Developments in Statistics
Andrej Mrvar and Anuška Ferligoj (Editors)
Metodološki zvezki, 17, Ljubljana: FDV, 2002

A Class of Indices of Equality of a Sport Championship: Definition, Properties and Inference

Maurizio Brizzi[1]

Abstract

This paper deals with the measure of equality of a sport championship, where each participant plays against all the others with a home and away match. We define a class of indices derived by normalising some measures of dispersion. Four of such indices are considered and studied: EQ_1 (based on the standard deviation), EQ_2 (Gini's concentration ratio), EQ_3 (mean absolute deviation) and EQ_4 (mean letter spread). We refer to two extreme schemes: the perfect balance and the completely unbalanced position. The distribution of such indices in 30 European national soccer leagues is studied, jointly with the correlation between the indices. Then, a simulation is made, under the hypothesis that the participants have the same level of skill, and some statistical features of the sample distribution are pointed out. Finally, a Beta model is fitted to the sample distribution, and it seems to be an adequate representative.

1 Introduction

Sport games are an inexhaustible source of data, and in the recent years have been increasingly accompanied by a thorough statistical support. The peculiar rules of each sport competition have generated a great deal of papers and articles on this subject. In these works, a lot of statistical and probabilistic tools have been applied, such as discrete and continuous distributions, stochastic processes, logistic regression, graphical models, and so on. We can mention, for instance, the following contributions, related to some popular sports:

[1] Dipartimento di Scienze statistiche, Via delle Belle Arti 41, I-40126 Bologna, Italy.

- General: Mosteller (1952), David (1959), Glenn (1960), Jackson (1994).
- Athletic: Blest (1996), Morton (1997), Grubb (1998), Cox and Dunn (2002).
- Baseball: Boronico (1999).
- Basketball: Carlin (1996), Schwertmann et al. (1996).
- Cricket: Kimber (1993), Preston and Thomas (2000, 2002).
- Football/Soccer: Maher (1982), Croucher (1984, 1994), Pollard (1986), Nevill *et al.* (1996), Dixon and Coles (1997), Wright (1997), Dixon and Robinson (1998), Rue and Salvesen (2000), Koning (2000).
- Golf: Holder and Nevill (1997), Ketzscher and Ringrose (2002).
- Tennis: Holder and Nevill (1997), Jackson and Mosurski (1997), Magnus and Klaassen (1999a, 1999b).

Let us now focus our attention on a sport championship, in which each participant plays against every other one, with a double match (home and away). Usually, but not always, this pattern is used in team sports rather than in individual ones. We are intended to apply our indices to soccer games, so we will consider the recent rule of assigning a team three points when it wins, one if it draws, no points if it loses. Generally, the championship is considered more interesting when it is well-balanced, and the result of each game is uncertain. We propose here four simple normalised indices of equality: they are equal to one if the championship is perfectly well-balanced, and each team gains the same final number of points; on the other side, each index is equal to zero if the final position is completely unbalanced. This happens when the first classified team wins all the matches, the second classified wins all the matches but two (it loses only when playing against the first team), the third wins all the matches except when playing against the first and the second, and so on, until we consider the last team, which always loses. We will denote this limit scheme with CUP (Completely Unbalanced Position).

Let n be the number of participants and P_j the score of the j-th classified team.Under the 3-1-0 rule, the total number T_n of points depends on the total number of draws. If there are no draws at all, we have three points for each match, so the total score will be:

$$T_n(max)=3n(n-1) \tag{1.1a}$$

On the opposite side, if all the matches finish with a draw, the total score is:

$$T_n(min)=2n(n-1) \tag{1.1b}$$

If the final position is perfectly balanced, each team has the same number of points T_n/n, which is included in the interval $[2(n-1), 3(n-1)]$. Under the CUP, the winner has a final score of $6(n-1)$ points, the second gets $6(n-2)$ points, the j-th gets $6(n-j)$ and so on. We have then two extreme patterns for the final position:

Perfect equality: $P_j = T_n / n, \ \forall j$ (1.2)

Maximum unbalance (CUP): $P_j = 6(n-j), j = 1,2,3\ldots,n$ (1.3)

We propose and develop here three normalised indices of equality, which are equal to zero under (1.3) and to one under (1.2).

2 Indices of equality

It is very well known that a dispersion measure *V*, whose value lies between a minimum *V'* and a maximum *V''*, may be normalised by applying the linear transformation:

$$V^* = \frac{V - V'}{V'' - V'} = \frac{V}{R(V)} - \frac{V'}{R(V)} \tag{2.1}$$

where $R(V) = V'' - V'$ is the maximum range of *V*.

When $V' = 0$, the formula (2.1) becomes simply $V^* = \frac{V}{V''}$.

We will then define a generic index of equality by choosing a measure of dispersion, normalise it by (2.1) and subtract the result from the value 1. If V is a regular measure of dispersion, equal to zero in absence of dispersion, we will define the corresponding index of equality EQ(V):

$$EQ(V) = 1 - \frac{V}{V \max} \tag{2.2}$$

We have considered and studied here four indices of equality belonging to the above defined class (2.2):

$$EQ_1 = 1 - \frac{SD}{\max SD} \tag{2.3}$$

where SD is the standard deviation of the final position.

$$EQ_2 = 1 - \frac{R}{\max R} \tag{2.4}$$

where R is the Gini concentration ratio of the final scores.

$$EQ_3 = 1 - \frac{MAD}{\max MAD} \tag{2.5}$$

where MAD is the mean absolute deviation about the median $\hat{m}$, and finally

$$EQ_4 = 1 - \frac{MLS^*}{\max \mathrm{MLS}^*} \qquad (2.6)$$

where MLS denotes the *mean letter spread*, which is the average of the letter spreads (Hoaglin, 1985), which are the differences of the corresponding letter values ($G^+ - G^-$, $F^+ - F^-$ etc.). We have calculated a modified version of the MLS, in which we include the G letter spread (Brizzi, 2000) and exclude the last but one letter spread, in order to reduce the weight of the extreme observations. We decided to exclude the last but one because it repeats the same information of the previous and following letter spreads.

Example: consider a CUP with n=9: the data are then:

48 - 42 – 36 – 30 – 24 – 18 – 12 – 6 – 0. Letter values and letter spreads are:

H (median): 24

G values: 30 and 18 → G spread = 12

F values: 36 and 12 → F spread = 24

E values: 42 and 6 → E spread = 36

D values: 45 and 3 → D spread = 42

C values: 48 and 0 → C spread = 48.

The mean letter spread is then:

$$MLS = \frac{12+24+36+42+48}{5} = \frac{162}{5} = 32.4$$

The modified MLS is calculated by excluding the D letter spread, which is the half sum of the E and C ones:

$$MLS^* = \frac{12+24+36+48}{4} = \frac{120}{4} = 30 .$$

3 Maximisation and normalisation

If we want to specify the maxima and to give an explicit expression of the indices proposed here, we need to study the behaviour of each index in the extreme situations (1.2) and (1.3). Sometimes the results, as shown below, are slightly different for even and odd values of ***n*****.** It is very easy to check what happens under (1.2): SD, R, MAD and MLS* are all equal to zero, and the corresponding indices of equality are equal to one.

