SORT 34 (2) July-December 2010, 181-200

On the use of simulation methods to compute probabilities: application to the first division Spanish soccer league

Ignacio Díaz-Emparanza* and Vicente Núñez-Antón**

Universidad del País Vasco

Abstract

We consider the problem of using the points a given team has in the First Division Spanish Soccer League to estimate its probabilities of achieving a specific objective, such as, for example, staying in the first division or playing the European Champions League. We started thinking about this specific problem and how to approach it after reading that some soccer coaches indicate that a team in the first division guarantees its staying in that division if it has a total of 42 points at the end of the regular season. This problem differs from the typical probability estimation problem because we only know the actual cumulative score a given team has at some point during the regular season. Under this setting a series of different assumptions can be made to predict the probability of interest at the end of the season. We describe the specific theoretical probability model using the multinomial distribution and, then, introduce two approximations to compute the probability of interest, as well as the exact method. The different proposed methods are then evaluated and also applied to the example that motivated them. One interesting result is that the predicted probabilities can then be dynamically evaluated by using data from the current soccer competition.

MSC: 62F05, 62P99, 6204, 6207

Keywords: Monte Carlo simulations, multinomial distribution, prediction, soccer league

Address for correspondence: Vicente Núñez-Antón, Departamento de Econometría y Estadística (E.A. III), Universidad del País Vasco, Avenida Lehendakari Aguirre, 83, E-48015 Bilbao, Spain. Phone: +34 94 601 3749; Fax: +34 94 601 3754; E-mail: vicente.nunezanton@ehu.es

* Departamento de Econometría y Estadística (E.A. III), Universidad del País Vasco, Bilbao, Spain. E-mail: ignacio.diaz-emparanza@ehu.es

** Departamento de Econometría y Estadística (E.A. III), Universidad del País Vasco, Bilbao, Spain. E-mail: vicente.nunezanton@ehu.es

Received: January 2010

Accepted: July 2010

1. Introduction

The *Liga de Fútbol Profesional* (LFP) stated that, starting at the regular season 1997-1998, there will be twenty teams competing in the First Division Spanish Soccer League. During the regular season, each team should play two games against each one of the remaining nineteen teams, one game at its own field and the other one at the other team's field. Therefore, during the regular season there will be a total of thirty-eight games played by each one of the teams participating in this league. After the 1995-1996 regular season the LFP stated the actual scoring system: a win gets a team three points, a draw one point and a defeat, no points. In this way, at the end of the regular season (i.e., after the thirty-eight games have been played), teams classified in the last three positions in the table (i.e., positions eighteenth to twentieth) will lose their place in the first division and will have to play the next regular season in the Second Division Spanish Soccer League. In addition, teams classified in the first four positions will play the European Champions League, the most prestigious soccer tournament in Europe (i.e., the one that only "the best" soccer teams in Europe will play), while teams classified in the fifth and sixth positions will play the UEFA tournament (nowadays called *Europa League*), an important soccer tournament for the so called "next-to-the-best" teams in Europe.

Soccer is the most important sport in Spain and there are several sports-related (TV and radio) programs that concentrate most of their attention and efforts on the Spanish soccer league. It is a fact that sports-related programs in Spain can very well be labelled as "soccer-centred programs." In the past few years, it has been very frequent to hear sport broadcasters indicating that "soccer coaches training teams in the first division believe that if a team obtains a total of 42 points at the end of the regular season, the team will remain in this division for the next regular season."[1] That is, the 42 points figure somehow represents the barrier that will determine if a team plays in the first or second division for the next regular season. In fact, after reading a specialized and well known sports newspaper, the first author's older son asked him the question of why some specific soccer coaches indicate that a team in the first division guarantees its staying in that division if it has a total of 42 points at the end of the regular season. This very simple and straightforward question originated our curiosity as to how one can propose some kind of approach to answer it. It also made us ask ourselves whether the question had been raised before by someone else.

