therefore seems to be a good approximation to the distribution of goals in soccer matches, and therefore can be viewed as a candidate for assessing the number of goals in a given soccer match.

Case Study

Before we can use Poisson to calculate the likely outcome of a match, we need to calculate the average number of goals each team is likely to score in that match. This can be calculated determining an "Attack" and "Defense Strength" for each team and comparing them.

Selecting a representative data range is vital when calculating Attack and Defense strengths – too long and the data will not be relevant for the teams current strength, while too short may allow outliers to skew the data. For this analysis we're using the 52 games played by each team in the English Premier League 2014/15 season including current season. The table 2 shows the final position in the league and the goals score and conceded by each team for 26 matches played at home and 26 matches played at away field.

#	Team	Played	Home		Away	
			Scored	Conceded	Scored	Conceded
1	Man City	26	53	30	29	32
2	Leicester	26	33	36	29	36
3	Man Utd	26	53	15	50	33
4	Arsenal	26	49	18	46	30
5	Tottenham	26	41	27	29	33
6	Liverpool	26	36	31	28	39
7	Crystal Palace	26	32	36	34	29
8	West Ham	26	37	29	32	39
9	Everton	26	43	32	32	37
10	Southampton	26	35	34	27	56
11	Watford	26	35	30	38	39
12	Stoke	26	42	20	23	27
13	West Bromwich	26	33	39	18	30
14	Chelsea	26	46	19	44	36
15	Swansea	26	24	36	21	36
16	Norwich	26	34	36	23	52
17	Sunderland	26	41	30	23	41
18	Bournemouth	26	22	33	23	50
19	Newcastle	26	52	27	24	40
20	Aston Villa	26	23	35	20	49
	Total	520	764	593	593	764

Table 2: Goals scored for each team in English Premier League season 2014/15 and current season. In total there are 520 matches with 1357 goals scored. In average 2.6096 goals was scored per match.

Results in the tables shows that was 764/520 at home and 454/520 away, equaling an average $\bar{G}_h = 1.47$goals per match at home and $\bar{G}_a = 1.14$ goals per match at away. The difference from average is what constitutes a team's Attack Strength. We'll also need the average number of goals a team concedes. This is simply the inverse of the above numbers (as the number of goals a home team scores will equal the same number that an away team concedes). Average number of goals conceded at home is 1.14 and average number of goals conceded away team from home is 1.47.

Using the numbers above and table we can calculate Attack and Defense Strengths for any match in English Premier League. The match we choose is Manchester United vs West Ham for their event on December 5th, 2015.

Predicting Manchester United's Goals

Calculate Manchester's attack strengthα_h:

1. Take the number of goals scored at home by Manchester 53 and divide by the number of home games (53/26): 2.04
2. Divide this value by average home goals scored per game (2.04/1.47), to get the attack strength $\alpha_h = 1.39$. This shows that Manchester scored 39% more goals at home than a hypothetical "average" Premier League side.

Calculate West Ham's defense strength β_a:

1. Take the number of goals conceded away by West Ham 39 and divide by the number of away games (39/26): 1.50.
2. Divide this by the season's average goals conceded by an away team per game (1.50/1.47) to get the "Defense Strength" $\beta_a = 1.02$. This therefore highlights West Ham conceded 2% more goals than an "average" Premier League side on the road.

Using equation (4) we can calculate the likely number of goals the home team might score:

$$\lambda_h = \alpha_h \cdot \beta_a \cdot \bar{G}_h = 1.39 \cdot 1.02 \cdot 1.47 = 2.08$$

Predicting West Ham's Goals

Calculate West Ham's attack strengthα_a:

1. Take the number of goals scored at away by West Ham 32 and divide by the number of home games (32/26): 1.23
2. Divide this value by average away goals scored per game (1.23/1.14), to get the attack strength $\alpha_a = 1.08$. This shows that West Ham scored 8% more goals at away than a hypothetical "average" Premier League side.

Calculate Manchester United's defense strength β_h:

1. Take the number of goals conceded home by Manchester United 15 and divide by the number of away games (15/26): 0.58.
2. Divide this by the season's average goals conceded by an away team per game (0.58/1.14) to get the defense strength $\beta_h = 0.51$. This therefore highlights West Ham conceded 49% fewer goals than an "average" Premier League side.