Let now consider the CUP for deriving the maximum. We start with the index EQ1, based on the standard deviation (SD), which is equal to zero under the condition (1.1). We need then to calculate the maximum of SD, which corresponds to (1.2) situation; as said before, under the CUP the average score is 3(n-1).

$$max\ SD = \sqrt{\frac{\sum_{j=1}^{n}[6(n-j)-3(n-1)]^2}{n}} = \sqrt{\frac{\sum_{j=1}^{n}[3(n-2j+1)]^2}{n}} =$$

$$= \sqrt{3(n+1)[3(n+1)-6(n+1)+2(2n+1)]} = \sqrt{3(n+1)(n-1)} \quad (3.1)$$

The index EQ_1 is then defined this way:

$$EQ_1 = 1 - \frac{SD}{\sqrt{3(n+1)(n-1)}} \quad (3.2)$$

We have also to calculate the maximum of R, i.e. the value of the Gini concentration ratio under the maximum unbalanced situation. Let p'1, p'2, …, p'n be the number of points of each team, arranged in an increasing order. The easiest formula for R (Brizzi, 1996, pag. 52) is the following:

$$R = 1 - \frac{\sum_{i=1}^{n-1} q_i}{\sum_{i=1}^{n-1} \frac{i}{n}} = 1 - \frac{2\sum_{i=1}^{n-1} q_i}{n-1}, \text{ where } q_i = \frac{p'_1 + p'_2 + ... + p'_i}{3n(n-1)} \quad (3.3)$$

Under the CUP, the value of q_i is:

$$q_i = \frac{0 + 6 + ... + 6(i-1)}{3n(n-1)} = \frac{3(i-1)i}{3n(n-1)} = \frac{i(i-1)}{n(n-1)}. \quad (3.4)$$

The corresponding value of R is then:

$$max\ R = 1 - \frac{2\sum_{i=1}^{n-1} \frac{i(i-1)}{n(n-1)}}{n-1} = 1 - \frac{2\sum_{i=1}^{n-1} i(i-1)}{n(n-1)^2} =$$

$$= 1 - \frac{2}{n(n-1)^2}\left(\frac{n(n-1)(2n-1)}{6} - \frac{n(n-1)}{2}\right) = 1 - \frac{1}{n-1}\left(\frac{2n-1}{3} - 1\right) =$$

$$= 1 - \frac{1}{n-1}\left(\frac{2(n-2)}{3}\right) = 1 - \frac{2}{3}\left(\frac{n-2}{n-1}\right) = \frac{n+1}{3(n-1)}. \quad (3.5)$$

The index of equality based on R is therefore:

$$\text{EQ}_2 = 1 - \frac{R}{\max R} = 1 - \frac{3(n-1)}{n+1} R \tag{3.6}$$

The third index is based on the MAD. Under the situation (1.3), the median is equal to the average (the CUP is symmetric), i.e. to 3(n-1). The maximum value of MAD differs, depending on n being even or odd.

Let n be even, then n = 2h, with h a positive integer. We have then:

$$max\ MAD = \frac{1}{2h}\sum_{j=1}^{2h}\left|6(2h-j)-3(2h-1)\right| = \frac{3}{2h}\sum_{j=1}^{2h}\left|2h-2j+1\right| =$$

$$= \frac{3}{2h}\sum_{j=1}^{h}(2h-2j+1) + \frac{3}{2h}\sum_{j=h+1}^{2h}(2j-2h-1) = \frac{3}{2h}\left[\sum_{j=h+1}^{2h} 2j - \sum_{j=1}^{h} 2j\right] =$$

$$= \frac{3}{2h}\left[\sum_{j=1}^{2h} 2j - 2\sum_{j=1}^{h} 2j\right] = \frac{3}{2h}\left[2h(2h-1) - 2h(h-1)\right] = 3(2h\text{-}1)\text{-}3(h\text{-}1) = 3h = \frac{3}{2}n \tag{3.7}$$

Let now n be odd, then n = 2h+1, with h a positive integer. The maximum value of MAD becomes:

$$max\ MAD = \frac{1}{2h+1}\sum_{j=1}^{2h+1}\left|6(2h+1-j)-6h\right| = \frac{6}{2h+1}\sum_{j=1}^{2h+1}\left|h+1-j\right| =$$

$$= \frac{6}{2h+1}\left[\sum_{j=1}^{h}(h+1-j) - \sum_{j=h+2}^{2h+1}(h+1-j)\right] = \frac{6}{2h+1}\left[\sum_{j=h+2}^{2h+1} j - \sum_{j=1}^{h} j\right] =$$

$$= \frac{6}{2h+1}\left[\sum_{j=1}^{2h+1} j - \sum_{j=1}^{h+1} j - \sum_{j=1}^{h} j\right] = \frac{6}{2h+1}\left[\frac{(2h+1)(2h+2)}{2} - \frac{(h+1)(h+2)}{2} - \frac{h(h+1)}{2}\right] =$$

$$= \frac{3(h+1)}{2h+1}\left[2(2h+1)-(h+2)-h\right] = \frac{3h(2h+2)}{2h+1} = \frac{3}{2}\left(n - \frac{1}{n}\right) \tag{3.8}$$

The index EQ3 is then defined this way:

$$\text{EQ}_3 = 1 - \frac{MAD}{\max MAD} = \begin{cases} 1 - \dfrac{3 \cdot MAD}{2n} & \text{if } n \text{ is even} \\[2ex] 1 - \dfrac{3 \cdot MAD}{2\left(n - \frac{1}{n}\right)} & \text{if } n \text{ is odd} \end{cases} \tag{3.9}$$

Finally, we considered the fourth index EQ_4. Since it is not easy to find an univoque expression for the maximum, we have calculated the value of the modified MLS (indicated with MLS*) under the CUP, for *n* between 10 and 20, which is the range that includes the value of *n* in the great majority of actual sport championships. The resulting maxima are reported in the following table:

Table 1: Maximum value of MLS* for each value of n between 10 and 20.

N	10	11	12	13	14	15	16	17	18	19	20
max MLS*	36	39	43.5	45	51	54	58.5	64.8	70.8	74.4	79.2

Table 2: National Soccer Leagues 1998/99: indices of equality.

Country	n	EQ_1	Rank	EQ_2	Rank	EQ_3	Rank	EQ_4	Rank
Sweden	14	0.730	1	0.707	1	0.755	1	0.743	1
Czech Rep.	16	0.658	2	0.638	2	0.698	2	0.675	2
Germany	18	0.657	3	0.544	10	0.690	3	0.588	9
Italy	18	0.622	4	0.605	3	0.656	4	0.630	3
Spain	20	0.617	5	0.586	4	0.615	9	0.615	5
England	20	0.606	6	0.583	5	0.652	6	0.619	4
Switzer.	12	0.596	7	0.553	8	0.611	10	0.612	6
Russia	16	0.583	8	0.580	6	0.656	5	0.611	7
France	18	0.579	9	0.562	7	0.623	7	0.602	8
Wales	17	0.574	10	0.474	14	0.592	13	0.522	16
Poland	16	0.552	11	0.547	9	0.620	8	0.583	10
Ireland	12	0.529	12	0.519	12	0.593	12	0.560	11
Turkey	18	0.524	13	0.523	11	0.597	11	0.554	12
Portugal	18	0.511	14	0.474	15	0.556	15	0.531	13
Hungary	18	0.507	15	0.482	13	0.549	16	0.528	14
Holland	18	0.504	16	0.464	17	0.504	19	0.508	17
Belgium	18	0.487	17	0.450	18	0.494	20	0.497	18
Norway	14	0.479	18	0.471	16	0.558	14	0.522	15
Israel	16	0.444	19	0.434	20	0.516	17	0.474	20
Ukraine	16	0.439	20	0.410	21	0.479	21	0.451	22
Croatia	12	0.438	21	0.445	19	0.505	18	0.489	19
Macedonia	14	0.437	22	0.408	22	0.459	24	0.456	21
Belarus	15	0.433	23	0.400	23	0.461	22	0.440	24
Bulgaria	16	0.403	24	0.395	24	0.461	23	0.440	23
Georgia	16	0.394	25	0.359	26	0.443	25	0.406	25
Greece	18	0.390	26	0.365	25	0.424	26,5	0.401	26
Romania	18	0.367	27	0.351	27	0.424	26,5	0.373	27
Slovakia	16	0.365	28	0.325	28	0.385	28	0.370	28
Luxemb.	12	0.344	29	0.319	29	0.310	29	0.336	29
Cyprus	14	0.255	30	0.232	30	0.282	30	0.314	30

