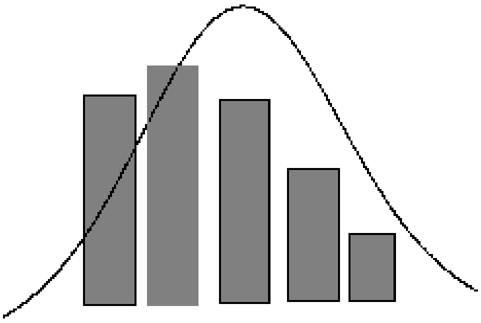
ATHENS UNIVERSITY OF
ECONOMICS AND BUSINESS

Distributions Based on Poisson Differences with Applications in Sports

Dimitris Karlis and Ioannis Ntzoufras
Department of Statistics
Athens University of Economics and Business

Technical Report No 101, April 2000



DEPARTMENT OF STATISTICS

TECHNICAL REPORT

Discrete Distributions with Applications in Sports

Dimitris Karlis [1] and Ioannis Ntzoufras
Department of Statistics
Athens University of Economics and Business
76, Patission Str., 10434, Athens, Greece

April 2000

Abstract

In this paper, the distributions of the difference and the absolute difference of two dependent Poisson variates are examined. It is shown that if two random variables jointly follow a bivariate Poisson distribution, then the probability distribution of their difference is the same as the distribution of the difference of two independent Poisson variables. The importance of this finding for sport modelling purposes is exploited. Additionally, the properties of these distributions as well as estimation methods are provided. Finally, applications and extensions of the derived distributions are also presented and discussed in detail.

Keywords: Absolute difference; Soccer; Bivariate Poisson distribution; Difference of random variables; EM algorithm; World Cup; Zero-inflated distributions.

1 Introduction

Sports is a blooming field for applying current and developing new statistical methodologies. In addition, betting on the outcome of soccer matches has a long tradition in United Kingdom and other European countries. Statistical models can be helpful tools for such purposes; see, for example, Jackson (1994).

Modelling the outcome of a soccer game is based on the number of goals scored by each team. Suppose that X and Y denote the number of goals scored by each team. The Poisson

[1] Corresponding author, email address: karlis@stat-athens.aueb.gr

distribution has been widely accepted as a simple and initial modelling approach for the distribution of the number of goals, despite the small overdispersion that has been found to exist for such data. Modelling simultaneously the number of goals of both teams, is a more sophisticated approach. Although, several authors (see, for example, Lee, 1997, Karlis and Ntzoufras, 1998) have shown a small dependence (but in fact not statistical significant) between the number of goals scored by the two opponents, this dependence has been ignored for modelling purposes.

Let us define the random variable $Z = X - Y$; then the first team wins a game if $Z > 0$, looses if $Z < 0$ and there is a draw between the two opponents if $Z = 0$. Keller (1994) calculated the probability of winning a game assuming independent Poisson distributions for both X and Y. Alternatively, we may consider a bivariate model, assuming that the two variables follow a bivariate Poisson distribution (see Kocherlakota and Kocherlakota, 1992 and references therein). The marginal distributions are simple Poisson distributions, while the random variables are not independent.

The aim of this paper is dual. Firstly, a result concerning the dependence of two Poisson random variables with respect to their difference is provided. It is found that, starting from a bivariate Poisson distribution, the distribution of the difference is not influenced by the dependence between the two variables. Secondly, the distributions of the difference and the absolute difference of two Poisson random variables are described. Illustrative applications to sports modelling are also provided.

The difference of two independent Poisson random variables has been discussed by Irwin (1937) for the case of equal means and Skellam (1946) for the case of different Poisson means. In this paper we examine further the properties and the behaviour of the distribution of the difference of two Poisson (dependent or not), as well as the distribution of the absolute difference of the two variables. The former distribution can be directly used for modelling soccer games and for calculating the probabilities of the game outcomes ('win', draw' or 'loss').

The remaining of the paper proceeds as follows. In section 2 the probability distribution of the difference Z is derived and its properties are discussed in detail. Section 3 deals with the distribution of the absolute difference of two Poisson distributed variables, while in Section 4 the importance of this new result for soccer modelling is exploited. Some extended models are described in Section 5. Finally, concluding remarks are given in Section 6.

2 The Distribution of the Bivariate Poisson Difference

2.1 The Probability Distribution

Suppose that the discrete random variables X, Y jointly follow the bivariate Poisson distribution. The probability function is given by

$$P_{X,Y}(x,y) = P(X=x, Y=y) = e^{-(\theta_1+\theta_2+\theta_3)} \frac{\theta_1^x}{x!} \frac{\theta_2^y}{y!} \sum_{i=0}^{min(x,y)} \begin{pmatrix} x \\ i \end{pmatrix} \begin{pmatrix} y \\ i \end{pmatrix} i! \left(\frac{\theta_3}{\theta_1 \theta_2} \right)^i . \quad (1)$$

The bivariate Poisson distribution defined in (1) allows for dependence between the two random variables. Marginally each random variable follows a Poisson distribution with parameters $\theta_1 + \theta_3$ and $\theta_2 + \theta_3$ respectively. Parameter θ_3 is the covariance between X and Y and hence it can be considered as a parameter that incorporates dependence between the two random variables. If $\theta_3 = 0$ then the two variables are independent and the bivariate Poisson distribution reduces to the product of two independent Poisson distributions. The marginal distributions of both X and Y are Poisson, while the conditional distribution of Y given X is the convolution of a Poisson with a binomial distribution (see Kocherlakota and Kocherlakota, 1992).

Define, further, the difference between X and Y as $Z = X - Y$. The distribution of the random variable Z is given by

$$\begin{aligned} P_Z(z) = P(Z=z) &= \sum_{r=0}^{\infty} P(X=r, Y=r-z) \\ &= \sum_{r=0}^{\infty} P(X=r) P(Y=r-z|X=r). \quad (2) \end{aligned}$$

If we substitute (2) in (1) we can derive the probability function of Z.

$$\begin{aligned} P(Z=z) &= e^{-(\theta_1+\theta_2+\theta_3)} \sum_{r=0}^{\infty} \frac{\theta_1^r}{r!} \frac{\theta_2^{r-z}}{(r-z)!} \sum_{i=0}^{min(r,r-z)} \begin{pmatrix} r \\ i \end{pmatrix} \begin{pmatrix} r-z \\ i \end{pmatrix} i! \left(\frac{\theta_3}{\theta_1\theta_2} \right)^i \\ &= e^{-(\theta_1+\theta_2+\theta_3)} \sum_{i=0}^{\infty} \sum_{r=i}^{\infty} \frac{\theta_1^{r-i} \theta_2^{r-z-i}}{(r-i)! i! (r-z-i)!} \theta_3^i . \end{aligned}$$

