Are Goals Poisson Distributed?

Georgi Boshnakov[1], Tarak Kharrat[1,2], and Ian G. McHale[2]

[1]School of Mathematics, University of Manchester, UK.
[2]Centre for Sports Business, Salford Business School, University of Salford, UK.

> *All models are wrong but some are useful.*
>
> George Box, *1976*

George Box's statement is perhaps one of the most famous quotes ever made by a statistician, and is something that is often forgotten. Bearing it in mind, let us look at a question that has no doubt crossed many quants minds when modelling football scores: "are goals actually Poisson?". In this article, we provide our own slant on the debate and show how we try to solve the issue...

1 Download the data

We will be using results data from the wonderful resource that is football-data.co.uk. We have five seasons of data from the 2010-11 season through to the 2014-15 season. The data are stored in an `R` data frame called `theData` and we provide the code used below.

```
> getDataFromFootballData <- function(season = "1415") {
+     temp <- read.csv(paste0("http://www.football-data.co.uk/mmz4281/",
+                             season, "/E0.csv"))
+     subset(temp,
+            select = c("Date", "HomeTeam", "AwayTeam", "FTHG", "FTAG")
+            )
+ }
> Seasons <- paste0(10:14, 11:15)
> theData <- data.frame()
> for (i in seq(along = Seasons))
+     theData <- rbind(theData, getDataFromFootballData(Seasons[i]))
```

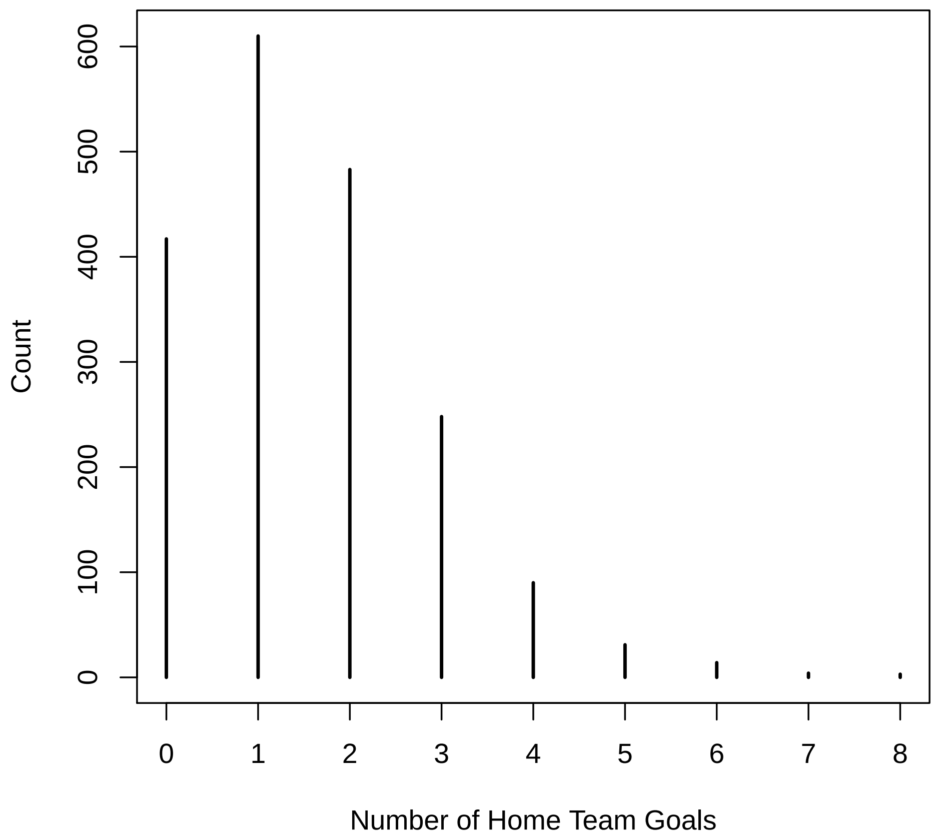
For convenience, we create variables `HG` (home goals) and `AG` (away goals):

```
> HG <- theData$FTHG
> AG <- theData$FTAG
```

2 Explore the data

First, let us start by plotting a bar chart of goals (let's concentrate on home goals for now).

```
> hg_freq <- table(HG)
> plot(hg_freq, xlab = 'Number of Home Team Goals', ylab = 'Count')
```



This looks as one would expect. The most common number of home team goals is 1 (occurring in about 600 of the 1900 games), and as the number of goals increases from there, the observed frequency decreases.

