Bayesian and Non-Bayesian Analysis of Soccer Data using Bivariate Poisson Regression Models

Dimitris Karlis
Department of Statistics
Athens University of Economics

John Ntzoufras
Dept. of Business Administration
University of the Aegean

Kavala, April 2003

Outline

- Statistical Models and soccer
- Bivariate Poisson model, pros and cons
- Bivariate Poisson regression model
- ML estimation through EM
- Bayesian estimation through MCMC
- Inflated Models
- Application

Statistical models for football: Motivation

- Insight into game characteristics
  (e.g. game behavior, coaching tactics , strategies, injury prevention etc)
- Team as companies
  (e.g. human resources, investment analysis etc)
- Betting purposes
  (e.g. betting on the outcome, on score or on any other characteristic)

Statistical models for football: type of models

- Model win-loss (no score included)
  (e.g. Paired comparison models, logistic regression etc)
- Model score
  (e.g Independent Poisson model, negative binomial alternative, our newly proposed bivariate Poisson model)
- Model game characteristics
  (e.g. effect of red card, artificial field, passing game etc)

Important questions to be answered

- Poisson or not Poisson
  Real data show small overdispersion. In practice the overdispersion is negligible especially if covariates are included
- Independence between the goals of the two competing teams
  Empirical evidence show small and not significant correlation (usually less than 0.05). We will show that even so small correlation can have impact to the results.

Existing models

Let X and Y the number of goals scored by the home and guest team respectively. The usual model is

$$\begin{aligned} X &\sim Poisson(\lambda_1) \\ Y &\sim Poisson(\lambda_2) \end{aligned}$$

independently and λ_1, λ_2 depend on some parameters associated to the offensive and defensive strength of the two teams.

We relax the independence assumption

Bivariate Poisson model

Let $X_i \sim Poisson(\theta_i)$, $i = 0, 1, 2$

Consider the random variables

$$\begin{aligned} X &= X_1 + X_0 \\ Y &= X_2 + X_0 \end{aligned}$$

$(X, Y) \sim BP(\theta_1, \theta_2, \theta_0)$,

Joint probability function given:

$$P(X = x, Y = y) = e^{-(\theta_1+\theta_2+\theta_0)} \frac{\theta_1^x}{x!} \frac{\theta_2^y}{y!} \sum_{i=0}^{min(x,y)} \binom{x}{i} \binom{y}{i} i! \left(\frac{\theta_0}{\theta_1\theta_2}\right)^i .$$

Properties of Bivariate Poisson model

- Marginal distributions are Poisson, i.e.

$$\begin{aligned} X &\sim Poisson(\theta_1 + \theta_0) \\ Y &\sim Poisson(\theta_2 + \theta_0) \end{aligned}$$

- Conditional Distributions : Convolution of a Poisson with a Binomial
- Covariance: $Cov(X, Y) = \theta_0$

  For a full account see Kocherlakota and Kocherlakota (1992) and Johnson, Kotz and Balakrishnan (1997)

Bivariate Poisson regression model

$$\begin{aligned}
(X_i, Y_i) &\sim BP(\lambda_{1i}, \lambda_{2i}, \lambda_{3i}), \\
\log(\lambda_{1i}) &= \boldsymbol{w}_{1i}\boldsymbol{\beta}_1, \\
\log(\lambda_{2i}) &= \boldsymbol{w}_{2i}\boldsymbol{\beta}_2, \\
\log(\lambda_{3i}) &= \boldsymbol{w}_{3i}\boldsymbol{\beta}_3,
\end{aligned} \tag{1}$$

$i = 1, \ldots, n$, denotes the observation number, $\boldsymbol{w}_{\kappa i}$ denotes a vector of explanatory variables for the i-th observation used to model $\lambda_{\kappa i}$ and $\boldsymbol{\beta}_\kappa$ denotes the corresponding vector of regression coefficients.

Explanatory variables used to model each parameter $\lambda_{\kappa i}$ may not be the same.