A very simple analysis of the available data for the last twelve regular seasons reveals that only in three occasions a team obtaining a total of 42 points at the end of the regular season lost its right to play in the first division for the next season. More specifically, during the 1999-2000 and the 2008-2009 regular seasons, Betis, a soccer team from Seville and, during the 2007-2008 regular season, Zaragoza, another soccer team from a city with the same name, were both sent to play the corresponding next regular season in

1. For more details on this see, for example: http://bit.ly/biPsGx, http://bit.ly/maportugal, http://bit.ly/brindis-osasuna, or http://bit.ly/racing42.

the second division. Even though there are some statistical papers that propose to model scores in soccer leagues (see, e.g., Lee, 1997; Karlis and Ntzoufras, 2000; Rue and Salvensen, 2000; Brillinger, 2008; and Karlis and Ntzoufras, 2009), we are not aware of any scientific study or attempt that has tried to find out the actual probability that a team playing in the first division with 42 points at the end of the regular season stays in that division during the next season. Along the same lines, we do not know of any attempt that has been able to establish, using a rigorous statistical reasoning or tool, the total number of points a team should obtain during the regular season, so that it can stay in the first division for the next season. These represent two of the main objectives that have led us to put forward some of the proposals included in the next sections.

The rest of the paper is organized as follows. Section 2 introduces some basic notation and contains a brief description of all of the possible classifications at the end of a regular season for a four and a twenty team league. Section 3 describes the use of the multinomial distribution in the context of the soccer league under study, as well as the normal approximation, Monte Carlo simulations approximation and the exact probability computation for the different probabilities of interest. In Section 4, we include the proposed method to compute the probability of a team staying in the first division for the next regular season. Section 5 puts forward a dynamic probability computation method that allows the researcher or individual to compute different probabilities of interest during the regular season. Finally, Section 6 ends with some conclusions and practical recommendations. All of the proposals contained in the different sections of the paper are illustrated and evaluated with data from the First Division Spanish Soccer League.

2. Basic notation and possible classifications settings

Let A_i represent each of the i ($i = 1,\ldots,N$) teams participating in a given league. That is, all teams will be denoted by $A_1,\ldots,A_N$. The order of the team is not relevant and it could be, in fact, alphabetical, per region, or sorted by any other criteria. Let E_{ik} represent the points obtained by team i on its k-th game during the regular season ($k = 1,\ldots,2(N-1)$). In this way, the result of the game played by teams A_i and A_j in the k-th of the regular season can be easily summarized by the 2×1 score vector $(E_{ik},E_{jk})'$. In the following sections we will analyze these results for the case of a league of four and twenty teams, which is the actual size of the First Division Spanish Soccer League under study.

As we will see in later sections and without loss of generality, we assume equiprobability. That is, in each game we assume that the probability that the local team wins, loses or that the result is a draw are all equal. This implies that all possible final classifications have the same probability. This assumption implies that no additional a priori information is needed to be able to compute, for example, the probability that a team loses its category when having 42 points at the end of the regular season, or the probabil-

ity of winning the league with a given number of points at the end of the regular season. In this sense, all of the results reported here could be applied not only to the First Division Spanish Soccer League, but also to any second division or to any other division or league using the scoring system proposed in this league. In any case and given that it is very unlikely that all teams in this league have the same constant probability of winning a given soccer game, this is clearly a restrictive hypothesis that may be considered too strong in some cases for practical reasons but, at the same time, it may also be considered simple enough to be interesting from a didactic point of view. In fact, this is the main reason to start analyzing this problem under this assumption because, in our view, it clearly simplifies its solution and, in addition, it will also provide reference values for the probabilities of interest that may then be useful for the analysis of any other soccer league one wishes to study in the future.