Goals scored by each team are a type of 'count data', and since goals can be considered to be relatively rare events, the natural distribution to use for modelling goals is the Poisson distribution. Indeed since as early as Moroney (1951), people have used the Poisson distribution as the basis of models for goals in football.

The Poisson distribution arises from a process (e.g. goal scoring) that is assumed to be homogeneous over time and the probability that a team will score x goals by time t is given by

$$\mathbb{P}(X(t) = x) = \frac{e^{-\lambda \cdot t}(\lambda \cdot t)^x}{x!}, \tag{1}$$

where λ is an intensity parameter which, here, is the rate at which the team scores goals (number of goals per unit time). It is convenient to rescale time so that the end of the match is at $t = 1$. With this convention $P(X(1) = x) = \lambda^x e^{-\lambda}/x!$ gives the probability that the team will score x goals in the match, and this is the familiar textbook definition of Poisson distribution.

The first thing students often learn about the Poisson distribution is that the mean of the distribution (the expected number of goals per match) is equal to the variance. If we are to use the Poisson distribution for home goals, this should be true:

```
> c(Mean = mean(HG), Var = var(HG))

    Mean      Var
1.563684 1.703683
```

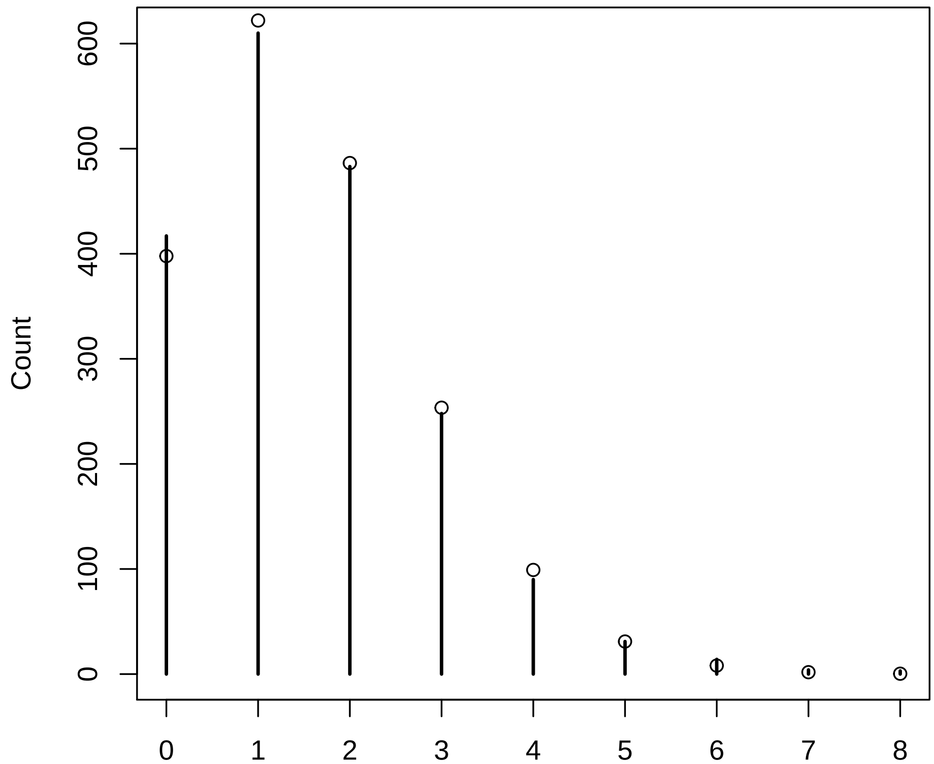
Maybe the mean is not equal to the variance here then — the variance is a little bigger than the mean, suggesting there is a little 'over-dispersion' in the data.

To actually fit the Poisson distribution to the data we need to know which value of λ to use. Fortunately, this could barely be any simpler and we just need the mean of the data as this is the *maximum likelihood estimate* of λ. For home goals we saw the mean was 1.56 goals per game on average, so $\hat{\lambda} = 1.56$.