By setting $t = r - i$ we obtain that

$$\begin{aligned} P(Z = z) &= e^{-(\theta_1+\theta_2+\theta_3)} \sum_{i=0}^{\infty}\sum_{t=0}^{\infty} \frac{\theta_1^t \theta_2^{t-z}}{t!i!(t-z)!}\theta_3^i \\ &= e^{-(\theta_1+\theta_2+\theta_3)} \sum_{i=0}^{\infty} \frac{\theta_3^i}{i!} \sum_{t=0}^{\infty} \frac{\theta_1^t \theta_2^{t-z}}{t!(t-z)!} \\ &= e^{-(\theta_1+\theta_2+\theta_3)}\ e^{\theta_3} \sum_{t=0}^{\infty} \frac{\theta_1^t \theta_2^{t-z}}{t!(t-z)!} \\ &= e^{-(\theta_1+\theta_2)} \sum_{t=0}^{\infty} \frac{\theta_1^t \theta_2^{t-z}}{t!(t-z)!}. \end{aligned}$$

The sum in the right hand side is the Modified Bessel function (see Abramowitz and Stegun, 1974, pp. 375) defined by

$$I_r(x) = \left(\frac{1}{2}x\right)^r \sum_{k=0}^{\infty} \frac{\left(\frac{1}{4}x^2\right)^k}{k!\Gamma(r+k+1)}$$

and hence the probability distribution of the difference Z is given by

$$P_Z(z) = P(Z = z) = e^{-(\theta_1+\theta_2)} \left(\frac{\theta_1}{\theta_2}\right)^{z/2} I_z\left(2\sqrt{\theta_1\theta_2}\right) \tag{3}$$

for $z = \ldots -2, -1, 0, 1, 2, \ldots$, $\theta_1, \theta_2 > 0$. It is easily verified that the above probability mass function does not depend on θ_3.

The derived distribution is the same with the one derived from independent Poisson variates with parameters θ_1 and θ_2 respectively, by Skellam (1946). Some references about this distribution can be found in Johnson *et al.* (1992, pp.191). Note a misprint in Johnson *et al.* (1992) about the order of the Modified Bessel function.

For simplicity we will refer to this distribution as the Poisson Difference distribution and will be denoted by $PD(\theta_1, \theta_2)$.

2.2 Properties and Estimation for the Poisson Difference Distribution

2.2.1 Properties

The expected value of the above distribution is given by $E(Z) = \theta_1 - \theta_2$ while the variance is $Var(Z) = \theta_1 + \theta_2$. In general, the odd cummulants are equal to $\theta_1 - \theta_2$ while the even cummulants are equal to $\theta_1 + \theta_2$. The skewness coefficient $\beta_1 = \frac{\mu_3'}{\mu_2'^{3/2}}$, with μ_r' denoting the r-th central moment, is given by

$$\beta_1 = \frac{(\theta_1 - \theta_2)}{(\theta_1 + \theta_2)^{3/2}}$$

and hence the type of the skewness is determined by the sign of $\theta_1 - \theta_2$. Thus, if $\theta_1 > \theta_2$ then positive skewness is present. The distribution is symmetric only when $\theta_1 = \theta_2$ (the case discussed by Irwin, 1937).

Similarly the kurtosis coefficient $\beta_2 = \frac{\mu_4'}{\mu_2'^2}$ is given by

$$\beta_2 = \frac{1 + 3(\theta_1 + \theta_2)}{\theta_1 + \theta_2} = 3 + \frac{1}{\theta_1 + \theta_2}.$$

It is clear that as $\theta_1 + \theta_2$ tends to infinity the kurtosis coefficient tends to 3, the value of the normal distribution. Moreover, for a constant difference $\theta_1 - \theta_2$, the skewness coefficient also tends to zero, indicating that for large values of the $\theta_1 + \theta_2$ the distribution can be sufficiently approximated by the normal distribution.

Figure 1 depicts the probability distribution in comparison to the normal distribution along with a normal density with the same mean and variance. For small parameter values, the normal approximation is poor. However, as we can see in Figure 1, the Poisson difference distribution approaches normality rather quickly.

If the parameter θ_2 is very close to 0, then the distribution tends to a Poisson distribution. On the contrary if the parameter θ_1 approaches 0, then the distribution is the negative of a Poisson distribution (i.e. a Poisson distribution at the negative axis).

Figure 1 about here

An interesting property is a 'type' of symmetry given by

$$P(Z = z|\theta_1, \theta_2) = P(Z = -z|\theta_2, \theta_1).$$

This property can be useful for fast probability calculations. For practical purposes we may compute the probabilities, via the convolution formula given in (2) using independent Poisson variates. Alternatively, we may use the recursive relationships for the Modified Bessel functions given by Abramowitz and Stegun (1974).

Skellam's (1946) original derivation of the distribution, used the probability generating functions. He showed that the probability generating function $G(t)$ is given by

$$G(t) = exp\left(-(\theta_1 + \theta_2) + \theta_1 t + \theta_2 t^{-1}\right).$$

The unimodality of the Poisson difference distribution can be proved using the results of Keilson and Gerber (1971). According to them, the convolution of two strongly unimodal distributions is again a strongly unimodal distribution. The distribution in (3) is a convolution of a Poisson distribution with a 'negative' Poisson distribution. The Poisson distribution is clearly unimodal while the 'negative' Poisson distribution is also unimodal because it just reflects the simple Poisson distribution to the negative axis. Therefore, according to the argument of Keilson and Gerber (1971), the Poisson difference distribution is also unimodal.

The sum and the difference of two Poisson difference random variables also follow the same distribution as the following theorem shows.

Theorem 1: Let us consider two random variables say $Z_1 \sim PD(\theta_1, \theta_2)$ and $Z_2 \sim PD(\theta_3, \theta_4)$. Then the sum $S_1 = Z_1 + Z_2$ also is given by $S_1 \sim PD(\theta_1 + \theta_3, \theta_2 + \theta_4)$.

Proof: Since Z_1 and Z_2 can be written as $Z_1 = X_1 - X_2$ and $Z_2 = X_3 - X_4$ with $X_i \sim Poisson(\theta_i)$ for $i = 1, 2, 3, 4$, results to

$$S_1 = (X_1 - X_2) + (X_3 - X_4) = (X_1 + X_3) - (X_2 + X_4)$$

with $(X_1 + X_3) \sim Poisson(\theta_1 + \theta_3)$ and $(X_2 + X_4) \sim Poisson(\theta_2 + \theta_4)$. Similarly, the variable $S_2 = Z_1 - Z_2$ follows a Poisson difference distribution with parameters $\theta_1 + \theta_4$ and $\theta_2 + \theta_3$.

Remark If we consider a random sample of size n of i.i.d variables $Z_i, i =, 1..., n$ then we can straightforwardly show that

$$S_3 = \sum_{i=1}^{n} Z_i \sim PD(n\theta_1, n\theta_2).$$

2.3 Estimation

In this section we focus on the estimation of the parameters θ_1 and θ_2 of a Poisson difference distribution. Two approaches are presented: the moment method and the maximum likelihood estimation. The first is straightforward and is described briefly, while the second is more interesting and detailed discussion is provided.