4 Application to European soccer data

We have applied the indices of equality EQ_1, EQ_2, EQ_3, EQ_4 to a set of soccer data, and more precisely to the final positions of 30 European National Soccer Leagues (including Cyprus, Israel and Turkey) in 1998/99 season. In Table 2 we have reported the name of the Country, the number of teams *n*, the equality indices and the corresponding rank (1 for the first, 30 for the last).

Swedish national league (14 teams) seems to be the most balanced, since it shows the highest value of all the indices, followed by the Czech Republic (16 teams), while the most unbalanced national leagues, with respect to all the indices, are held in Cyprus and Luxembourg

We have then studied the degree of association of the indices in this set of 30 data, by computing the Bravais-Pearson correlation coefficient r and the Spearman rank correlation index r_S between each pair of indices, in order to verify the degree of consistency between the indices. We have obtained the following correlation matrices:

Table 3: Linear correlation between the indices of equality (Bravais-Pearson coefficient).

	EQ_1	EQ_2	EQ_3	EQ_4
EQ_1	*1.000*	*0.973*	*0.973*	*0.973*
EQ_2		*1.000*	*0.974*	*0.994*
EQ_3			*1.000*	*0.975*
EQ_4				*1.000*

Table 4: Rank correlation between the indices of equality (Spearman coefficient).

	EQ_1	EQ_2	EQ_3	EQ_4
EQ_1	*1.000*	*0.977*	*0.974*	*0.976*
EQ_2		*1.000*	*0.972*	*0.994*
EQ_3			*1.000*	*0.970*
EQ_4				*1.000*

Therefore, the information given by the four indices about the degree of equality is quite similar. In particular, EQ2 and EQ4 have a correlation very close to one; so, they seem to be almost equivalent. The remaining 5 pairs of indices show almost the same level of correlation, about 0.975 (i.e. 39/40).

5 Sample distribution of the indices

Finally, we tried to study the sample distribution of the indices defined above**.** We simulated with GAUSS package 25,000 soccer championships for each value of k and for each index of equality. We worked under the simple hypothesis that all the participants have the same level of skill, and took into consideration the "home advantage", which is very relevant in European soccer, giving a probability of 5/8 to the event "home team wins", 1/4 to the event "draw match", 1/8 to the event "guest team wins". These probabilities are quite suitable to the real soccer world, at least according to our set of data.

5.1 Index EQ1 based on standard deviation

In Table 5 we show some statistical features of the simulated sample distribution of the index of equality EQ_1, for some values of n between 10 and 20 (all the n's reported in Table 1 are included in this range): average, standard deviation, median and some tail percentiles.

Table 5: Simulated sample distribution of the index of equality EQ_1**.**

	Aver.	StD.	Median	Left tail percentiles			Right tail percentiles		
n	**$M_1(n)$**	**$S_1(n)$**	**$Me_1(n)$**	**1%**	**2.5%**	**5%**	**95%**	**97.5%**	**99%**
10	*0.7413*	*0.0615*	*0.7438*	*0.587*	*0.614*	*0.636*	*0.838*	*0.854*	*0.872*
12	*0.7613*	*0.0511*	*0.7627*	*0.637*	*0.658*	*0.675*	*0.842*	*0.857*	*0.873*
13	*0.7695*	*0.0473*	*0.7712*	*0.654*	*0.672*	*0.690*	*0.845*	*0.858*	*0.872*
14	*0.7766*	*0.0438*	*0.7781*	*0.670*	*0.687*	*0.702*	*0.847*	*0.859*	*0.872*
15	*0.7838*	*0.0413*	*0.7851*	*0.681*	*0.699*	*0.714*	*0.850*	*0.861*	*0.873*
16	*0.7897*	*0.0387*	*0.7912*	*0.695*	*0.711*	*0.724*	*0.851*	*0.862*	*0.874*
18	*0.8010*	*0.0342*	*0.8018*	*0.717*	*0.731*	*0.744*	*0.856*	*0.865*	*0.877*
20	*0.8103*	*0.0308*	*0.8110*	*0.736*	*0.747*	*0.758*	*0.859*	*0.868*	*0.878*

We can use the tail percentiles of Table 5 as critical values for testing the null hypothesis that all the participants have the same technical level, and consequently the probability of winning a game depends only on the home factor. The simulated sample distribution of EQ_1 is plotted in Figure 1.

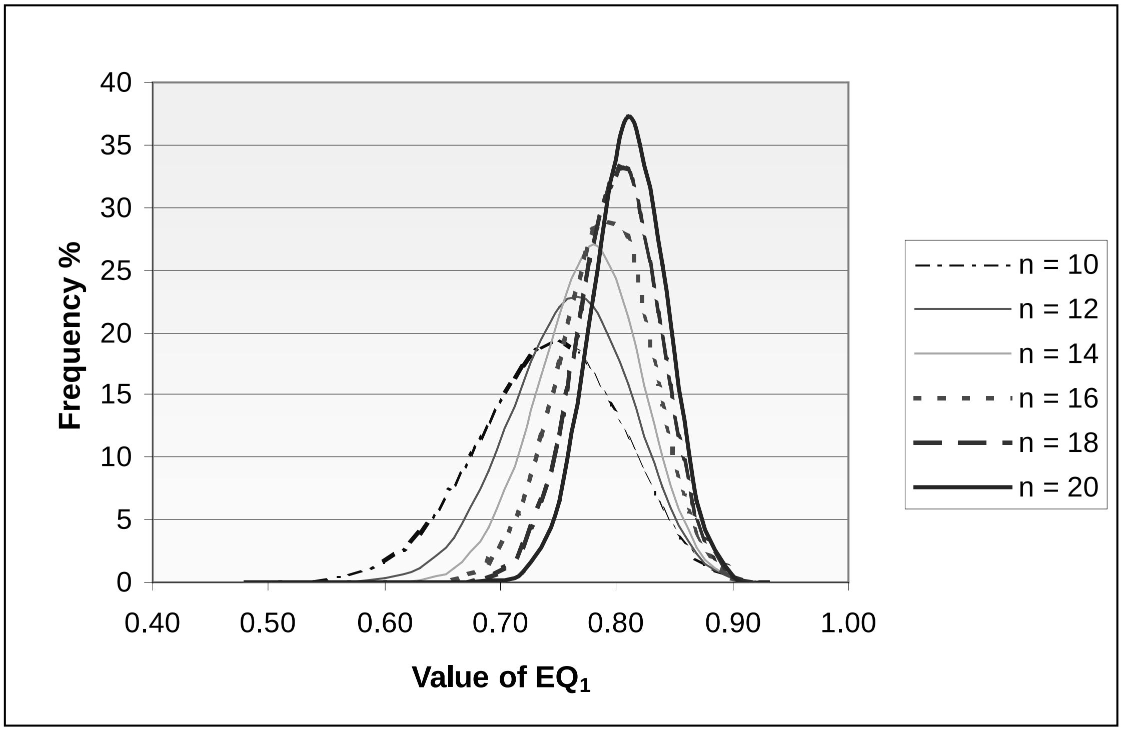

Figure 1: Sample distribution of EQ_1.