To assess the goodness-of-fit, we need to compare the observed frequencies to the expected frequencies calculated for the fitted Poisson distribution, using formula (1) and the estimated value of λ. We add the expected frequencies to our bar plot as follows:

```
> hg_n <- length(HG)
> lambdaHat <- mean(HG)
> hg_count <- as.numeric(names(hg_freq))
> hg_E <- hg_n * dpois(hg_count, lambda = lambdaHat)
> hg_eo <- rbind(hg_freq, hg_E)
> row.names(hg_eo) <- c("Observed", "Expected")

> plot(hg_freq, xlab = 'Observed and Expected frequencies of Home Team Goals',
+      ylab = 'Count')
> points(hg_count, hg_E)
```



Observed and Expected frequencies of Home Team Goals

The fitted Poisson appears to be under-estimating the number of matches in which the home team scores 0 goals and over-estimating the number of matches in which the home team scores 1 goal. But on-the-whole, to the naked eye, this looks like a decent fit.

To test the goodness-of-fit more formally, we can use the chi-square goodness-of-fit test.

The chi-square test for the Poisson distribution fitted to home goals gives a test statistic of 13.23 with 5 degrees of freedom. The p-value is 0.02. Ouch, this means that despite the Poisson 'looking' like a decent fit, it is rejected as the distribution of home goals (at the 5% level of signicance). To check what our 'naked eyes' might have missed, let us look look at the actual values depicted in the previous plot:

```
> print(hg_eo, digits = 4)
```

```
             0   1     2     3     4     5      6     7      8
Observed 417.0 610 483.0 248.0 90.00 31.00 14.000 4.000 3.0000
Expected 397.8 622 486.3 253.5 99.09 30.99  8.076 1.804 0.3526
```

We notice that the difference between the observed and expected counts for $x = 0$, i.e. the number of games in which the gome team scores no goals, is 19.21, i.e. the observed frequency for $x = 0$ is about 4.8% larger than the one predicted by the model. Maybe this is not so small after all. Since the chi-squre test pulls together the discrepancies between what we expect and what is observed for a number of frequencies it is worth taking note of this result, even if we are not prepared to rely entirely on formal statistical inference.

For away goals, the p-value of the test was less than 0.001 and we confidently reject the Poisson model as the distribution of away goals, as well.

The seeds of doubt have firmly been planted...

3 Other models

What next? The negative binomial distribution relaxes the homogeneity assumption in the Poisson assumption by assuming that the intensity parameter is gamma distributed. This seems sensible as we are trying to model goals scored in matches in which each team is scoring goals at a possibly different rate. We will use this distribution later. But what other distributions are there?

In search of peace of mind, let's think about the process by which the Poisson distribution is generated. As with most models, we need to make some assumptions before we can declare that goals are a "textbook" case of a Poisson process. Here is an outline of a standard setup. Assume that we observe the counting process over a time interval of length τ and that the events occur completely at random, more specifically that the number of occurences in any time interval is proportional to the length of the interval and the numbers of occurences in non-intersecting time intervals are independent. The Poisson count probabilty is obtained by dividing the observation interval of length τ into tiny little intervals of length δ, such that the probability of observing k arrivals in each tiny interval is given by:

$$\mathbb{P}(k,\delta) \approx \begin{cases} 1-\lambda\delta & k=0 \\ \lambda\delta & k=1 \\ 0 & k>1 \end{cases} \tag{2}$$

and then summing up the contributions of all these tiny intervalls to obtain the probability $\mathbb{P}(k,\tau)$. The Poisson process can hence be seen as a sequence of Bernoulli trials over a large number of (tiny) intervals of length δ with a small probability of success $p = \lambda\delta$. It can be thought of as the continuous version of the binomial distribution.

This sounds like a football match — consider each possession a trial. An event occurs when a goal is scored. The probability to score a goal in any small time interval is surely small. So far, so good, but the seemingly abstract discussion above reveals possible problems. Indeed, the scoring probability is hardly constant during a match. This violation alone doesn't invalidate the Poisson model however, it can be accommodated by allowing the rate λ to vary during matches (Smith and Zandt, 2000). The critical assumption is that the success probability of each trial (possession) is independent from the last. Is this true for football? It possibly isn't. We see this all of the time as games open up as play gets more and more stretched as teams commit more and more players forwards after each (unsuccessful) possession.