The estimation of the parameters can be easily accomplished using the moment method. Matching the sample mean and variance with the theoretical values of section 2.2.1 results in

$$\hat{\theta}_1 = (\bar{z} + s_z^2)/2 \quad \text{and} \quad \hat{\theta}_2 = (s_z^2 - \bar{z})/2,$$

where $\bar{z}$ and s_z^2 are the sample mean and variance respectively. The moment estimates do not exist if $s_z^2 - |\bar{z}| < 0$. Asymptotic variances and covariances of the estimates can be derived in the usual manner.

Maximum Likelihood (ML) estimation is much more complicated, since the likelihood involves the modified Bessel function. If the true data X_i and Y_i were observed then the estimation is straightforward since their means would be the ML estimates for the Poisson parameters. Here an EM type algorithm is constructed, based on the missing data representation of the difference Z. In the EM algorithm, one calculates at the E-step the expectation of X_i and Y_i conditional on the data and the current values of the parameters and then, at the M-step, the means of the pseudodata X_i and Y_i are simply equated to the parameters. More formally the EM can be described as follows

E-step: With the current values of the parameters $\theta_1^{(k)}$ and $\theta_2^{(k)}$ from the k-th iteration, calculate the expected values of X_i and Y_i given the current values of the parameters. Using simple probability calculus is obtained that

$$\begin{aligned} t_i = E(X_i|z_i, \theta_1^{(k)}, \theta_2^{(k)}) &= \sum_{x=0}^{\infty} x P(X_i = x|Z_i = z_i) \\ &= \sum_{x=0}^{\infty} x \frac{P(X_i = x)P(Y = x - z_i)}{P(Z_i = z_i)} \\ &= \theta_1^{(k)} \frac{P_Z(z_i - 1)}{P_Z(z_i)}, \end{aligned}$$

where $P_Z(z)$ is defined in (3).

M-step: Update the estimates by $\theta_1^{(k+1)} = n^{-1} \sum_{i=1}^{n} t_i$ and $\theta_2^{(k+1)} = \theta_1^{(k+1)} - \bar{z}$

It is not necessary to calculate both conditional expectations, since second parameter can be updated by simple subtraction.

The above scheme is straightforward. It possesses all the properties of the standard EM algorithm (see, McLachlan and Krishnan, 1997). The global maximum was obtained for all the cases we used either with real or simulated data. However, starting from several starting points is useful to ensure that the global maximum has been obtained.

2.4 Mixture Distributions of Bivariate Poisson Variables

In this subsection we examine whether we can generalize some of the results of Poisson difference distribution in other discrete distributions and specifically in mixtures of the Bivariate Poisson distribution.

Suppose that X and Y conditional on the parameters follow the bivariate Poisson distribution as given by (1). Further assume that the parameters θ_1 and θ_2 are random variables that follow jointly a distribution function, say $g(\theta_1, \theta_2)$. Then the unconditional probability is given by

$$P(x, y|\theta_3) = \int \int P(x, y|\theta_1, \theta_2, \theta_3) g(\theta_1, \theta_2) d\theta_1 d\theta_2.$$

Interchanging the integrations with the summations we can easily find that the distribution of the difference Z is given by

$$P(Z = k) = \int \int e^{-(\theta_1+\theta_2)} \left(\frac{\theta_1}{\theta_2}\right)^{k/2} I_k(2\sqrt{\theta_1\theta_2}) g(\theta_1, \theta_2) d\theta_1 d\theta_2.$$

which does not depend on θ_3. However the above result must be used with caution because mixing a Bivariate Poisson distribution with the joint density $g(\theta_1, \theta_2)$ will result to a covariance which can be decomposed in two parts: one stemming from the intrinsic covariance of X and Y measured by θ_3 and one stemming from the covariance of the joint density $g(\theta_1, \theta_2)$ (see Stein and Juritz, 1987). Although the correlation due to the first part has been eliminated, correlation due the second one is still present unless the assumed mixing distribution $g(\theta_1, \theta_2)$ assumes independence for θ_1 and θ_2. For example, assuming that $g(\theta_1, \theta_2)$ is the product of two independent gamma distributions, giving rise to the bivariate negative binomial distribution (Kocherlakota and Kocherlakota, 1992), the probability of $Z > 0$ does not depend on the correlation parameter of the two initial random variables. Hence, the result of section 2 can also be extended for certain mixtures of bivariate Poisson distributions.

3 The Distribution of the Bivariate Poisson Absolute Difference

3.1 The Probability Distribution

In association with soccer data, the difference between the variables representing the number of goals scored by each team, represents the outcome of the game. The first variable X may represent the number of goals of the home team while the second Y the goals scored by the second team. Thus a positive sign for the difference represents the winning of the home team with a score difference given by $Z = X - Y$. For games played in neutral ground, for example in a World Cup, such differences are not well defined and hence analysis of the data can be speculative and inaccurate. In such cases the absolute difference can be used, which ignores the order in which the scores are referred. In this section we investigate and discuss the distribution of the absolute difference of two Poisson variates. Assuming the bivariate Poisson distribution then for the distribution of the absolute difference we conclude in a result similar to the one of section 3; that is, the absolute difference $W = |X - Y|$ does not depend on the covariance parameter θ_3.

The probability distribution function of the absolute difference W is given by

$$P_W(w) = P(W = w) = e^{-(\theta_1+\theta_2)} \; I_k\left(2\sqrt{\theta_1\theta_2}\right) \left(\left(\frac{\theta_1}{\theta_2}\right)^{k/2} + \left(\frac{\theta_2}{\theta_1}\right)^{k/2} \right) \tag{4}$$

since $P(W = w) = P(Z = -w) + P(Z = w)$ and from (3) we have that $P(Z = k|\theta_1, \theta_2) = P(Z = -k|\theta_2, \theta_1)$ and $I_k(2\sqrt{\theta_1\theta_2}) = I_{-k}(2\sqrt{\theta_1\theta_2})$. For the rest of the paper, this distribution will be called 'absolute Poisson difference' distribution and will be noted by $APD(\theta_1, \theta_2)$.

Note that interchanging the parameters lead to the same APD distribution. So, to avoid non-identifiability we assume $\theta_1 > \theta_2$. An alternative reparametrization can be adopted by setting $\theta_1 = \theta$ and $\theta_2 = \theta + \delta$; $\theta, \delta > 0$. In such case (4) can be rewritten as

$$P_W(w) = P(W = w) = e^{-(2\theta+\delta)} \; I_k\left(2\sqrt{\theta(\theta+\delta)}\right) \left(\left(1 + \frac{\delta}{\theta}\right)^{-k/2} + \left(1 + \frac{\delta}{\theta}\right)^{k/2} \right). \tag{5}$$

Calculation of the probabilities can be accomplished using the Poisson difference distribution. Histograms of the probability mass functions for various combinations of the parameters θ_1 and θ_2 are given in Figure 2.