A less restrictive assumption that also allows us to obtain interesting statistical results, can be that of *equal strength*. In order for this assumption to hold, the probability that the local team wins and the probability that it loses should be the same. That is, if we let p_1 be the probability that the local team wins, p_2 the probability that the result is a draw, and p_3 the probability that the local team loses the game, *equal strength* will occur if $p_1 = p_3 = (1-p_2)/2$. One interesting fact about this assumption is that it includes the equiprobability case as a particular case (i.e., if $p_1 = p_2 = p_3 = 1/3$), but it also includes additional possibilities that could also be analyzed. Along these lines, if we consider the First Division Spanish Soccer League historical data for the 11,242 games played from the 1976-1977 up to the 2008-2009 seasons, the estimated value we obtain for p_2, if we use the relative frequency for the event that the result of the game is a draw is, approximately, $\hat{p}_2 = 0.25$. In the following sections, we will use both the equiprobability and the *equal strength* assumptions. Finally, we should also mention that the equiprobability and equal strength assumptions imply that the probability that a given team wins, loses or that the result of its game is a draw, does not depend on which team it is playing against and that, therefore, there is an underlying independence assumption between games. This may also be a restrictive assumption but, in our view, it simplifies the solution to the problem of interest and, in addition, it provides the reader some very useful insights about the solution to a more complex problem.

2.1. A four-team soccer league

If we have four teams in the league, A_1, A_2, A_3, A_4, there will be three games in which a given team plays at home and three games in which it plays away from it, as a visiting team. That is, the regular season will have a total of six games. In this case, each date for which games are scheduled will have two games being played at the same time. If we let $a = (E_{11}, E_{21})'$ be the score for the game played by teams A_1 and A_2, and $b = (E_{31}, E_{41})'$ be the game played by teams A_3 and A_4, the possible scores for the first set of games to be played is listed in Table 1. After this first set of games is played, there are $3^2 = 9$ possible score vectors that are listed in the corresponding columns of Table 2. In order

Table 1: *Possible score vectors for a two-team soccer league.*

Possibilities	1	2	3
Score vector a:	$(3,0)'$	$(1,1)'$	$(0,3)'$
Score vector b:	$(3,0)'$	$(1,1)'$	$(0,3)'$

Table 2: *Possible scores for a four-team soccer league after the first set of games have been played.*

Result	a1b1	a1b2	a1b3	a2b1	a2b2	a2b3	a3b1	a3b2	a3b3
Team A_1	3	3	3	1	1	1	0	0	0
Team A_2	0	0	0	1	1	1	3	3	3
Team A_3	3	1	0	3	1	0	3	1	0
Team A_4	0	1	3	0	1	3	0	1	3

to better understand both the notation and contents in Table 2, let us describe one of the results provided therein (i.e., axby). The result in the fourth column of Table 2 (i.e., a2b1) indicates that in the game between teams A_1 and A_2 the result was a draw (i.e., the second possible result for the score vector a in Table 1, or a2), and in the game between teams A_3 and A_4 the result was that team A_3 won (i.e., the first possible result for the score vector b in Table 1, or b1). For this specific case, the final scores obtained by each of the teams A_1, A_2, A_3 and A_4 after the two games have been played would be of 1, 1, 3, and 0 points, respectively and, thus, the score vector would then be $(1,1,3,0)'$ (see the fourth column in Table 2). In summary, after the regular season ends (i.e., after each team has played its six corresponding games), there could be $3^{2\times 6} = 9^6 = 531,441$ *different*[2] results that will, in turn, generate their corresponding score vectors for these four teams.

The aforementioned number of possible results is clearly quite large. However, it can be easily managed by a computer. As the reader may have already guessed, this is exactly the situation in the first round of the Champions League competition, where only the first two teams in each group of four teams advance to the next round. Therefore, it would not be difficult to compute, for example, what would be the exact probability, for each possible score, that a team finishes the competition in the first two positions (i.e., the probability that the team advances to the next round in the Champions League):

Score:	≤ 6	7	8	9	≥ 10
Prob(next round):	0	0.0047477	0.18593	0.97050	1

That is, in all of the possible $531,441$ *different* results, we look for all results where a team having a given score finishes the competition in the first two positions (i.e., these

2. These results are different in the sense that, even though the scores could end up being equal for some of the cases, they were generated from different results in the games the teams have played.

will the favourable cases) and divide this absolute frequency by the total number of cases where teams had obtained this score (i.e., these will be the possible cases).