Without modelling the possessions themselves, one can resort to trying other distributions. In a paper by McShane et al. (2008), a discrete distribution based on a Weibull (hence relaxing the time-constant intensity assumption) renewal process is derived. This effectively means that the increasing rates of scoring can be accomodated by the model. This makes sense as it means that the scoring rates (the probability of each possession being successful) can be allowed to increase during the match. The resulting distribution is called the 'Weibull count model' and its probability mass function is given by

$$\mathbb{P}(X(t) = x) = \sum_{j=x}^{\infty} \frac{(-1)^{x+j}(\lambda t^c)^j \alpha_j^x}{\Gamma(cj+1)}, \tag{3}$$

where $\alpha_j^0 = \Gamma(cj+1)/\Gamma(j+1), j = 0, 1, 2, \ldots,$ and $\alpha_j^{x+1} = \sum_{m=x}^{j-1} \alpha_m^x \Gamma(cj - cm + 1)/\Gamma(j - m + 1)$, for $x = 0, 1, 2, \ldots,$ for $j = x+1, x+2, x+3, \ldots$. In equation (3), λ is a 'rate' parameter and c is the 'shape' parameter of the distribution. Here, the observation unit is the match which we take as having a duration of 1 time unit. The rate, λ, is thus the scoring rate per match.

So far we have three candidate distributions: the Poisson, the negative binomial, and the Weibull count models. We can generate a fourth distribution by allowing the rate parameter in the Weibull count model to be a gamma random variable (in the same way the negative binomial distribution was derived from the Poisson distribution). We call this distribution the 'Weibull-gamma count' distribution.

We have written an `R` package that computes the various quantities used when dealing with these four distributions and it is available from us upon request. We now fit the Poisson, the negative binomial and the two Weibull count models to the goals data...

4 Comparing the four distributions

Within the R package, the four distributions are referred to as `"poissonC"`, `"negBinC"`, `"weibullC"` and `"weiGamC"` for the Poisson, the negative binomial, the Weibull count, and the Weibull-gamma count models respectively.

To fit one of the distributions to our data, we use the code:

```
> parWeibullC <- fitDist(theData, home =  TRUE, "weibullC")
```

where we set `home` equal to `TRUE` to fit the model to home goals.

The estimated coefficients can be viewed using:

```
> coef(parWeibullC)

   lambda        cc
1.5008982 0.9340265
```

To perform the chi-square goodness of fit tests we use the function `chisqGof()`. It performs tests for all four models and returns a list with five elements: the observed frequencies, and one element per distribution.

```
> gofHome <- chisqGof(theData, home = TRUE)
> gofAway <- chisqGof(theData, home = FALSE)
```

Tables 1 and 2 indicate that the Poisson model is inadequate for both the home and away goals. The three other distributions are OK, at least at the conventional 5% significance level and even the 10% significance level for the home goals.

Thinking about George Box's quote at the start of this article then: all four of these models are wrong, but the Poisson is the 'most' wrong! It is not realistic to expect particularly good models by

	Test statistic	degrees of freedom	p-value
Poisson	13.23	5	0.02
Negative Binomial	5.75	4	0.22
Weibull Count	7.65	4	0.11
Weibull-Gamma Count	4.77	3	0.19

Table 1: Results of chi-square goodness-of-fit tests for four distributions fitted to home goals.

	Test statistic	degrees of freedom	p-value
Poisson	23.49	5	0.00
Negative Binomial	8.53	4	0.07
Weibull Count	6.60	4	0.16
Weibull-Gamma Count	6.61	3	0.09

Table 2: Results of chi-square goodness-of-fit tests for four distributions fitted to away goals.

only using an overall distribution but the clear superiority of the Weibull-count models illustrated here and their ability to accommodate a relaxation of the homogeneity and time-constant scoring rate assumptions suggest using them as the basis of more sophisticated models.

In our next article, we are going to describe a Maher-type model (where each team has parameters representing its attack and defence abilities) for football scores based on the Weibull count model, and instead of using the chi-square test to examine the ‘goodness’ of the model, we are going to test the model properly and use it for betting.

References

McShane, B., Adrian, M., Bradlow, E. T., and Fader, P. S. (2008). Count models based on weibull interarrival times. *Journal of Business and Economic Statistics*, 26(3):369–378.

Moroney, M. J. (1951). *Facts from Figures.* Penguin.

Smith, P. L. and Zandt, T. (2000). Time-dependent poisson counter models of response latency in simple judgment. *British Journal of Mathematical and Statistical Psychology*, 53(2):293–315.