5.2 Index EQ_2 based on Gini concentration ratio

In Table 6 we give some statistical features of the simulated sample distribution of the index of equality EQ_2, for some values of *n* between 10 and 20 (all the *n*'s reported in Table 1 are included in this range): average, standard deviation, median and some tail percentiles.

Table 6: Simulated sample distribution of the index of equality EQ_2**.**

	Aver.	StD	Median	Left tail percentiles			Right tail percentiles		
n	**$M_1(n)$**	**$S_1(n)$**	**$Me_1(n)$**	**1%**	**2.5%**	**5%**	**95%**	**97.5%**	**99%**
10	*0.7298*	*0.0646*	*0.7317*	*0.570*	*0.596*	*0.619*	*0.832*	*0.850*	*0.868*
12	*0.7508*	*0.0541*	*0.7527*	*0.618*	*0.640*	*0.658*	*0.836*	*0.850*	*0.865*
13	*0.7583*	*0.0498*	*0.7602*	*0.634*	*0.655*	*0.673*	*0.837*	*0.850*	*0.865*
14	*0.7656*	*0.0469*	*0.7675*	*0.650*	*0.669*	*0.685*	*0.839*	*0.852*	*0.866*
15	*0.7724*	*0.0435*	*0.7736*	*0.665*	*0.684*	*0.698*	*0.841*	*0.854*	*0.867*
16	*0.7789*	*0.0411*	*0.7804*	*0.679*	*0.696*	*0.710*	*0.844*	*0.855*	*0.869*
18	*0.7897*	*0.0366*	*0.7909*	*0.698*	*0.715*	*0.727*	*0.848*	*0.858*	*0.869*
20	*0.8001*	*0.0333*	*0.8012*	*0.718*	*0.732*	*0.743*	*0.853*	*0.862*	*0.872*

We can also use the tail percentiles of Table 5 as critical values for testing the null hypothesis that all the participants have are at the same level, and consequently the probability of winning a game depends only on the home factor. The simulated sample distribution of EQ_1 is plotted in Figure 2.

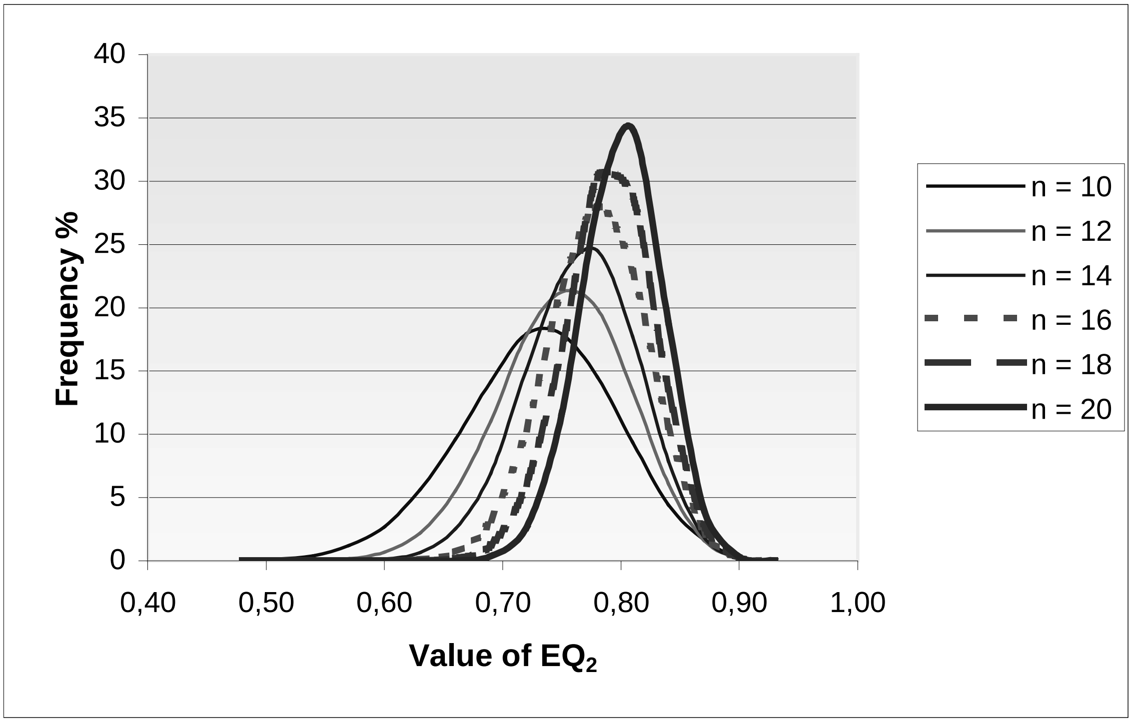

Figure 2: Sample distribution of EQ_2.

5.3 Index EQ_3 based on mean absolute deviation

In Table 7 we report some statistical features of the simulated sample distribution of the index of equality EQ_2, for some values of *n* between 10 and 20 (all the *n*'s reported in Table 1 lie in this range): average, standard deviation, median and some tail percentiles.

As done before, we can use the tail percentiles of Table 7 as critical values for testing the null hypothesis that all the participants have the same technical level, and consequently the probability of winning a game depends only on the home factor. The simulated sample distribution of EQ_3 is plotted in Figure 3.

Table 7: Simulated sample distribution of the index of equality EQ_3.

	Aver.	StD	Median	Left tail percentiles			Right tail percentiles		
n	$M_1(n)$	$S_1(n)$	$Me_1(n)$	1%	2.5%	5%	95%	97.5%	99%
10	*0.6443*	*0.0900*	*0.6500*	*0.420*	*0.450*	*0.490*	*0.780*	*0.810*	*0.830*
12	*0.6708*	*0.0750*	*0.6736*	*0.486*	*0.514*	*0.542*	*0.785*	*0.806*	*0.826*
13	*0.6813*	*0.0701*	*0.6845*	*0.500*	*0.536*	*0.560*	*0.792*	*0.810*	*0.827*
14	*0.6924*	*0.0645*	*0.6939*	*0.525*	*0.556*	*0.582*	*0.791*	*0.811*	*0.826*
15	*0.7014*	*0.0604*	*0.7054*	*0.549*	*0.576*	*0.598*	*0.795*	*0.808*	*0.826*
16	*0.7102*	*0.0563*	*0.7109*	*0.570*	*0.594*	*0.613*	*0.797*	*0.812*	*0.828*
18	*0.7251*	*0.0507*	*0.7284*	*0.599*	*0.620*	*0.639*	*0.806*	*0.818*	*0.833*
20	*0.7386*	*0.0453*	*0.7400*	*0.625*	*0.645*	*0.660*	*0.810*	*0.822*	*0.838*