Using equation (4) we can calculate the likely number of goals the away team might score:

$$\lambda_a = \alpha_a \cdot \beta_h \cdot \bar{G}_a = 1.08 \cdot 0.58 \cdot 1.14 = 0.62$$

Of course, no match ends 2.08 vs 0.62, this is simply the average. Equation (3) allow us to distribute 100 of probability across a range of goal outcomes for each side. Using the equation (5) with assumption that scoring a goal by teams are independent we can calculate each possible final score probability by multiply each home goals probability by each away goal probability. In table 2 are shown results:

		λ_a	West Ham Goals							
		0.62	0	1	2	3	4	5	6	7+
λ_h	2.08		**53.7**	**33.3**	**10.3**	**2.14**	**0.33**	**0.04**	**0.00**	**0.00**
Manchester Goals	0	**12.4**	6.72	4.17	1.29	0.27	0.04	0.01	0.00	0.00
	1	**25.9**	13.9	8.67	2.69	0.56	0.09	0.01	0.00	0.00
	2	**27.0**	14.5	9.01	2.79	0.58	0.09	0.01	0.00	0.00
	3	**18.7**	10.1	6.25	1.94	0.40	0.06	0.01	0.00	0.00
	4	**9.74**	5.24	3.25	1.01	0.21	0.03	0.00	0.00	0.00
	5	**4.05**	2.18	1.35	0.42	0.09	0.01	0.00	0.00	0.00
	6	**1.41**	0.76	0.47	0.15	0.03	0.00	0.00	0.00	0.00
	7+	**0.56**	0.30	0.19	0.06	0.01	0.00	0.00	0.00	0.00

Table 3: Probabilities of goal outcomes for each side and final score probabilities of the match.

The table 3 shows that there is a 12.49% chance that Manchester United will not score, 25.99% chance to score one and 27.02% chance to score two. West Ham, on the other hand, are at 23.79% not to score, 33.35% to score one and 10.34% to score two. Under the assumption that both score are independent, we can multiply the two probabilities together and we will get probability of 2-0 outcome 14.4% chance.

Using the table 3 we can predict the match winner in terms of probability. If we add up the probability of all results where home team wins (e.g. 1-0, 2-0, 2-1, 3-2 etc.) than we will have the overall likelihood of a home win.

$$P(home\ win) = \sum_{x>y} P(X = x,\ Y = y) = 71.51\%$$

$$P(draw) = \sum_{x=y} P(X = x, Y = y), = 18.62\% \; P(away\ win) = \sum_{x>y} P(X = x, Y = y) = 9.87\%$$

The probability that Manchester win the match is 71.51%, the match can end in draw with 18.62% chance and the West Ham wins the match with 9.87% chance. We can convert the probability into decimal odds using equation (2) and we will get the odds 1.40 for Manchester, 5.37 for draw and 10.13 for West Ham with 100% payback.

The table 3 is all we need to generate also other types of markets. Some of the most know markets are shown in the below for our match prediction between Manchester and West Ham including match winner:

Market	Probability	Odds
Home Win	71.85%	1.40
Draw	18.62%	5.37
Away Win	9.87%	10.13
Over 2.5 Goals	50.64%	1.97
Under 2.5 Goals	49.36%	2.03
Both Teams to Score (Yes)	75.13%	1.33
Both Teams to Score (No)	24.87%	4.02

Table 4. Probability and odds most known markets

In table 5 there are shown our odds compared with some of the biggest bookmakers.

Bookmaker/Providers	Manchester Win	Draw	West Ham Win	Payback
Poisson Distribution	1.40	5.37	10.13	100.00%
Bet365.com	1.40	4.50	7.5	93.47%
Sbobet.com	1.44	4.20	7.20	93.33%
Williamhill.com	1.44	4.33	7.50	94.45%
Igcbet.com	1.43	4.20	7.20	92.91%
Betradar.com	1.46	4.55	7.38	96.13%

Table 5. Odds from our model compare to other bookies online.

Conclusions

Having implemented and evaluated the Poisson distribution assessor for predicting football match outcomes. Statistical analytics says that Poisson distribution is well studied for assessing number of goals scored in a match and also for estimating odds in betting industry.

The model uses past data to predict future results. Poisson distribution is a simple predictive model and easy to use. But it has some limitation that doesn't allow for a lot of factors. Situational factors such as club circumstances, game status and subjective evaluation of the change of each team during the transfer window are completely ignored. The accuracy of this Poisson model is open to debate. Does something that happened 6 months ago with different players in different weather conditions really help us understand what will happen?

Correlations are also ignored; such as the widely recognized pitch affect that shows certain matches have a tendency to be either high or low scoring.

These are particularly important areas in lower league games, which can give punters an edge against bookmakers, while it's harder to gain an edge in major leagues, given the expertise that modern bookmakers.

Poisson model suggest that Manchester United will win the match against West Ham United with 71.51% chance. The match will ends in draw by 18.62% chance and the victory for West Ham is 9.87% chance. The prediction is what the past says but the future is full of surprises, if we can predict the match it won't be interested to watch it.

Poisson distribution can be used also on other sports like Ice-Hockey, Basketball to predict the match winner and estimate odds for betting industry.

Reference

[Bet] http://www.pinnaclesports.com.

[IGC] http://www.igcbet.com.

[CH] Tobias Christensen and Rasmus Hansen. Odds assessment on football matches. Master's thesis, Aalborg University.

[DC97] Mark J. Dixon and Stuart G. Coles. Modelling association football scores and inefficiencies in the football betting market. *Journal of Royal Statistical Society series, Series C. Vol. 46, No. 2, p. 265-280, 1997.*