Bivariate Poisson regression model (continued)

- Allows for covariate-dependent covariance.
- Separate modelling of means and covariance
- Standard estimation methods not easy to apply.
- Computationally demanding.
- Application of an easily programmable EM algorithm

Applications

- Paired count data in medical research
- Number of accidents in sites before and after infrastructure changes
- Marketing: Joint purchases of two products (customer characteristics as covariates)
- Epidemiology: Joint concurrence of two different diseases.
- Engineering: Faults due to different causes
- Sports especially soccer, waterpolo, handball etc
- etc

Important Result

Let X, Y the number of goals for the home and the guest teams respectively. Define $Z = X - Y$. The sign of Z determines the winner. What is the probability function of Z if X, Y jointly follow a bivariate Poisson distribution?

Solution:

$$P_Z(z) = P(Z = z) = e^{-(\lambda_1+\lambda_2)} \left(\frac{\lambda_1}{\lambda_2}\right)^{z/2} I_z\left(2\sqrt{\lambda_1\lambda_2}\right), \qquad (2)$$

$z = \ldots, -3, -2, -1, 0, 1, 2, 3, \ldots$, where $I_r(x)$ denotes the Modified Bessel function

Remark 1: The distribution has the same form as the one for the difference of two independent Poisson variates (Skellam, 1946)

Remark 2: The distribution does not depend on the correlation parameter!

Summarizing: Irrespectively the correlation between X, Y the distribution of $X - Y$ has the same form!

Important difference: There is a large difference in the interpretation of the parameters. So, for the given data the two different models (independent Poisson, bivariate Poisson) lead to different estimate for winning a game.

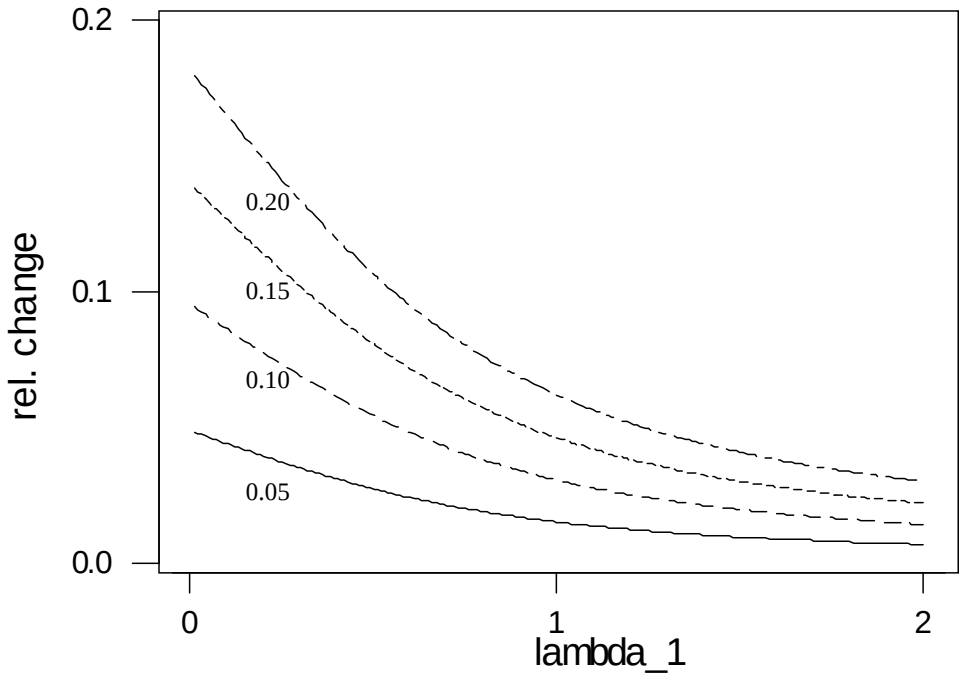


Figure 1: The relative change of the probability of a draw, when the two competing teams have marginal means equal to $\lambda_1 = 1$ and λ_2 ranging from 0 to 2. The different lines correspond to different levels of correlation.

Table 1: The gain for betting using a misspecified model. We have set $\lambda_1 = 1$, and we vary the values of λ_2, λ_3. The entries of the table are the expected gain per unit of bet.