2.2. The twenty-team or first division Spanish soccer league

The Spanish First Division Soccer League, as well as, for example, the ones in France or the United Kingdom, has a total of twenty teams (i.e., $N = 20$), so that every round there will be ten different games played at the same time. In addition, every team should play nineteen games at home and another nineteen games as a visiting team, so that the regular season will have a total of thirty-eight different rounds.

If we let $(E_{ik}, E_{jk})'$ be the score vector representing the result of the game between teams A_i and A_j in the k-th round of games during the regular season, we have that:

$$(E_{ik}, E_{jk}) = \begin{cases} (3,0) & \text{if } A_i \text{ wins} \\ (1,1) & \text{if the result is a draw} \\ (0,3) & \text{if } A_j \text{ wins} \end{cases}$$

Therefore, after the k-th round of games is over (i.e., after the ten scheduled games have been played by the twenty teams in the league), we will have that $E_k = (E_{1k}, E_{2k}, \ldots, E_{20k})'$ represents the score vector assigned to all teams for the games played that date. Moreover and given that for each one of the ten games played that date there are only three possible different results, the number of different score vectors that one can obtain for that specific date is equal to $3^{10} = 59,049$.

Let C_k be the 20×1 score vector containing the sum of the scores from the first up to the k-th round of games, so that

$$C_k = \sum_{l=1}^{k} E_l,$$

and $C_k = (C_{1k}, \ldots, C_{20k})'$. Therefore, $C_{ik},\ i = 1, \ldots, 20;\ k = 1, \ldots, 38$ represents the score team A_i has after playing k games. If we place the elements of C_k in descending order and denote this new score vector by $C_k^o = (C_{(1)k}, \ldots, C_{(20)k})' = (C_{1k}^o, \ldots, C_{20k}^o)'$, we will have in the elements of the ordered score vector C_k^o the complete information about *teams classification or standings after k games have been played*, which will be very relevant to compute the probabilities of interest.

If, for example, we wish to analyze the number of vectors with possible *different* scores after two rounds of games have been played (i.e., C_2), we have to consider that each one of the 59,049 resulting score vectors for the second date for which games were scheduled can be added to each one of the 59,049 score vectors for the first date for which games were also scheduled, making a total of $3^{10\times 2} = 59,049 \times 59,049 = 3,486,784,401$ possible results, even though we know that a large number of them will be basically equal. If we follow the same reasoning, we can find out that the number of

vectors with possible *different* scores for the score vector C_{38} at the end of the regular season would then be $3^{10\times 38} = 3^{380} \simeq 2.023376E+181$.

In order to be able to compute the exact probability of losing the category for a team having 42 points, just as we did in Section 2.1 for the probability of advancing to the next round in the Champions League for the four teams' case, we would have to find out for how many of these $2.023376E+181$ score vectors a team having 42 points stays away from the last three positions in the table (i.e., stays away from the last three positions in the ordered score vector C^o_{38}). It is clear that, even with the current capabilities large computers have to compute this probability, it is not reasonable to think about working with such a large number of possibilities. For example, if the computer is able to compute 1,000 score vectors per second, after a year of computations, the computer would have only computed about $3.1536E+10$ score vectors. Therefore, there is a need to look for efficient and reasonable proposals that can make such a complicated computation of probabilities possible.