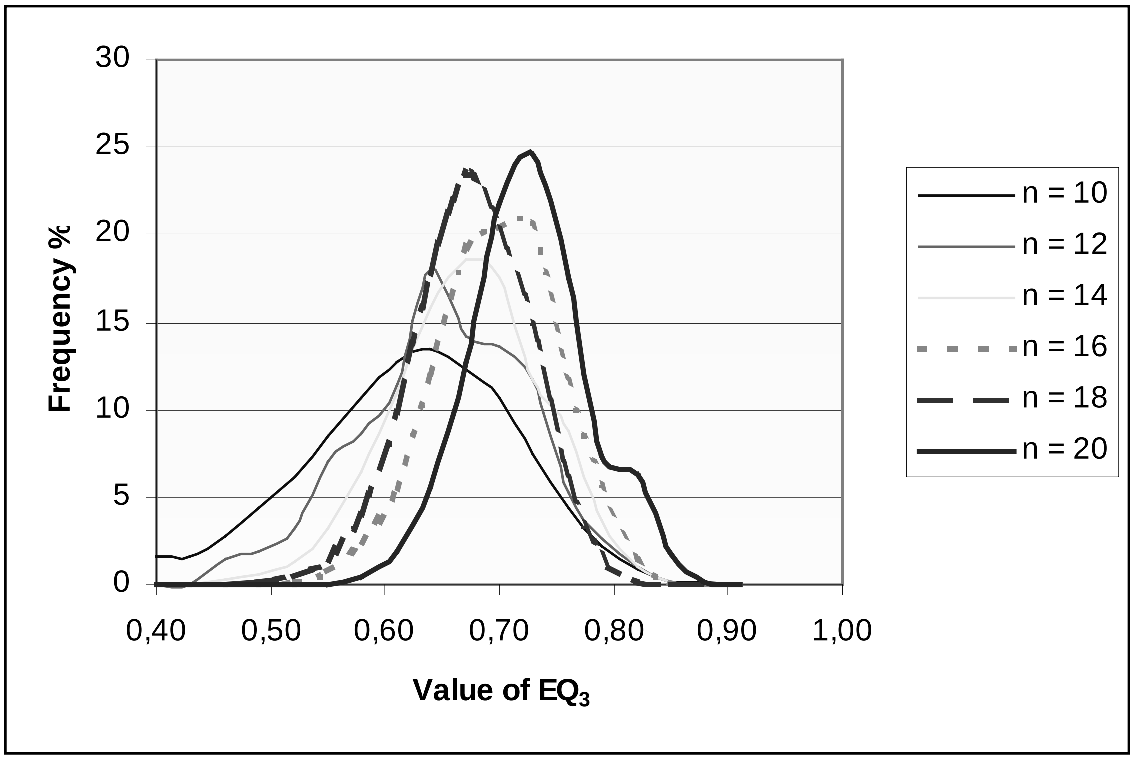

Figure 3: Sample distribution of EQ_3.

5.4 Index EQ_4 based on mean letter spread

We have given in Table 8 some statistical features of the simulated sample distribution of the index of equality EQ_2, for some values of *n* between 10 and 20 (all the *n*'s reported in Table 1 are included in this range): average, standard deviation, median and some tail percentiles.

Table 8: Simulated sample distribution of the index of equality EQ_4.

n	Aver. $M_1(n)$	StD $S(n)$	Median $Me_1(n)$	Left tail percentiles 1%	2.5%	5%	Right tail percentiles 95%	97.5%	99%
10	*0.7606*	*0.0595*	*0.7639*	*0.611*	*0.639*	*0.660*	*0.854*	*0.868*	*0.882*
12	*0.7741*	*0.0502*	*0.7759*	*0.649*	*0.670*	*0.690*	*0.853*	*0.868*	*0.879*
13	*0.7822*	*0.0461*	*0.7833*	*0.664*	*0.686*	*0.703*	*0.856*	*0.868*	*0.881*
14	*0.7878*	*0.0430*	*0.7892*	*0.681*	*0.699*	*0.716*	*0.855*	*0.868*	*0.880*
15	*0.7930*	*0.0404*	*0.7940*	*0.692*	*0.711*	*0.725*	*0.856*	*0.868*	*0.880*
16	*0.7972*	*0.0379*	*0.7970*	*0.703*	*0.720*	*0.733*	*0.857*	*0.868*	*0.880*
18	*0.8079*	*0.0338*	*0.8079*	*0.726*	*0.740*	*0.751*	*0.862*	*0.870*	*0.881*
20	*0.8142*	*0.0307*	*0.8150*	*0.739*	*0.752*	*0.762*	*0.864*	*0.872*	*0.883*

We can use the tail percentiles of Table 8 as critical values for testing the null hypothesis that all the participants have the same technical level, and consequently the probability of winning a game depends only on the home factor. The simulated sample distribution of EQ_4 is plotted in Figure 4.

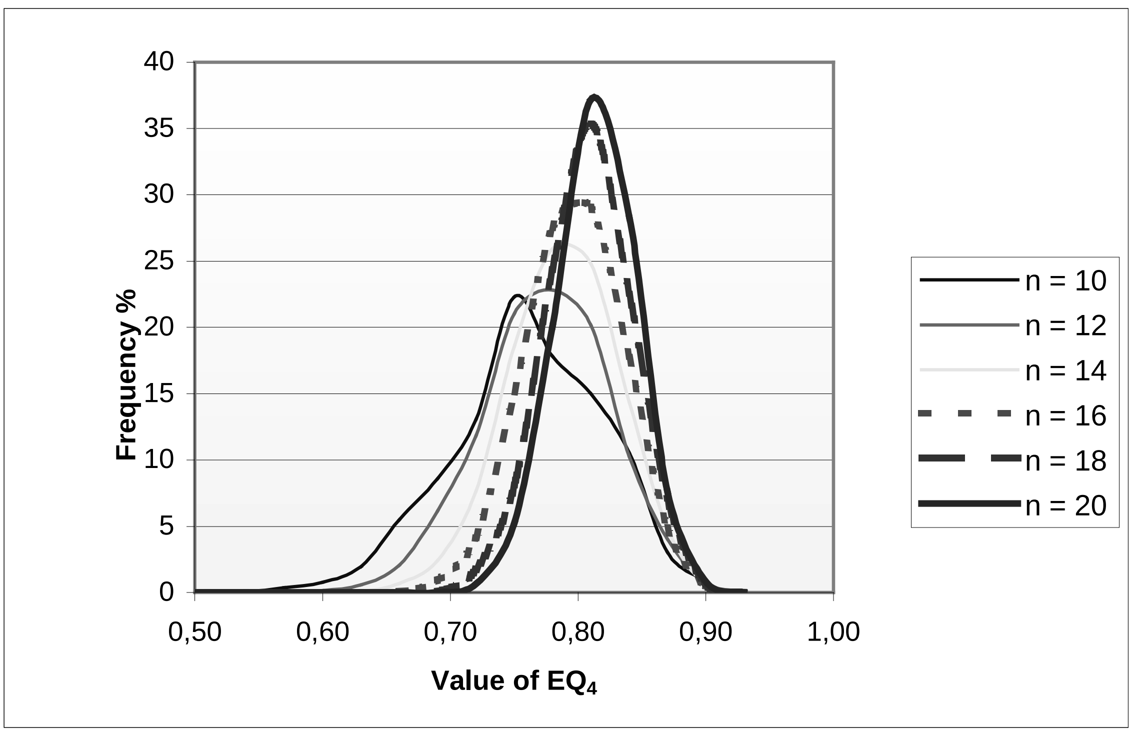

Figure 4: Sample distribution of EQ_4.