Figure 2 about here

3.2 Moments

The moments of the distribution are not straightforward. Katti (1960) provided the moments of the absolute differences for several combinations of discrete distributions. For the case of two Poisson variates he showed that the r-th simple moment is given by

$$E(W^r) = e^{-(\theta_1+\theta_2)}[A_r(\theta_1,\theta_2) + A_r(\theta_2,\theta_1)]$$

where $A_s(\theta_1,\theta_2)$ is calculated by

$$A_s(\theta_1,\theta_2) = \sum_{i=0}^{s} \left(\begin{array}{c} s \\ i \end{array} \right) \left(\theta_1 + (-1)^{i+1}\frac{\theta_1\theta_2}{\theta_1}\right) A_{s-i}(\theta_1,\theta_2) + (-1)^s \frac{\theta_1\theta_2}{\theta_1} e^{2\sqrt{\theta_1\theta_2}} M\left(1/2; 1; -4\sqrt{\theta_1\theta_2}\right)$$

with

$$A_0(\theta_1,\theta_2) = \sum_{i=0}^{\infty} \frac{(\theta_1\theta_2)^i}{(i!)^2} M(1; i+1; \theta_2)$$

and

$$A_1(\theta_1,\theta_2) = (\theta_1-\theta_2)A_0(\theta_1,\theta_2) + \theta_1\theta_2 e^{2\sqrt{\theta_1\theta_2}} M\left(3/2; 3; -4\sqrt{\theta_1\theta_2}\right) + \theta_2 e^{2\sqrt{\theta_1\theta_2}} M\left(1/2; 1; -4\sqrt{\theta_1\theta_2}\right)$$

where $M(\alpha;\beta;z)$ denotes the confluent hypergeometric function (see Abramowitz and Stegun, 1974).

Unfortunately the above formulas are very complicated and do not provide a clear picture of how the mean varies with the parameters. Figure 3 demonstrates how the mean behaves for several combinations of the parameters. We have considered the alternative parametrization (θ,δ) and the probability function (5) to examine the index of dispersion of the distribution.

Figure 3 about here

Note that, as the difference of the parameters $(\theta_1 - \theta_2)$ increases, the mean tends to their absolute difference. This is due to the fact that the differences $X - Y$ have the same sign due to the large difference of the means. For such cases, the variance is high, leading to large index of dispersion. Similarly, if θ_1 tends to 0, the distributions tends to a simple Poisson distribution.

In the special case of $\theta_1 = \theta_2 = \theta$ the mean simplifies to

$$E(W) = 2\theta \left\{ \theta M(3/2; 3; -4\theta) + M(1/2; 1; -4\theta) \right\}.$$

Note a misprint in the original paper of Katti (1960). This distribution is interesting as it has only a single parameter and thus a direct comparison with the Poisson distribution can be made. The distribution is overdispersed for values of the parameter that exceeds 1.7389. For selected values of the parameter one can see the mean, the variance and the index of dispersion in Table 1.

Table 1 about here

3.3 Estimation of the Parameters

Estimation of the parameters is much more complicated than the simpler case of the difference. Moment estimates are rather cumbersome for practical purposes, since the theoretical moments involve the confluent hypergeometric function. On the contrary, maximum likelihood estimation can be carried out by constructing an EM algorithm. This algorithm is briefly described below:

E-step: Assuming that the current values of the parameters are given by $\theta_1^{(k)}$ and $\theta_2^{(k)}$ from the k-th iteration, the expected values of X_i and Y_i given the current values of the parameters are calculated. Using simple probability calculus we have that

$$\begin{aligned} t_i = E(X_i | w_i, \theta_1^{(k)}, \theta_2^{(k)}) &= \sum_{x=0}^{\infty} x P(X_i = x | W_i = w_i) \\ &= \sum_{x=0}^{\infty} x \frac{P(X_i = x)\left(P(Y = x - w_i) + P(Y = x + w_i)\right)}{P(W_i = w_i)} \\ &= \theta_1 \frac{P_Z(z_i - 1) + P_Z(-z_i - 1)}{P_W(w_i)}. \end{aligned}$$

Similarly

$$s_i = E(Y_i | w_i, \theta_1^{(k)}, \theta_2^{(k)}) = \theta_2 \frac{P_Z(z_i - 1) + P_Z(-z_i - 1)}{P_W(w_i)}.$$

where $P_Z(z)$, $P_W(w)$ are defined in (3) and (4) respectively and $P_{-Z}(z)$ is the probability function of $Y - X$, that is $PD(\theta_2.\theta_1)$.

M-step: Update the estimates by $\theta_1^{(k+1)} = n^{-1} \sum_{i=1}^{n} t_i$ and $\theta_2^{(k+1)} = n^{-1} \sum_{i=1}^{n} s_i$ Again, the above scheme has all the properties of the standard EM algorithm. No specific problems were encountered when applying the algorithm.

4 Application in Soccer

4.1 The Poisson Assumption

Sports related industry is an increasing area and vast amounts of money are invested in such activities. Betting on soccer games outcomes is common in many countries. Often, we wish to be able to predict the outcome of a soccer game (or batches of games). The possible outcomes are: win, loss, or draw, and a gambler gains money according to the number of successes in his predictions (for an extended review of statistics in soccer see Bennet, 1998). Other kinds of bets, involve the prediction of the 'lead' of the winning team, measured as the difference in the number of goals scored by each team.

The Poisson distribution has a formal theoretical basis and is naturally used for events that occur randomly at a constant rate over the observed time period. Thus if we assume a constant ability across time of a team to score, the Poisson distribution seems a plausible model. The majority of the papers in this field assume that the number of goals scored by each team follows a Poisson distribution (Maher, 1982, Lee, 1987, Rue and Salvesen, 1997 among others). The observed overdispersion led some authors to advocate the use of an overdispersed alternative to the simple Poisson distribution, such as the negative binomial distribution (Baxter and Stevenson, 1988). However, the negative binomial does not offer a significant improvement upon the results. On the contrary, interpretation of the results becomes more complicated.

Another crucial question, arising in modelling soccer games, is whether the number of goals scored by each team are dependent. Since the two teams interact into the game it seems plausible that some kind of dependence exist. Lee (1997) reported that the dependence is rejected via a chi-square test and in any case it is rather small to be considered as important.

Our results from section (2) offer another theoretical explanation for ignoring the dependence (if any) on the number of goals scored by each team.