	λ_3									
λ_2	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
0.5	0.0079	0.0160	0.0242	0.0326	0.0412	0.0500	0.0589	0.0681	0.0774	0.0870
0.6	0.0075	0.0152	0.0230	0.0310	0.0391	0.0474	0.0559	0.0646	0.0734	0.0824
0.7	0.0071	0.0144	0.0218	0.0294	0.0371	0.0450	0.0530	0.0612	0.0696	0.0781
0.8	0.0068	0.0137	0.0207	0.0279	0.0352	0.0426	0.0502	0.0580	0.0659	0.0739
0.9	0.0064	0.0130	0.0196	0.0264	0.0333	0.0404	0.0476	0.0549	0.0623	0.0699
1	0.0061	0.0123	0.0186	0.0250	0.0316	0.0382	0.0450	0.0519	0.0589	0.0660
1.1	0.0058	0.0116	0.0176	0.0237	0.0298	0.0361	0.0425	0.0489	0.0555	0.0623
1.2	0.0054	0.0110	0.0166	0.0223	0.0281	0.0340	0.0400	0.0461	0.0523	0.0586
1.3	0.0051	0.0103	0.0156	0.0210	0.0265	0.0320	0.0377	0.0434	0.0492	0.0551
1.4	0.0048	0.0097	0.0147	0.0197	0.0249	0.0301	0.0353	0.0407	0.0462	0.0517
1.5	0.0045	0.0091	0.0138	0.0185	0.0233	0.0282	0.0331	0.0381	0.0432	0.0483

Estimation - ML method

- Likelihood is intractable as it involves multiple summation
- The trivariate reduction derivation allows for an easy EM type algorithm.
- Same augmentation will be used for Bayesian analysis

- **Recall:** If X_1, X_2, S independent Poisson variates then $X = X_1 + S, Y = X_2 + S$ follow a bivariate Poisson distribution.

  Complete data $Y_{com} = (X_1, X_2, S)$
  Incomplete (observed) data $Y_{inc} = (X, Y)$

So, if we knew X_0 the estimation task would be straightforward.

EM algorithm

E-step: With the current values of the parameters $\boldsymbol{\lambda}_1^{(k)}$, $\boldsymbol{\lambda}_2^{(k)}$ and $\boldsymbol{\lambda}_3^{(k)}$ from the k-th iteration, calculate the expected values of S_i given the current values of the parameters:

$$\begin{aligned} s_i &= E(S_i \mid X_i, Y_i, \boldsymbol{\lambda}_1^{(k)}, \boldsymbol{\lambda}_2^{(k)}, \boldsymbol{\lambda}_3^{(k)}) \\ &= \begin{cases} \lambda_{3i}^{(k)} \frac{BP(x_i-1, y_i-1 \mid \lambda_{1i}^{(k)}, \lambda_{2i}^{(k)}, \lambda_{3i}^{(k)})}{BP(x_i, y_i \mid \lambda_{1i}^{(k)}, \lambda_{2i}^{(k)}, \lambda_{3i}^{(k)})}, & \text{if } min(x_i, y_i) > 0 \\ 0 & \text{if } min(x_i, y_i) = 0 \end{cases} \end{aligned}$$

where $BP(x, y \mid \lambda_1, \lambda_2, \lambda_3)$ is the joint probability function distribution of the $BP(\lambda_1, \lambda_2, \lambda_3)$ distribution.

EM algorithm - M-step

M-step: Update the estimates by

$$\begin{aligned}
\boldsymbol{\beta}_1^{(k+1)} &= \hat{\boldsymbol{\beta}}(\boldsymbol{x} - \boldsymbol{s}, \boldsymbol{W}_1), \\
\boldsymbol{\beta}_2^{(k+1)} &= \hat{\boldsymbol{\beta}}(\boldsymbol{y} - \boldsymbol{s}, \boldsymbol{W}_2), \\
\boldsymbol{\beta}_3^{(k+1)} &= \hat{\boldsymbol{\beta}}(\boldsymbol{s}, \boldsymbol{W}_3);
\end{aligned}$$

where $\boldsymbol{s} = [s_1, \ldots, s_n]^T$ is the $n \times 1$ vector, $\hat{\boldsymbol{\beta}}(\boldsymbol{x}, \boldsymbol{W})$ are the maximum likelihood estimated parameters of a Poisson model with response the vector $\boldsymbol{x}$ and design or data matrix given by $\boldsymbol{W}$. The parameters $\boldsymbol{\lambda}_\ell^{(k+1)}$, $\ell = 1, 2, 3$ are calculated directly from (1). Note that one may use different covariates for each λ, for example different data or design matrices.