We have finally tried to see what would happen without the "home advantage": in Table 9 we have shown the results of the index EQ_1 by giving the same probability to each possible outcome (home wins, draw match, guest wins).

Table 9: Simulated sample distribution of the index of equality EQ_1.

	Aver.	StD	Median	Left tail percentiles			Right tail percentiles		
n	**$M_1(n)$**	**$S_1(n)$**	**$Me_1(n)$**	**1%**	**2.5%**	**5%**	**95%**	**97.5%**	**99%**
10	*0.7023*	*0.0686*	*0.7044*	*0.534*	*0.561*	*0.586*	*0.812*	*0.830*	*0.849*
12	*0.7250*	*0.0578*	*0.7270*	*0.581*	*0.606*	*0.627*	*0.817*	*0.833*	*0.851*
13	*0.7335*	*0.0533*	*0.7353*	*0.604*	*0.625*	*0.643*	*0.818*	*0.834*	*0.849*
14	*0.7421*	*0.0499*	*0.7443*	*0.620*	*0.639*	*0.657*	*0.821*	*0.835*	*0.849*
15	*0.7502*	*0.0466*	*0.7516*	*0.639*	*0.656*	*0.671*	*0.825*	*0.838*	*0.852*
16	*0.7572*	*0.0437*	*0.7585*	*0.650*	*0.669*	*0.684*	*0.828*	*0.840*	*0.852*
18	*0.7697*	*0.0393*	*0.7710*	*0.674*	*0.689*	*0.703*	*0.832*	*0.843*	*0.855*
20	*0.7810*	*0.0353*	*0.7818*	*0.698*	*0.710*	*0.721*	*0.838*	*0.847*	*0.859*

If we compare Table 5 and Table 9 we can notice that the general trend does not seem to change dramatically. The average values are proportionally lower when eliminating the home advantage (about 5% less), while the standard deviation increases (from 12% to 15% depending on *n*), and this is not surprising if we consider that the probabilities (1/3, 1/3, 1/3) given in Table 9 to the possible outcomes of a game are more heterogeneous than (5/8, 2/8, 1/8) given before. The same happens with the other indices.

5.5 Analysis and comparison of the sample distributions

Looking at the results, we have observed that the sample average of each index of equality increases with *n* following an approximately linear pattern, while the sample standard deviation is almost perfectly proportional to *n*. Actually, if we fit the recorded sample average of each index with a least squares straight line, we have the following results:

$M_1^*(n) = 0{,}6793 + 0{,}0068\, n \quad (r=+\,0{,}989)$ (5.1)
$M2^*(n) = 0{,}6672 + 0{,}0069\, n \quad (r = +\,0{,}990)$ (5.2)
$M3^*(n) = 0{,}5585 + 0{,}0093\, n \quad (r = +\,0{,}991)$ (5.3)
$M4^*(n) = 0{,}7100 + 0{,}0054\, n \quad (r = +\,0{,}990)$, (5.4)

where $M_i^*(n)$, $i = 1, 2, 3, 4$ are the theoretical mean values, which are actually not far from the observed ones, and the values of r_i , all very close to one, give us even more evidence that the relation between n and M_i *(n)* may be satisfactorily represented with a linear model.

Referring now to the sample standard deviation $S_i(n)$ of the simulated values, we noticed that the product $n \cdot S_i(n)$ is approximately constant for each i. We can then define a simple but satisfactory approximation for the standard deviation , by using the following expressions:

$$S_1*(n)=\frac{0.616}{n} \cong \frac{8}{13n}. \tag{5.5}$$

$$S_2*(n)=\frac{0.655}{n} \cong \frac{15}{23n} \tag{5.6}$$

$$S_3*(n)=\frac{0.905}{n} \cong \frac{19}{21n} \tag{5.7}$$

$$S_4*(n)=\frac{0.605}{n} \cong \frac{17}{28n} \tag{5.8}$$

Finally, if we consider the tail percentiles shown in Tables 5, 6, 7, 8 as critical values for a statistical test where the null hypothesis is that all the teams have the same probability to win a game, the only country for which all the indices of equality lead us to keep the null hypothesis is Sweden.

6 Fitting the sample distribution with a Beta model

Now, we have tried to fit the sample distributions to the standard two-parameters Beta model:

$$f(y)=\frac{y^{p-1}(1-y)^{q-1}}{B(p,q)},\ p>0,\ q>0,\ 0<y<1. \tag{6.1}$$

The normalising constant $B(p,q)$ in (6.1) is equal to:

$$B(p,q)=\int_0^1 y^{p-1}(1-y)^{q-1}dy=\frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)} \tag{6.2a}$$

If p and q are both integer numbers, we have:

$$B(p,q)=\frac{(p-1)!(q-1)!}{(p+q-1)!} \tag{6.2b}$$

We considered even values of n from 10 to 20, estimated p and q with the method of moments, rounded the result to the nearest half, in order to have an easier task in calculating the Beta operator, and reported the estimates $\hat{p}$, $\hat{q}$ in Table 10. The estimators are:

$$\hat{\alpha} = \frac{\overline{y}(\overline{y} - m_2)}{s^2}; \ \hat{\beta} = \frac{(1-\overline{y})(\overline{y} - m_2)}{s^2} \qquad (6.3)$$

where $\overline{y}$ is the sample average, s^2 is the sample variance and m_2 is the sample moment of the second order ($m_2 = \overline{y}^2 + s^2$).

Table 10: Estimates of the Beta parameters for different values of n.

	EQ_1		EQ_2		EQ_3		EQ_4	
n	$\hat{p}$	$\hat{q}$	$\hat{p}$	$\hat{q}$	$\hat{p}$	$\hat{q}$	$\hat{p}$	$\hat{q}$
10	37	13	34	12.5	18	10	38	12
12	51	16	48	16	25,5	12,5	53	15.5
14	70	20	62	19	34	15	70.5	19
16	86.5	23	79	22.5	45.5	18.5	90	23
18	109	27	97.5	26	55.5	21	109	26
20	132	31	116	29	68	24	131	30

Looking at Table 10, we notice that the estimated values of the Beta parameters increase almost linearly as *n* increases, and that the distribution is almost equal for the three indices. We have plotted the empirical frequency function jointly with the correspondent Beta theoretical function, and we saw that it fits very well. In the following figures we have represented the empirical and theoretical distribution for EQ_1 and $k=10$ (Figure 5), for EQ_2 and $k=12$ (Figure 6), for EQ_3 and $k=14$ (Figure 7), for EQ_4 and $k=16$ (Figure 8). Looking at the figures, it is evident that Beta model can be considered quite appropriate for the sample distribution of the indices, and that estimation based on moments gives in this context very good results.