Specifically, assume that the random variables X and Y represent the number of goals scored by each team, respectively. The joint distribution of these variables can be assumed as a bivariate Poisson distribution, allowing for the dependence between the goals scored by each team. In such a model, the marginal distributions are Poisson as it is commonly assumed for sports data. A natural interpretation for the parameters is that θ_1 and θ_2 are the parameters reflecting the scoring ability of the two teams while parameter θ_3 reflects game conditions.

The sign of the random variable $Z = X - Y$ reveals the winner of the game while a zero value corresponds to a draw. The result of section 2 implies that the winning probbaility does not depend on this parameter. Treating the number of goals for each team separately is easier and allows for calculation via simple statistical packages, expanding the applicability of statistical methods for betting purposes. It should be kept in mind that since the parameters θ_1 and θ_2 are estimated from the marginal distributions, the covarianec parameter θ_3 is confounded. Using the Poisson differene distribution when the data are the differences in the goals, the 'net' scoring abilities are estimated. To do so with the original data one has to find ML estimates from a bivariate Poisson distribution which is relatively more difficult.

For example, Maher (1982) developed Poisson regression models for soccer outcomes. Such models take the general form

$$n_{ijk} \sim Poisson(\lambda_{ijk}), \quad log(\lambda_{ijk}) = \mu + h_i + a_j + d_k + h.a_{ij} + h.d_{ik} + a.d_{jk} + h.a.d_{ijk}$$

where n_{ijk} and λ_{ijk} are the observed and the mean, respectively, of the goals scored by team j, with opponent team k, playing in soccer ground i (away/home); μ is a constant parameter, h_i is the home effect parameter, a_j is the parameter for the offensive performance of j team and d_k encapsulates the defensive performance of k team. Karlis and Ntzoufras (1998) examined such models in a general log-linear setting allowing also for model selection. Our result verifies the commonly used approach that handles the two teams as independent and hence the data from N games consist of $2N$ data points.

In addition, Keller (1994) examined the probability of a win starting from independent Poisson variates. Clearly, the results presented in Keller (1994) are also valid for dependent Poisson variates.

In conclusion, violation of the independence assumption in Poisson distribution does not affect the probabilities of soccer outcomes and therefore simple Poisson models can be utilized to calculate these probabilities. On the other hand, this result is not useful when we wish to predict the number of goals scored by each team or the total number of goals scored in the game.

Recently, Kuonen and Rohl (2000) modelled the difference on the number of goals as a Normal variate. For small values of θ_1 and θ_2 as those expected for soccer data, the approximation of the Normal distribution is rather poor. It would more appropriate to consider a model based on the distribution defined by (3).

4.2 Implementation of the new Distributions

4.2.1 The Poisson Difference Distribution

In this section we use game results of twenty four (24) different European leagues of six (6) different countries from the period 1989-1995 in order to fit the Poisson difference distribution. The number of teams participating in each league varies. Moreover, the Poisson distribution was also used for the number of goals scored by the home and the guest teams. Preliminary investigation of the data revealed that some sort of overdispersion is present in almost all cases though it varies from league to league.

The EM algorithm was used to obtain the ML estimates of the PD distribution. The assumption of a Poisson distribution was tested using a bootstrap version of the index of dispersion test statistic (see, for example, Stute *et al.*, 1993, Karlis and Xekalaki, 2000). Results are reported in Table 2. Covariance is rather small for all the cases. In almost all cases it is not significantly different from a zero value.

Note that in many cases the Poisson difference assumption is rejected by the usual χ^2-goodness of fit test. The distribution fits well the data only in a few cases. This can be attributed both to the observed feature that the component distributions were not Poisson in the majority of cases and the observation that the larger residual was observed at zero value in all the fitted distributions. This fact of an excess in the expected number of draws is well known in the literature (see, for example, Dixon and Coles, 1997).

Table 2 about here

4.2.2 The Absolute Poisson Distribution Distribution

The absolute Poisson difference (APD) distribution is particularly suitable for games where there is not home effect and hence the two teams cannot be ordered. Such an example is the World Cup competition, held in a neutral country and where home effect does not exist (apart from the home team). With such data the interest lies in the absolute difference revealing the lead of the winner team.

The data from the matches of the World Cup' 98, held in France, were used for illustrative purposes. The score at the end of the game has been considered ignoring extra-time wherever exist. The data representing the differences in goals can be seen in Table 3. The mean equals 1.28 while the variance is 1.73. We fitted the APD distribution using the EM algorithm described. The obtained estimates were $\hat{\theta}_1 = 1.517$ and $\hat{\theta}_2 = 0.531$. The p-value of a χ^2 goodness of fit test was 0.05. The APD distribution improves the likelihood with respect to the Poisson distribution, as it was expected due to the overdispersion present at the data. The log-likelihood for the APD distribution were -96.83 while the Poisson log-likelihood was -97.71. Note that for this data set the zero frequency is not much higher than the expected. This can be attributed to the purpose of these matches and to the design of the tournament which does not support draws.

Table 3 about here

5 Extended Models

5.1 Zero Inflated Distributions

From section 4 we observe that the zero frequency in almost all the cases is larger than the expected frequency for draws under the PD model. The same observations have been made by other authors when modelling soccer games (see for example Dixon and Coles, 1997).

Zero inflated models can be adeqaute to overcome this problem. Such models assume that the probability of observing a 0 increases with adequet decrease of all the rest probabilities. So, we define the zero inflated-PD (Z-PD) distribution as

$$\begin{aligned} P(Z=0) &= p + (1-p)PD(Z=0, \theta_1, \theta_2) \\ P(Z=k) &= (1-p)PD(Z=k, \theta_1, \theta_2), \end{aligned}$$

for $k = \ldots, -3, -2, -1, 1, 2, 3, \ldots$ where $PD(Z=k, \theta_1, \theta_2)$ is given in (3) and $0 < p < 1$.

Zero inflated distributions has been described in Johnson *et al.* (1992) and the references therein. Recently Bohning *et al.* recast interest in such distribution proposing zero inflated Poisson distributions allowing for covariates (see also Lambert, 1992).

Denoting as X_{ZI} the zero-inflated random variable stemming from the random Variable X it can be seen that the simple moments of X_{ZI} are given as $E(X_{ZI} = (1-p)E(X)$ and thus for the zero inflated Poisson difference distribution given in (6) one obtains that $E(Z) = (1-p)(\theta_1 - \theta_2)$ and $Var(Z) = (1-p)(\theta_1 + \theta_2) + p(1-p)(\theta_1 - \theta_2)^2$.