Estimation- Bayesian estimation via MCMC algorithm

Closed form Bayesian estimation is impossible

Need to use MCMC methods

Implementation details

- Use the same data augmentation
- Jeffrey priors for regression coefficients
- The posterior distributions of β_r, $r = 1, 2, 3$ are non-standard and, hence, Metropolis-Hastings steps are needed within the Gibbs sampler,

More details

The conditional posterior of the latent variable S_i is given as

$$s_i \mid \cdot \propto \left(\frac{\lambda_{3i}}{\lambda_{1i}\lambda_{2i}}\right)^{s_i} \frac{1}{(x_i - s_i)!(y_i - s_i)!}, \quad s_i = 0, ..., min(x_i, y_i)$$

The conditional posteriors for β_i, $i = 1, 2, 3$, are the usual for Poisson GLM using as responses $\boldsymbol{x} - \boldsymbol{s}$, $\boldsymbol{y} - \boldsymbol{s}$ and $\boldsymbol{s}$ respectively. Metropolis algorithm is used to update the parameters.

Hint: At the EM the E-step calculates the posterior expectation, while at the MCMC we simulate merely from it. The M-step in the EM is a maximization while for the MCMC generation form the conditional posterior.

Application of Bivariate Poisson regression model

Champions league data of season 2000/01

The model

$$\begin{aligned}(X,Y)_i &\sim BP(\lambda_{1i},\lambda_{2i},\lambda_{0i})\\ \log(\lambda_{1i}) &= \mu + home + att_{h_i} + def_{g_i}\\ \log(\lambda_{2i}) &= \mu + att_{g_i} + def_{h_i}.\end{aligned}$$

Use of sum-to-zero or corner constraints

Interpretation

- the overall constant parameter specifies λ_1 and λ_2 when two teams of the same strength play on a neutral field.
- Offensive and defensive parameters are expressed as departures from a team of average offensive or defensive ability.

Application of Bivariate Poisson regression model (2)

Modelling the covariance term

$$\log(\lambda_{0i}) = \beta^{con} + \gamma_1 \beta^{home}_{h_i} + \gamma_2 \beta^{away}_{g_i}$$

γ_1 and γ_2 are dummy binary indicators taking values zero or one depending on the model we consider. Hence when $\gamma_1 = \gamma_2 = 0$ we consider constant covariance, when $(\gamma_1, \gamma_2) = (1, 0)$ we assume that the covariance depends on the home team only etc.

Results(1)

Table 2: Details of Fitted Models for Champions League 2000/01 Data ($^1H_0 : \lambda_0 = 0$ and $^2H_0 : \lambda_0 = constant$, B.P. stands for the Bivariate Poisson).

	Model Distribution	Model Details	Log-Lik	Param.	p.value	AIC	BIC
1	Poisson		-432.65	64		996.4	1185.8
		$\underline{\lambda_0}$					
2	Biv. Poisson	constant	-430.59	65	0.042^1	994.3	1186.8
3	Biv. Poisson	Home Team	-414.71	96	0.438^2	1024.5	1311.8
4	Biv. Poisson	Away Team	-416.92	96	0.655^2	1029.0	1316.2
5	Biv. Poisson	Home and Away	-393.85	127	0.151^2	1034.8	1428.8

Results(2)

Table 3:

Home	Away Goals						
	0	1	2	3	4	5	Total
0	10(17.3)	11(10.5)	5(4.2)	3(1.4)	0(0.4)	1(0.1)	30(33.9)
1	20(17.9)	17(14.8)	2(6.8)	3(2.5)	1(0.8)	0(0.2)	43(43.0)
2	14(12.8)	13(11.9)	6(6.1)	2(2.4)	0(0.8)	0(0.2)	35(34.2)
3	10 (7.6)	8 (7.6)	8(4.1)	2(1.7)	0(0.6)	0(0.2)	28(21.8)
4	3 (4.1)	4 (4.2)	3(2.4)	1(1.0)	1(0.4)	0(0.1)	12(12.2)
5	3 (2.0)	2 (2.2)	0(1.3)	1(0.5)	0(0.2)	0(0.1)	6 (6.3)
6	1 (1.0)	1 (1.1)	0(0.6)	0(0.3)	0(0.1)	0(0.0)	2 (3.1)
7	0 (0.4)	0 (0.5)	1(0.3)	0(0.1)	0(0.0)	0(0.0)	1 (1.3)
Total	61 (63.1)	56(52.8)	25 (25.8)	12(9.9)	2 (3.3)	1(0.9)	157 (155.8)*