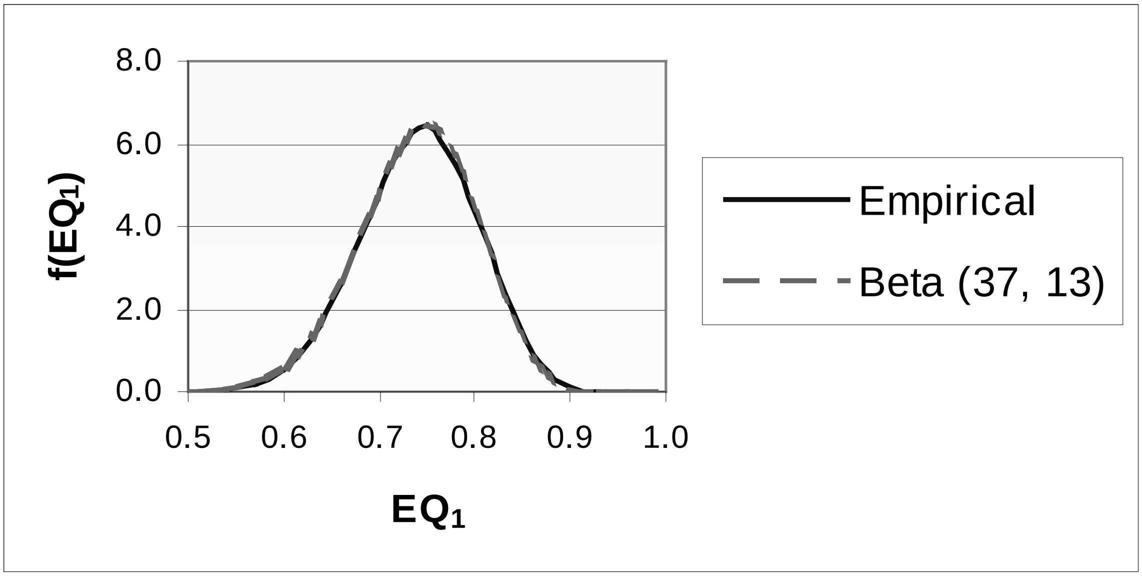

Figure 5: Empirical and theoretical Beta d.f.: EQ_1 with n=10.

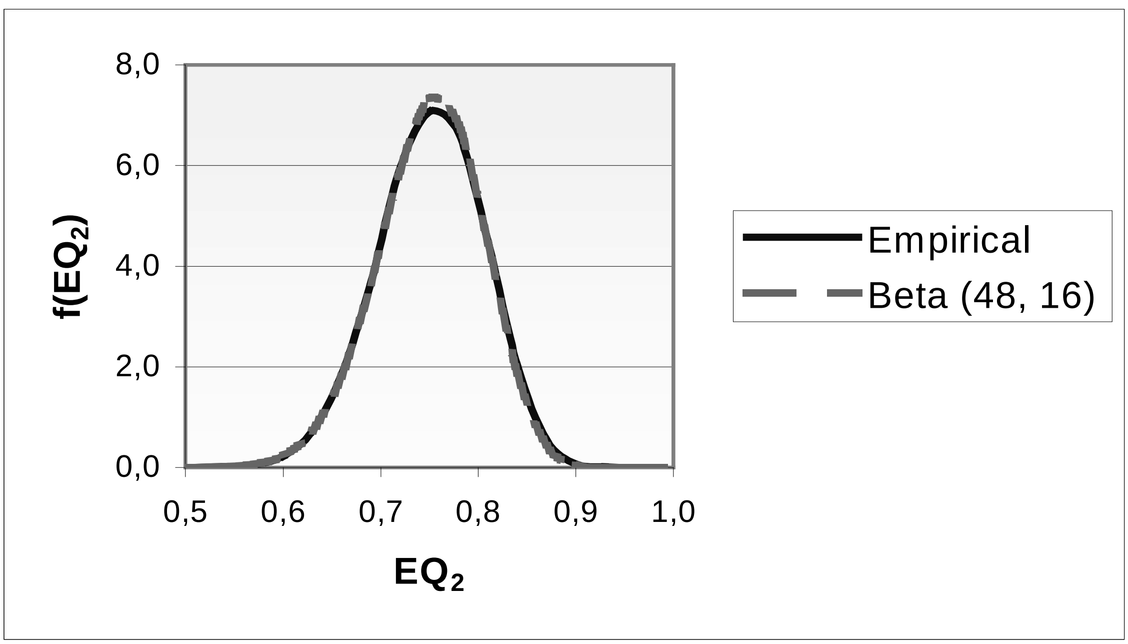

Figure 6: Empirical and theoretical Beta d.f.: EQ_2 with n=12.

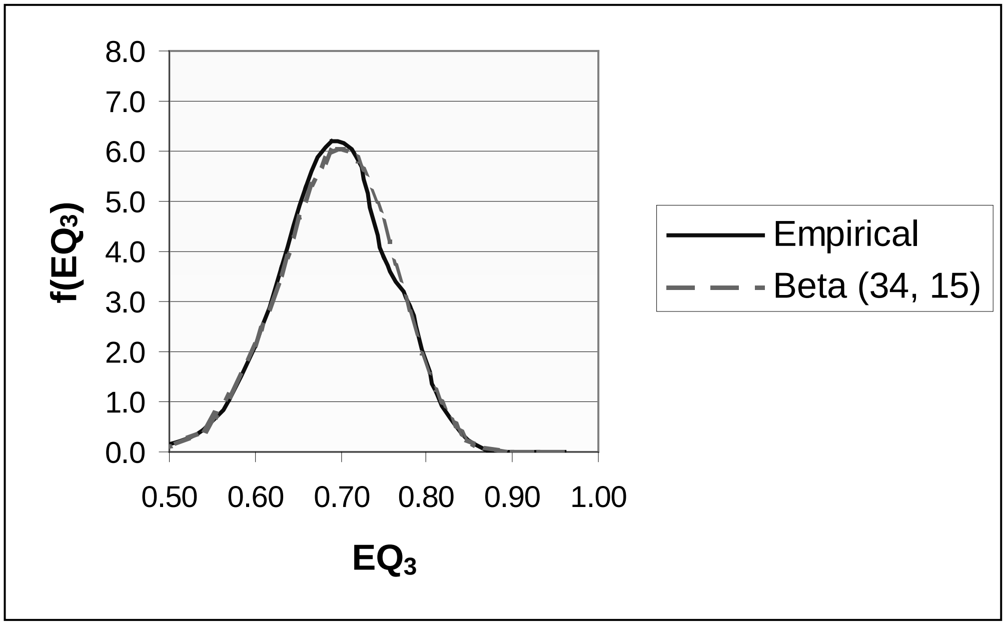

Figure 7: Empirical and theoretical Beta d.f.: EQ_3 with n=14.