3. Multinomial distribution

The settings we have introduced in the previous sections allow us to state that, for each game and team, the set of possible results can be classified in the disjoint events: R_1 (winning the game), R_2 (game ends in a draw) or R_3 (losing the game). We now define the probabilities for these events as follows:

$$\Pr(R_j) = p_j \qquad \text{with} \quad 0 < p_j < 1, \quad j = 1,2,3 \quad \text{and} \quad p_1 + p_2 + p_3 = 1$$

To start with a simple setting, we can consider a discrete uniform probability distribution for the three alternatives, so that it is assumed that $p_1 = p_2 = p_3 = 1/3$. That is, we start with the initial aforementioned equiprobability assumption. Under this assumption, for any team in the league, the random variable $X = (X_1, X_2, X_3)'$, describing the event that after n dates for which games were scheduled during the regular season, there were x_1 times where the event R_1 occurred, x_2 times where the event R_2, and x_3 times where the event R_3 occurred, follows a multinomial distribution with probability mass function given by (see, e.g., Morris, 1975)

$$\Pr(X_1 = x_1, X_2 = x_2, X_3 = x_3) = \frac{n!}{x_1!x_2!x_3!}\, p_1^{x_1} p_2^{x_2} p_3^{x_3}, \tag{1}$$

$$x_1 + x_2 + x_3 = n, \quad 0 < p_j < 1, \quad j = 1,2,3 \quad \text{and} \quad p_1 + p_2 + p_3 = 1$$

It is well known that the marginal distribution of each of the X_j variables from a multinomial distribution follows a binomial distribution with parameters n and p_j; that is, $X_j \sim B(n, p_j)$, $j = 1,2,3$ with $E(X_j) = np_j$, $Var(X_j) = np_j(1-p_j)$ and $Cov(X_i, X_j) =$

$-np_ip_j,\ i,j=1,2,3,\ i\neq j$. As we have already mentioned, we are under the equiprobability assumption. However, the distribution of the random variable $X=(X_1,X_2,X_3)'$ is multinomial as long as the assumed probabilities p_1, p_2 and p_3 remain unchanged for all games in the league. This implies that, for example, under the *equal strength* assumption, the distribution of the random variable X is also multinomial.

3.1. Normal distribution approximation

If we use the multivariate normal central limit (see, e.g., Agresti, 1990, p. 424; or Rao, 1973, p. 128), we can see that the multinomial distribution converges to the multivariate normal distribution, so that $X=(X_1,X_2,X_3)'$ converges in distribution to a multivariate normal distribution with mean vector given by $\mu_X=(np_1,np_2,np_3)'$ and variance-covariance matrix Σ_X with elements given by:

$$\Sigma_{X_{ij}}=\begin{cases} np_i(1-p_i) & \text{if} \quad i=j \\ -np_ip_j & \text{if} \quad i\neq j \end{cases} \tag{2}$$

The score a given team A_i obtains after playing n games is $C_{in}=3X_1+X_2$, a linear combination of the components of the asymptotic multivariate normal random variable X, which can be written as $C_{in}=d'X$, with $d'=(3,1,0)$. Therefore, C_{in} converges to the univariate normal distribution $N(d'\mu_X,d'\Sigma_Xd)$.

For the specific case under study, we have that $n=38$ and $p_1=p_2=p_3=1/3$, so that the standard conditions (i.e., $np_i>5$ and $n(1-p_i)>5$, see, e.g., Hogg and Tanis, 1988 or Cryer and Miller, 1991) for a valid approximation hold and, therefore, we have that

$$C_{i38}\approx N(50.67,59.11) \tag{3}$$

The probability that a given team in the league loses its category is the probability that its ordered position after the regular season ends in one of the last three out of the twenty possible positions. Thus, we are interested in computing the critical score value, say C_{38c}, such that there would be three teams below it (around 15% or 3 out of 20 teams) or three teams having a score smaller than C_{38c}. In other words, C_{38c} should be such that $\Pr(C_{i38}\leq C_{38c})\geq 0.15$. In order to compute C_{38c} and using the result in (3), we have that

$$\frac{C_{i38}-50.67}{\sqrt{59.11}}\approx N(0,1)$$

Therefore,

$$\Pr\left(\frac{C_{i38}-50.67}{\sqrt{59.11}}\leq -z_{0.15}=-1.0364\right)=0.15 \tag{4}$$

being $-z_{0.15} = -1.0364$, the 15-th percentile of the standard normal distribution, $N(0,1)$. Solving for C_{i38} in the left hand side inside the parenthesis, leads us to obtain that

$$\Pr\left[C_{i38} \leq 50.67 - (1.0364)\sqrt{59.11}\right] = 0.15$$

and, thus, $\Pr(C_{i38} \leq 42.70) = 0.15$.