Comparison of the models

DP : independent Poisson with means 1.1 and 1

BP : Bivariate Poisson with parameters 1, 0.9 and 0.1

score	DP	BP	Result	DP	BP
0-0	0.122	0.135	team A wins	0.376	0.367
1-0	0.134	0.135	draw	0.299	0.318
2-0	0.074	0.067	team B wins	0.324	0.314
0-1	0.122	0.122			
0-2	0.062	0.054			
1-1	0.134	0.135			
2-2	0.037	0.040			
2-1	0.074	0.074			
1-2	0.067	0.066			

Extended models - Inflated models

Empirical results show a problem in estimating the number of draws. The probability of a draw is underestimated. The bivariate Poisson model improves on this point.

Alternative models: Diagonal inflated bivariate Poisson models

Inflated models

- Popular models in the univariate setting. Some specific values have more probability than that predicted by the model, this probability is removed from other points. Very flexible model occur.
- Most common model the zero-inflated model. i.e. the probability of observing a 0 values is larger than what the model predicts.
- Sparse literature in more dimensions (e.g. Li *et al.*, 1999). Inflation only in the $(0, 0)$ cell. Inflation in larger dimensions more difficult to handle.

Diagonal Inflated model

Since the draws are represented in the diagonal of the 2way probability table of the BP model we propose to inflate only the diagonal.

The model:

$$P_D(x,y) = \begin{cases} (1-p)BP(x,y \mid \lambda_1,\lambda_2,\lambda_3), & x \neq y \\ (1-p)BP(x,y \mid \lambda_1,\lambda_2,\lambda_3) + pD(x,\boldsymbol{\theta}), & x = y, \end{cases} \tag{3}$$

where $D(x,\boldsymbol{\theta})$ is discrete distribution with parameter vector $\boldsymbol{\theta}$.

Useful Properties

- Choices for $D(x, \boldsymbol{\theta})$ are the Poisson, the Geometric or simple discrete distributions such as the Bernoulli. The Geometric distribution might be of great interest since it has mode at zero and decays quickly.
- The marginal distributions of a diagonal inflated model are not Poisson distributions but mixtures of distributions with one Poisson component.
- Secondly, if $\lambda_3 = 0$ the resulting inflated distribution introduces a degree of dependence between the two variables under consideration. For this reason, diagonal inflation may correct both overdispersion and correlation problems.
- Model can be fitted using the EM algorithm.

Table 4: Details of Fitted Models for Italian Serie A 1991/92 Data ($^1H_0 : \lambda_3 = 0$, $^2H_0 : \lambda_3 = constant$, $^3H_0 : p = 0.0$, $^4H_0 : \theta_2 = 0.0$ and $^5H_0 : \theta_3 = 0.0$; B.P. stands for the bivariate Poisson, Arrow indicates best fitted model).

	Model Distribution	Additional Model Details	LL	m	p-value	AIC	BIC
1	Poisson		-771.5	36		1614.9	1774.2
		Covariates on λ_3					
2	Bivariate Poisson	constant ($\gamma_1 = \gamma_2 = 0$)	-764.9	37	0.00^1	1603.9	1767.5
3	Bivariate Poisson	Home Team ($\gamma_1 = 1, \gamma_2 = 0$)	-758.9	55	0.84^2	1627.8	1871.1
4	Bivariate Poisson	Away Team ($\gamma_1 = 0, \gamma_2 = 1$)	-755.6	55	0.41^2	1621.2	1864.5
5	Bivariate Poisson	Home and Away ($\gamma_1 = \gamma_2 = 1$)	-745.9	72	0.33^2	1635.7	1954.3
6	Zero Inflated B.P.	constant	-764.9	38	1.00^3	1605.9	1773.9
		Diagonal Distribution					
7	Diag.Inflated B.P.	Geometric	-764.9	39	1.00^3	1607.9	1780.3
→ 8	Diag.Inflated B.P.	Discrete (1)	-756.6	39	0.00^3	1591.1	1763.7
9	Diag.Inflated B.P.	Discrete (2)	-756.6	40	1.00^4	1593.1	1770.1
10	Diag.Inflated B.P.	Discrete (3)	-756.4	41	0.54^5	1594.8	1776.2
11	Diag.Inflated B.P.	Poisson	-763.5	39	0.25^3	1605.1	1777.5
12	Diag.Inflated Poisson	Poisson	-767.0	38	0.01^3	1610.0	1778.1