Fortunatelly maximum likelihood estimation acan be accompolished easily via an EM type algorithm. At first one can recognize that zero-inflated distribution are in fact finite mixture distributions (see e.g Titterington *et al.*, 1985) with a degenerate at 0 component. The additional parameter p can be considered as the mixing proportion and then the standard EM algorithm for mixture models apply. More formally the EM can be described as follows

E-step: With the current values of the parameters $\theta_1^{(k)}$, $\theta_2^{(k)}$ and $p^{(k)}$ from the k-th iteration, calculate the expected values of X_i and Y_i given the current values of the parameters. Using simple probability calculus is obtained that

$$t_i = E(X_i|z_i, \theta_1^{(k)}, \theta_2^{(k)}, p^{(k)}) \quad = \quad \sum_{x=0}^{\infty} x \frac{P(X_i = x)P(Y = x - z_i)}{P(Z_i = z_i)}$$

where $P_Z(z)$ is defined in (6). One can recognize that this step is the same as the E-step made for the PD distribution apart from the denominator which is, now, the zero-inflated counterpart.

M-step: Update the estimates by

$$p^{(k+1)} = \frac{p^{(k)} f(0)}{nP(Z=0)} \quad \text{and} \quad \theta_1^{(k+1)} = n^{-1} \sum_{i=1}^{n} t_i \quad \text{and} \quad \theta_2^{(k+1)} = \theta_1^{(k+1)} - \frac{\bar{z}}{1 - p^{(k+1)}}$$

where $f(0)$ stands for the observed frequency of the zero value. It is not necessary to calculate both conditional expectations, since second parameter can be updated by simple subtraction.

Again the above step has all the properties of the standard EM algorithm.

It must be pointed out that the zero-inflated Poisson difference distribution given in (6) is not the same with the distribution arising as the difference of two zero-inflated Poisson distribtuions. The following theorem eamines the more general case of the distribution of the difference of two Finite Poisson mixtures. A zero-inflated Poisson distribution can be seen in this context as it arises when one component is a Poisson distribution with zero parameter, i.e. a degenerate at 0 distribution.

Theorem 2 : Suppose that X follows a k_1 finite Poisson mixture distribution with parameters λ_{1i} and mixing proportions $\phi_i, i = 1, \ldots, k_1$. Suppose also that Y follows a k_2 finite Poisson mixture distribution with parameters λ_{2j} and mixing proportions π_j, $j = 1, \ldots, k_2$. Then their difference $Z = X - Y$ follows a $k_1 k_2$ finite Poisson Difference mixture where the $ij - th$ component has parameters $\lambda_{1i}, \lambda_{2j}$ and mixing proportion $\phi_i \pi_j$, $i = 1, \ldots, k_1, j = 1, \ldots, k_2$.

Proof: The result can be deduced by using the fact that
$P(Z = z) = \sum\limits_{r=0}^{\infty} P(X = x)P(Y = x - z)$.
Since $P(X = x)$ has k_1 terms and $P(Y = x - z)$ has k_2 terms, then multiplying we obtain $k_1 k_2$ terms. The requisite parameters stems from this multiplication of the $i - th$ term of the first part with the $j - th$ term of the second.

Consider the simplest case with $k_1 = k_2 = 2$, i.e. the difference of two 2-finite Poisson mixtures. Then the resulting distribution has 4 Poisson difference components. Suppose further that each mixture is in fact a zero-inflated Poisson distribution, i.e. one parameter for each mixture is 0, and the other components have parameters,say λ_1 and λ_2 for each mixture respectively. Then the four components of the resulting Poisson difference mixture have parameters λ_1, λ_2, $\lambda_1, 0$, $\lambda_2, 0$, and 0. Thus the distribution has inflation at 0 as it is described by the component with 0 parameters, while the other components are a Poisson difference distribution and two simple Poisson distributions (those having a 0 parameters). Clearly this distribution is different form the zero-inflated distribution defined in (6).

5.2 A Negative Binomial -Poisson Model

A possible extension of the PD distribution is to assume that the one component is not a Poisson distribution. Assume that the parameter θ_1 in (3) follows itself a distribution, say $g(\theta_1)$. Then the resulting distribution is a mixture. Further assume that $g(\theta)$ is the Gamma distribution. Then the probability function is given by

$$P(Z = k) \;\; = \;\; \int_0^{\infty} PD(k; \theta_1, \theta_2) g(\theta_1) d\theta_1$$

$$= \frac{\beta^\alpha exp(-\theta_2)}{(1+\beta)^{k+\alpha}\Gamma(\alpha)} M\left(\alpha+k; k+1; \frac{\theta_2}{1+\beta}\right)$$

for $k = \ldots, -2, -1, 0, 1, 2, \ldots$, $\alpha, \beta > 0$. It can easily be seen that the above distribution is also the distribution of $Z = X - Y$, where X follows a Negative Binomial distribution with parameters α and β (the parameters correspond to the mixture formulation of the negative binomial as a mixture of the Poisson distribution with a $Gamma(\alpha, \beta)$ mixing distribution). Moreover, the above distribution can be derived from a bivariate model with negative binomial and Poisson marginals, which allows for dependence between the variables (see section 2.4).

Generalisation can be provided assuming a Gamma mixing distribution for the parameter θ_2 which correspond to the difference of two Negative binomial distributions, but we do not pursue further this issue.

6 Discussion

In the literature there are only few distributions for describing the difference of integer valued discrete distributions. In this paper we examine two distributions appropriate for modelling soccer data; the Poisson difference and the absolute Poisson difference distributions.

In soccer games, observed counts are rather small and hence the assumption of normality is far from realistic. On the other hand, such an assumption may be realistic for other games where the number of goals scored is larger (for example handball). Thus the distributions presented in this paper can be helpful for modelling purposes of soccer data. Moreover, it is proven that the dependence implied by a bivariate Poisson distribution is indifferent in respect to the distribution of the difference of the two random variables. This offers a theoretical background in order to use Poisson model formulations assuming that the number of goals of each team are independent.

References

[1] Abramowitz, M. and Stegum, I.A. (1974). *Handbook of Mathematical Functions.* New York: Dover

[2] Baxter, M. and Stevenson, R. (1988). Discriminating Between the Poisson and the Negative Binomial Distributions: An Application to Goal Scoring in Association Football. *Journal of Applied Statistics*, **15**, 347-348.

[3] Bennet, J. (1998). *Statistics in Sports.* First Edition, London: Edward Arnold.

[4] Bohning, D., Dietz, E., Schlattmann, P., Mendonca, L. and Kirchner, U. (1999). The Zero-Inflated Poisson Model and the Decayed, Missing and Filled Teeth Index in Dental Epidemiology. *Journal of the Royal Statistical Society A*, **162**, 265-280.

[5] Dixon, M.J. and Coles, S.G. (1997). Modelling Association Football Scored and Inefficiencies in Football Betting Market. *Applied Statistics*, **46**, 265-280.

[6] Irwin, W. (1937). The Frequency Distribution of the Difference Between Two Poisson Variates Following the Same Poisson Distribution. *Journal of the Royal Statistical Society A*, **100**, 415-416.

[7] Jackson, D.A. (1994). Index Betting on Sports. *The Statistician*, **43**, 309-315.