If we apply a standard continuity correction[3], we would have that the C_{38c} value we are searching for is $C_{38c} = 43$. This value leads us to obtain that $\Pr(C_{i38} \leq 43) = 0.1755$. Moreover, we can also easily verify that $\Pr(C_{i38} \leq 42) = 0.1439$. Therefore, the objective score for any team wishing not to lose its category in the First Division Spanish Soccer League should be of 43 points.

In addition, if we use the *equal strength* assumption with a probability that the result of a draw is $p_2 = 0.25$, we have that $C_{i38} \approx N(52.25, 65.92)$ and, therefore, $\Pr(C_{i38} \leq 43.83) = 0.15$. If we use again the aforementioned continuity correction and given that we can easily compute $\Pr(C_{i38} \leq 44) = 0.1699$ and $\Pr(C_{i38} \leq 43) = 0.1406$, we would now have that $C_{38c} = 44$.

3.2. Monte Carlo simulations approximation

A second alternative approach to obtain the distribution of C_{i38} consists of using a simulations approach. Let us begin by recalling that we are assuming equal probabilities for each one of the three possible results than a given game can have; that is, R_1 (winning the game), R_2 (game ends in a draw) and R_3 (losing the game):

$$\Pr(R_j) = p_j = \frac{1}{3} \qquad \text{for} \quad j = 1,2,3.$$

Most statistical packages include random number generators based on the uniform distribution and, thus, it is quite simple to simulate the result of a given game with the use of this software[4]. In this sense, if we assume independence among the games played at each round during the regular season, the results for ten independent games can be easily simulated in order to obtain the scores for all twenty teams after that specific date. We also assume that the probabilities p_j remain constant for each game so that, under the previous assumptions, the results for the different round of games are also independent. We can then repeat the whole simulation process thirty-eight times in order to be able to simulate the results for the final standings for all twenty teams in the league at the end of the regular season. The whole process can be easily summarized as follows:

---

3. We consider that for any $x \in \{0,1,\ldots,n\}$, if the conditions to consider the normal approximation a valid one hold, then $\Pr(X \leq x) = \Pr(X < x+1)$ can be well approximated by $\Pr(Y \leq x + \frac{1}{2})$, where Y is a normal random variable having the same mean and variance as the random variable X.

4. We have used the open source software package `gretl` (see, e.g., `http://gretl.sourceforge.net` or Cottrell and Lucchetti, 2009).

1. Based on the uniform distribution, we generate the results of the game between teams A_i and A_j in the first date for which games are scheduled, and obtain the corresponding score vector $(E_{i1}, E_{j1})'$.
2. Repeat step 1 ten times and obtain the results for all ten games played in the first date for which games are scheduled. At the end of this step, we obtain the 20×1 score vector $E_1 = (E_{11}, \ldots, E_{20,1})'$.
3. Repeat steps 1 and 2 for each one of the thirty-eight dates for which games are scheduled, generating the corresponding 20×1 score vectors $E_k = (E_{1k}, \ldots, E_{20k})'$, $k = 1, \ldots, 38$, and obtain the sum of the scores for all twenty teams after the "simulated" regular season ends; that is, obtain the 20×1 final scores vector $C_{38} = \sum_{l=1}^{38} E_l$.
4. Finally, repeat steps 1 to 3 a large number of times, say M, and obtain the simulated frequency distribution for $C_{38} = (C_{1,38}, \ldots, C_{20,38})'$, an approximation of the probability distribution for this random variable and, accordingly, of its individual components C_{i38}.