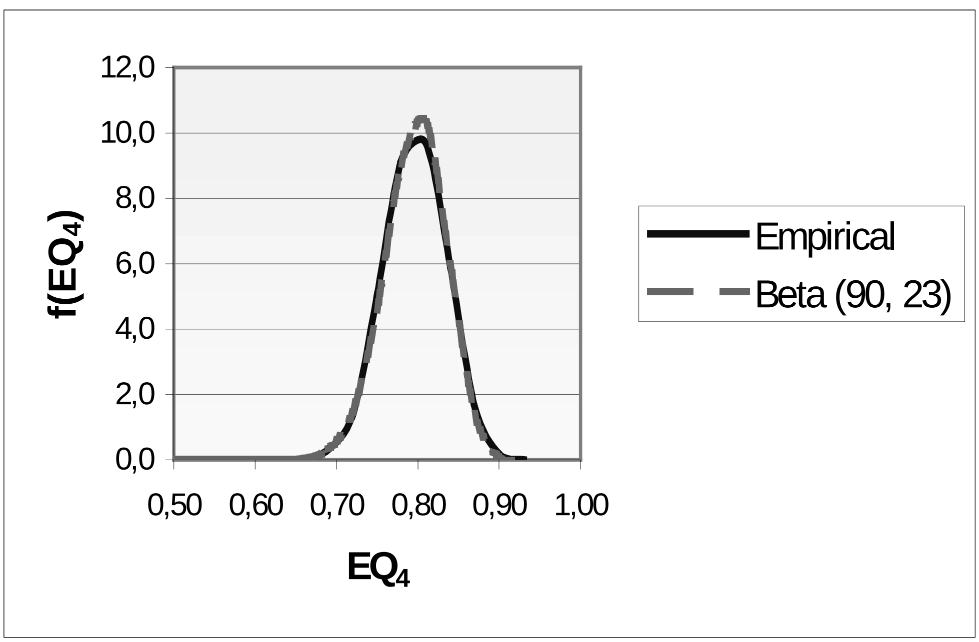

Figure 8: Empirical and theoretical Beta d.f.: EQ_3 with n=16.

7 Concluding remarks

Considering the results described above, we can conclude that all the indices of equality give similar (and strongly correlated) results in measuring the equality of a sport championship. Since we are looking for a “quick” index, perhaps EQ_1 and EQ_3 are preferable, being easier to calculate. The soccer data show that a “good level” of equality is reached when the indices are greater than 0.6, and this is frequently associated with a high technical level: countries like Spain, Italy, France, England, Germany, where the technical level is outstanding, show high values for all the indices. On the other side, countries which do not have a good soccer tradition, like Cyprus and Luxembourg, show the lowest values, like recent countries such as Belarus and Slovakia, whose best teams participated to other leagues in the past.

The simulated sample distribution of each of the indices, under the condition of equal level of all the competing teams, shows increasing values of the sample mean as n increases, while the standard deviation decreases. The link between n and the sample mean is well fitted by a straight line, see (5.1),(5.2),(5.3),(5.4); the sample standard deviation is approximately proportional to inverse of n, as shown in (5.4), (5.5) and Table 9. The tail percentiles indicated in Tables 5, 6, 7, 8 may be used as critical values of a test of significance (the null hypothesis is the perfect equality of the level). The Beta model, whose parameters may be estimated with a simple method (like the method of moments) seems very suitable for representing the sample distribution, as shown in Chapter 6.

The next step on this research topic should be an extension of the study of the sample distribution, under different conditions (not only the equal level), and/or an analytical approach to this study, beginning with a small number of participants to facilitate the analysis; other indices of equality may be proposed as well, by considering other measures of dispersion and applying (2.2). A possible extension of this procedure may be as well to study the equality of the participants by taking into account not only the number of points, but also the number of goals scored in each match: indeed, scores like 2-1 and 7-0, yield the same number of points to both teams in the final position, but give a completely different impression about their level of skill.

References

[1] Blest, D.C. (1996): Lower bounds for athletic performances. *The Statistician*, **45**, 243-253.

[2] Boronico, J.S. (1999): Multi-tiered playoffs and their impact on professional baseball. *The American Statistician*, **53**, 56-61.

[3] Brizzi, M. (1996): *Le misure relative di variabilità*. Bologna: Società Editrice Esculapio.

[4] Brizzi, M. (2000): Detecting skewness and kurtosis by letter values: a new proposal. *Statistica*, **LX**, 243 – 258.

[5] Carlin, B.P. (1996): Improved NCAA Basketball tournament modeling via point spread and team strength information. *The American Statistician*, **50**, 39-43.

[6] Cox, T.F. and Dunn, R.T.(2002): An analysis of decathlon data. *The Statistician*, 51, 179-187.

[7] Croucher, J. (1984): The effect of changing competition points in the English Football League. *Teaching Statistics*, **6**, 39-42**.**

[8] Croucher, J. (1994): Ranking soccer teams. *Teaching Statistics*, **16**, 90-92**.**

[9] David, H.A. (1959): Tournaments and paired comparisons. *Biometrika*, **46**, 139-149.

[10] Dixon, M.J. and Coles, S.G. (1997): Modelling association football scores and inefficiencies in the football betting market. *Applied Statistics*, **46**, 265-280.

[11] Dixon, M.J. and Robinson, M.E. (1998): A birth process for association football matches. *The Statistician*, **47**, 523-538.

[12] Glenn, W.A. (1960): A comparison of the effectiveness of tournaments. *Biometrika*, **47**, 253-262.

[13] Grubb, H.J. (1998): Models for comparing athletic performances. *The Statistician*, **47**, 509-521.

[14] Hoaglin, D.C., Mosteller, F., and Tukey, J.W. (1985): *Exploring Data Tables, Trends and Shapes*. New York: John Wiley & Sons; Chapter 10 by D.C. Hoaglin.

[15] Holder, R.L. and Nevill, A.M. (1997): Modelling preformance at International Tennis and Golf tournaments: is there a home advantage? *The Statistician*, **46,** 551-559.

[16] Jackson D.A. (1994): Index betting on sports. *The Statistician*, **43**, 309-315.

[17] Jackson, D.A. and Mosurski, K. (1997): Heavy defeats in tennis: psychological momentum or random effect? *Chance*, **10**, 27-34.

[18] Ketzscher, R. and Ringrose, T.J. (2002): Exploratory analysis of European Professional Golf Association statistics. *The Statistician*, **51**, 215-228.

[19] Kimber, A.C. (1993): A graphical display for comparing bowlers in cricket. *Teaching Statistics*, **15**, 84-86.

[20] Koning, R.H. (2000): Balance in competition in Dutch soccer. *The Statistician*, **49,** 419-432.

[21] Magnus, J.R. and Klaassen, F.J. (1999a): The effect of new balls in a tennis set: four years at Wimbledon. *The Statistician*, **48**, 239-246.

[22] Magnus, J.R. and Klaassen, F.J. (1999b): On the advantage of serving first in a tennis set: four years at Wimbledon. *The Statistician*, **48**, 247-256.

[23] Maher, M.J. (1982): Modeling association football scores. *Statistica Neerlandica*, **36**, 109-118.

[24] Morton, R.H. (1997): Statistical effects of lane allocation on times in running races. *The Statistician*, **46**, 101-104.

[25] Mosteller, F. (1952): The World Series competition, *JASA*, **47**, 355-380.

[26] Nevill, A.M., Newell, S.M., and Gale, S. (1996): Factors associated with home advantage in English and Scottish soccer. *Journal of Sport Science*, **14**, 181-186.

[27] Pollard, R. (1986): Home adventage in soccer: a retrospective analysis. *Journal of Sport Sciences*, **4**, 237-248.

[28] Preston, I., and Thomas, J. (2000): Batting strategy in limited overs cricket. *The Statistician*, **49**, 95-106.

[29] Preston, I.and Thomas, J. (2002): Rain rules for limited overs cricket and probabilities of victory. *The Statistician*, **51**, 189-202.

[30] Rue, H. and Salvesen, O. (2000): Prediction and retrospective analysis of soccer matches in a league. *The Statistician*, **49**, 399-418.

[31] Schwertmann, N.C., Schenk, K.L., and Holbrook, B.C. (1996): More probability models for NCAA regional Basketball tournaments. *The American Statistician*, **50**, 34-38.

[32] Wright, P.B. (1997): Football standings and measurement levels. *The Statistician*, **46**, 105-110.