[8] Karlis, D. and Ntzoufras, I. (1998). Statistical Modelling for Soccer Games. *Proceedings of the Fourth Hellenic-European Conference on Computer Mathematics and its Applications* (E.A.Lipitakis, eds.). Athens: LEA, 541-548.

[9] Karlis D. and Xekalaki, E. (2000). A Simulation Comparison of Several Tests for Goodness of Fit for the Poisson Distribution. *The Statistician, to appear.*

[10] Katti, S.K. (1960). Moments of the Absolute Difference and Absolute Deviation of Discrete Distributions. *Annals of Statistics*, **31**, 78-85.

[11] Keilson, J. and Gerber, H. (1971). Some Results in Discrete Unimodality. *Journal of the American Statistical Association*, **66** , 386-389.

[12] Keller, J. (1994). A Characterization of the Poisson Distribution and the Probability of Winning a Game. *The American Statistician*, **48**, 294-299

[13] Kocherlakota, S. and Kocherlakota, K. (1992). *Bivariate Discrete Distributions.* New York: Marcel Dekker

[14] Kuonen, D. and Rohl, A.S.A. (2000). Was France's World Cup Win Pure Chance? *Student* (to appear).

[15] Lambert, D. (1992). Zero-inflated Poisson Regression, with Applications to Defects in Manufacturing. *Technometrics*, **34**, 1-14.

[16] Lee, A.J. (1997). Modeling Scores in the Premier league: Is Manchester United Really the Best? *Chance*, **10**, 15-19.

[17] Maher, M.J. (1982). Modelling Association Football Scores. *Statistica Neerlandica*, **36**, 109–118.

[18] McLachlan, G. and Krishnan, N. (1997). *The EM Algorithm and Extensions.* Chichester: John Wiley & Sons.

[19] Rue, H. and Salvesen, O. (1997). Predicting Soccer Matches in a League. *Technical Report*, Department of Mathematical Sciences, Norwegian University of Science and Technology, Norway. *The Statistician*, (to appear).

[20] Skellam, J.G. (1946). The Frequency Distribution of the Difference Between Two Poisson Variates Belonging to Different Populations. *Journal of the Royal Statistical Society B*, **10**, 257-261.

[21] Stein, G.Z. and Juritz, J. (1987). Bivariate Compound Poisson Distributions. *Communication in Statistics: Theory and Methods*, **16**, 3591-3607.

[22] Stute, W., Manteiga, W.G. and Quindmil, M. P. (1993). Bootstrap Based Goodness of Fit Tests. *Metrika,* **40**, 243-256

[23] Titterington, M. , Makov, U. and Smith A.F.M. (1985) *Statistical Analysis of Finite Mixtures.* Wiley, London.

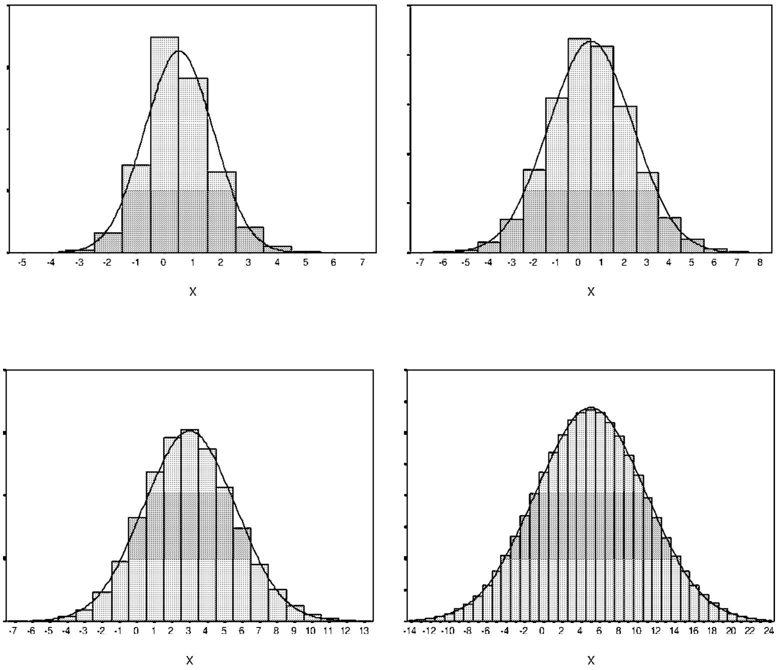


Figure 1: Probability Mass Function of Poisson Difference in Comparison to the Normal Density for a variety of parameter values.The parameters used were (2,1.5), (4,3.5),(5,2) and (20,15), respectively

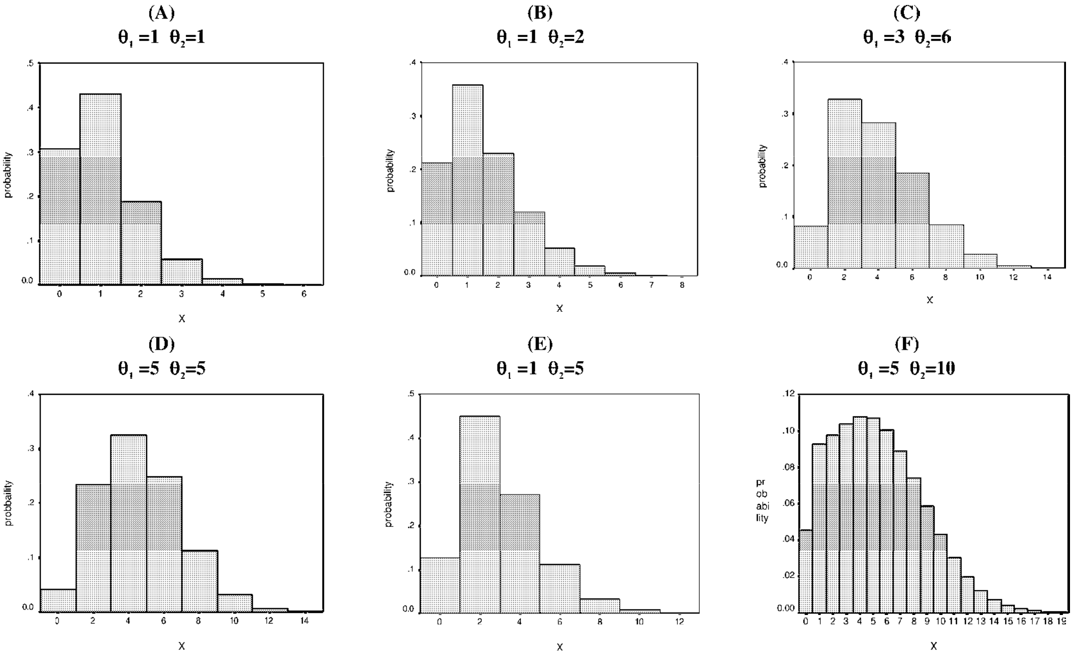


Figure 2: Histograms of the APD Probability Function for Various Combinations of the Parameters.