If we follow the procedure described in Díaz-Emparanza (2002, equation (8)), we see that, with a 95% confidence level, M =10,000 replications will suffice to guarantee a precision of ± 0.007 in the estimation of the 15% distribution percentile of interest.

After these simulations are performed, we can straightforwardly obtain that $\Pr(C_{i38} \leq 43) = 0.1770$ and $\Pr(C_{i38} \leq 42) = 0.1448$, values that, as can be easily verified, are very close to those obtained in Section 3.1 with the use of the normal approximation.

However, it is also possible to consider an alternative interpretation of the results obtained in this simulation approach. In Section 2.2 we have indicated that there is a large number of *different* possibilities for values in the final score vector C_{38}. Statistics usually tells us that if we wish to learn about the specific characteristics of a given population that is impossible to measure or compute, we can use statistical inferential methods. That is, based on the values obtained from a random sample of a "reasonable size" from the population under study, we can always extract information that allows us to estimate the characteristics of interest and, thus, be able to generalize the obtained conclusions to the population under study. In this specific case, we can interpret our proposed procedure as one that randomly extracts or samples possible final scores (or standings) among the set of all final scores (or standings) that we have in the First Division Spanish Soccer League. The use of the assumption of equal probabilities for the three possible results R_1, R_2 and R_3 in the simulations guarantees that, in the random extraction or sampling, all possible score or standing vectors will be equally likely or have the same probability of being selected in the sample.

In addition, if we use the *equal strength* assumption with a probability that the result of a draw is $p_2 = 0.25$, and also using M =10,000 replications, we obtain that $\Pr(C_{i38} \leq 44) = 0.1728$ and $\Pr(C_{i38} \leq 43) = 0.1427$.

3.3. Exact probability computation

The specific probability computation under study does not require the use of either the normal approximation or the Monte Carlo simulation approaches proposed in the previous sections. More specifically, if we use enumeration techniques it is possible to find the exact probability distribution for C_{i38}.

In Section 3 we have seen that the random variable $X = (X_1, X_2, X_3)'$ follows a multinomial distribution, with probability mass function given by equation (1). Therefore, as $n = 38$, each one of its individual components, $X_i,\ i = 1,2,3$ will take on values in the set $\{0,1,2,\ldots,38\}$, and the set of possible values for the random variable X is finite (i.e., there are $(n+1)(n+2)/2 = 780$ possible values), so that the probability for each one of its possible values can be easily computed by using (1). Once these probabilities have been obtained, for each one of them, we can compute $C_{i38} = 3X_1 + X_2$ and the probabilities for each one of the $(3n+1) = 115$ possible different values for C_{i38} can be easily added up together. We have done this in `Gretl` and list both the possible values and corresponding probabilities for C_{i38} (see Table 3, column with $p_2 = 1/3$). Table 3 includes the values $\Pr(C_{i38} \leq 43) = 0.1768$ and $\Pr(C_{i38} \leq 42) = 0.1444$, values that are very close to those reported in the approximation methods proposed in Sections 3.1 and 3.2 and, thus, these results confirm the conclusion (see the results reported in Sections 3.1 and 3.2) that a team wishing to stay in the first division should obtain at least 43 points at the end of the regular season. One of the anonymous reviewers suggested an alternative procedure to compute the exact probabilities by means of the probability generating function (pgf) of the random variable X, that measures the number of points gained in a match. That is,

$$g(t) = p_1 + p_2 t + p_3 t^3$$

In this sense, the sum of 38 such random variables has a pgf given by $g(t)^{38}$ and, thus, the probability values associated to each score in the distribution function is given by the coefficients of this polynomial.