Table 5: Estimated Parameters for Poisson and Bivariate Poisson Models for 1991/92 Italian Serie A League Data.

		Model 1 Poisson		Model 2 Bivariate Poisson	
	Team	Att	Def	Att	Def
1	Milan	0.68	-0.50	0.84	-1.18
2	Juventus	0.18	-0.50	0.22	-0.70
3	Torino	0.11	-0.60	0.18	-0.86
4	Napoli	0.43	0.12	0.51	0.19
5	Roma	0.00	-0.16	0.02	-0.17
6	Sampdoria	0.02	-0.16	0.10	-0.16
7	Parma	-0.15	-0.27	-0.14	-0.34
8	Inter	-0.29	-0.28	-0.37	-0.29
...	...	...	...	...	...
15	Verona	-0.40	0.43	-0.51	0.57
16	Bari	-0.33	0.24	-0.50	0.33
17	Cremonese	-0.29	0.28	-0.36	0.45
18	Ascoli	-0.34	0.61	-0.64	0.75
	Other Parameters				
	Intercept(μ)	-0.18		-0.57	
	Home	0.36		0.50	
	λ_3	0.00		0.23	
	Mixing Proportion	0.00		0.09	
	θ_1			1.00	

Table 6: Estimated Draws for Every Model (B.P. stands for the bivariate Poisson, Arrow indicates best fitted model).

	Model Distribution	Additional Model Details	0-0	1-1	2-2	3-3	4-4
	Observed Data		38	58	10	4	1
1	Double Poisson		38	33	9	1	0
		Covariates on λ_3					
2	Bivariate Poisson	constant	49	35	11	2	0
3	Bivariate Poisson	Home Team	51	34	11	3	0
4	Bivariate Poisson	Away Team	49	34	11	2	0
5	Bivariate Poisson	Home and Away	47	32	10	2	0
6	Zero Inflated B.P.		49	35	11	2	0
		Diagonal Distribution					
7	Diag.Inflated B.P.	Geometric	49	35	11	2	0
→ 8	Diag.Inflated B.P.	Discrete (1)	43	58	9	2	0
9	Diag.Inflated B.P.	Discrete (2)	43	58	9	2	0
10	Diag.Inflated B.P.	Discrete (3)	43	58	9	3	0
11	Diag.Inflated B.P.	Poisson	50	38	13	3	1
12	Diag.Inflated Poisson	Poisson	45	40	14	3	1

Conclusions for Bivariate Poisson regression models

- The results can be extended to multivariate Poisson regression
- The model can be used for several other disciplines apart form sports
- The data augmentation offers simple estimation via both ML and Bayesian techniques.
- Both algorithms easily programmable to any statistical software

Conclusions for sports modelling

For sports modelling purposes Bivariate Poisson model

- is more realistic,
- improves on the estimation of draws
- can be easily fitted to the data
- allows for other factors that may influence the outcome (e.g. neutral ground, weather conditions, information about players etc

Diagonal inflated models

- Imposes overdispersion and correlation at the same time, so in some sense resolves drawbacks of existing models.

Οι παράμετροι του μοντέλου μετά την 27η αγωνιστική

Constant		-0.168
Home effect		0.340

	attack	defense		attack	defense
ΑΕΚ	0.759	-0.111	ΟΦΗ	0.136	-0.139
Αιγάλεω	-0.272	0.210	Ολυμπιακός	0.700	-0.394
Ακράτητος	-0.037	0.580	Παναχαική	-1.010	0.701
Αρης	0.081	-0.051	Παναθηναικός	0.312	-0.768
Γιάννινα	-0.310	0.149	Πανιώνιος	-0.012	-0.508
Ιωνικός	-0.537	0.090	ΠΑΟΚ	0.512	0.075
Ηρακλής	0.206	0.016	Προοδευτική	-0.235	0.066
Καλλιθέα	-0.155	0.270	Ξάνθη	-0.137	-0.187