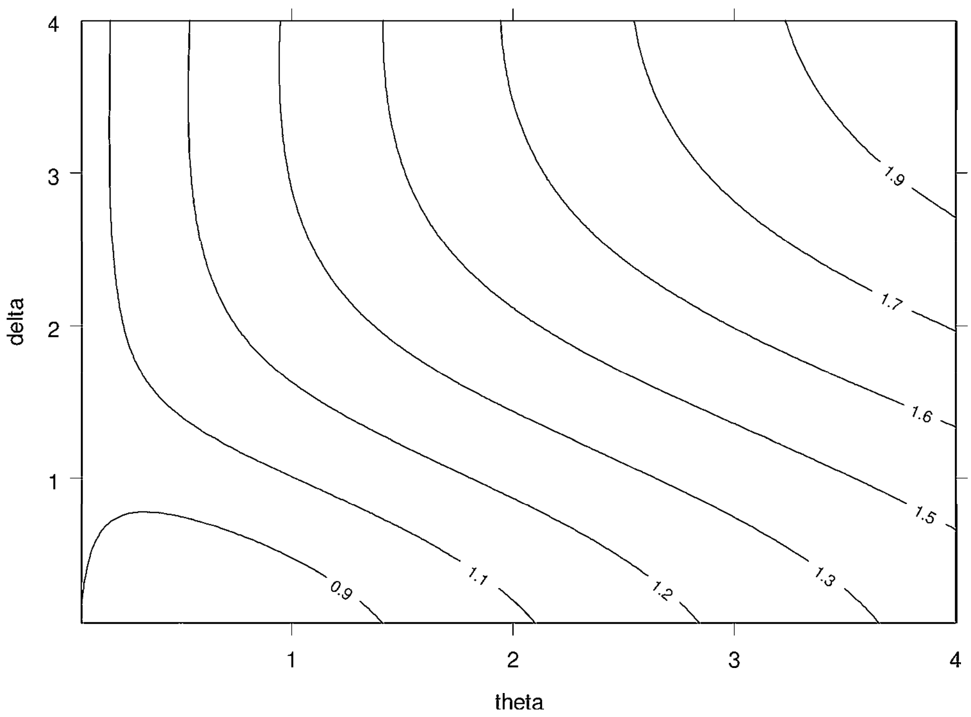


Figure 3: Contour Plot for the Overdispersion of APD Distribution.

θ	Mean	Variance	Index of Dispersion
0.1	0.181	0.166	0.918
0.2	0.333	0.288	0.865
0.3	0.462	0.385	0.833
0.4	0.574	0.469	0.816
0.5	0.673	0.546	0.810
0.7	0.842	0.690	0.820
1.0	1.047	0.902	0.861
1.2	1.164	1.044	0.897
1.5	1.319	1.259	0.954
2.0	1.543	1.619	1.049
2.5	1.737	1.980	1.140
3.0	1.912	2.343	1.225
3.5	2.072	2.706	1.305
4.0	2.220	3.069	1.382
5.0	2.490	3.795	1.523
6.0	2.734	4.521	1.653
7.0	2.958	5.248	1.773
8.0	3.165	5.980	1.889

Table 1: APD distribution with equal parameters. Underdisperesion is observed only for small parameter values (< 1.74).

	Home		Guest			Home	Guest				
League	Mean	Var	Mean	Var	Covar	p-val	p-val	$\hat{\theta}_1$	$\hat{\theta}_2$	χ^2	p-val
Dutch 93	1.73	1.79	1.21	1.63	-0.38	0.370	0.000	2.377	1.857	17.20	0.045
Dutch 94	1.90	2.40	1.31	1.50	-0.26	0.000	0.060	2.463	1.872	19.43	0.021
Dutch 95	1.75	2.39	1.25	1.48	-0.29	0.000	0.000	2.404	1.907	24.45	0.003
Ger 89	1.68	1.96	0.91	0.96	-0.04	0.020	0.210	1.824	1.053	25.26	0.002
Ger 93	1.75	1.72	1.17	1.39	-0.15	0.560	0.010	1.969	1.391	19.15	0.023
Ger2 93	1.53	1.70	0.97	1.07	0.04	0.120	0.050	1.597	1.031	17.22	0.045
Ger 94	1.76	1.81	1.24	1.24	-0.14	0.380	0.430	1.950	1.420	30.50	0.000
Ger 95	1.54	1.50	1.17	1.24	0.01	0.630	0.260	1.465	1.096	28.06	0.000
Italy 89	1.37	1.34	0.87	0.84	0.12	0.530	0.620	1.150	0.646	17.18	0.045
Italy 90	1.43	1.74	0.86	0.89	0.05	0.000	0.310	1.445	0.869	34.71	0.000
Italy 91	1.34	1.39	0.93	1.09	0.11	0.320	0.000	1.257	0.849	21.08	0.012
Italy 92	1.72	1.72	1.08	1.22	0.24	0.420	0.110	1.470	0.836	24.17	0.004
Italy 93	1.48	1.63	0.94	1.00	0.03	0.120	0.180	1.463	0.917	19.59	0.020
Italy 94	1.57	1.72	0.96	1.21	-0.05	0.140	0.000	1.816	1.205	4.95	0.838
Italy 95	1.66	1.87	0.97	0.97	0.09	0.090	0.520	1.635	0.945	14.49	0.105
France 93	1.42	1.42	0.81	0.81	0.01	0.520	0.500	1.348	0.738	14.54	0.104
France 94	1.59	1.63	0.92	1.04	0.01	0.210	0.040	1.632	0.956	11.18	0.263
France 95	1.44	1.41	0.84	0.98	0.03	0.490	0.010	1.445	0.850	18.90	0.026
Spain 93	1.59	2.04	1.01	1.10	0.10	0.000	0.110	1.683	1.099	12.19	0.202
Spain 94	1.54	1.78	1.00	1.28	-0.09	0.000	0.000	1.841	1.299	18.19	0.033
Spain 95	1.56	1.71	1.14	1.41	0.06	0.120	0.000	1.666	1.242	17.79	0.037
Engl 93	1.44	1.57	1.15	1.23	0.02	0.070	0.140	1.455	1.171	16.21	0.062
Engl 94	1.51	1.65	1.08	1.21	0.04	0.100	0.040	1.544	1.113	11.74	0.228
Engl 95	1.53	1.58	1.07	1.25	0.04	0.300	0.020	1.558	1.106	11.98	0.214

Table 2: Table of Fitted PD Distribution for 24 European Leagues.

Absolute Goal Difference	Observed	APD Expected
0	19	16.37
1	27	24.69
2	6	13.83
3	8	6.11
4	1	2.17
5	3	0.83
$X^2 = 5.99$	df=2	p-value=0.05

Table 3: Frequency Tabulation of Absolute Goal Differences of 1998 World Cup.