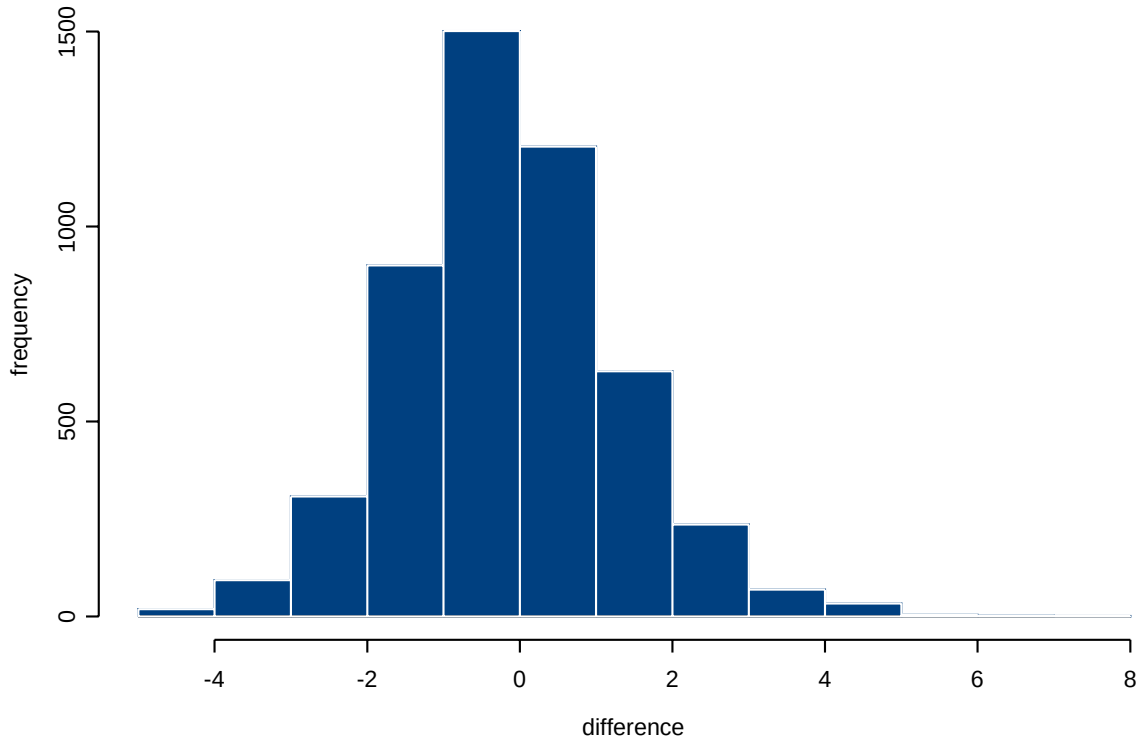
Πιθανότητες νίκης στον αγώνα και στο πρωτάθλημα

Νικητής στον αγώνα		Πρωταθλητής	
Παναθηναϊκος	26.40%	Παναθηναικός	0.72
Ισοπαλία	30.02%	Ολυμπιακός	0.15
Ολυμπιακος	43.58%	ΑΕΚ	0.02
		Ισοβαθμία	0.11

Πιθανότητες κάθε σκορ (Bayesian model)

		Παναθηναικός						
		0	1	2	3	4	5	6+
Ολυμπιακός	0	0.1472	0.1220	0.0454	0.0136	0.0032	0.0002	0.0000
	1	0.1622	0.1224	0.0496	0.0146	0.0048	0.0002	0.0000
	2	0.0890	0.0672	0.0262	0.0076	0.0014	0.0002	0.0002
	3	0.0364	0.0320	0.0104	0.0042	0.0008	0.0002	0.0000
	4	0.0104	0.0090	0.0044	0.0012	0.0002	0.0000	0.0000
	5	0.0054	0.0032	0.0018	0.0002	0.0000	0.0000	0.0000
	6+	0.0008	0.0016	0.0002	0.0000	0.0000	0.0000	0.0000

Histogram of the posterior values for the difference



THE END