The previously reported results have used the equiprobability assumption, something that may be a very restrictive assumption in the view of some researchers. However, as we have already mentioned in Section 3, under the *equal strength* assumption, the distribution of the random variable X is also multinomial. As above, from this resulting distribution and with the use of enumeration methods, we can obtain the exact probability distribution for $C_{i38} = 3X_1 + X_2$ as well. Moreover, we can easily compute the probability distribution function for C_{i38} for different values of p_2 (i.e., the probability that the result of the game is a draw). Figure 1 includes the probability distribution functions for different values of p_2 (i.e., for $p_2 = 0.1,\ldots,0.9$). In addition, Table 3 only includes the probability values for the so called central values of the distribution of the Score variable, also as a function of p_2. As can be seen in Table 3, we have included the corresponding probability values for several cases of the *equal strength* model, which contains two of its special cases we have been studying so far: the equiprobability case

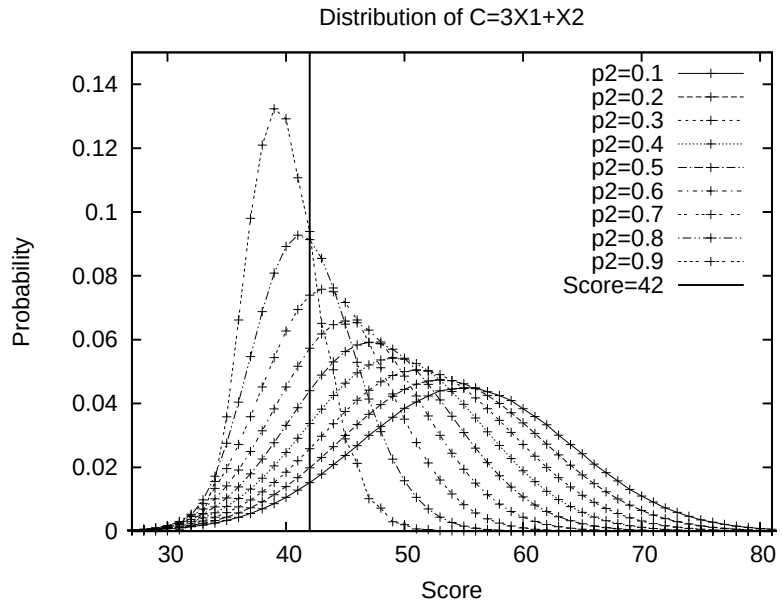

Figure 1: *Team final scores probability distributions as a function of the probability p_2 that the result of the game is a draw.*

Table 3: *Final scores cumulative probability values $C_{i38} = 3X_1 + X_2$ as a function of the probability of a draw, p_2. Reported results correspond to final scores ranging from* 36 *to* 50 *after the regular season has ended and for a twenty-team league. Boldfaced numbers indicate the required final score a team should have at the end of the regular season for not losing the category under the assumed p_2 probability.*

	Probability of a draw: p_2										
Score	0.10	0.20	0.25	0.30	1/3	0.40	0.50	0.60	0.70	0.80	0.90
36	0.0167	0.0217	0.0246	0.0281	0.0306	0.0364	0.0474	0.0617	0.0804	0.1039	0.1252
37	0.0223	0.0290	0.0331	0.0378	0.0414	0.0496	0.0653	0.0867	0.1162	**0.1586**	**0.2232**
38	0.0292	0.0382	0.0437	0.0501	0.0550	0.0661	0.0879	0.1181	**0.1614**	0.2274	0.3442
39	0.0379	0.0496	0.0569	0.0654	0.0717	0.0866	0.1156	**0.1564**	0.2157	0.3082	0.4765
40	0.0484	0.0636	0.0730	0.0838	0.0920	0.1112	0.1488	0.2016	0.2784	0.3974	0.6057
41	0.0613	0.0804	0.0922	0.1059	0.1162	0.1403	**0.1875**	0.2533	0.3479	0.4901	0.7164
42	0.0765	0.1002	0.1148	0.1317	0.1444	**0.1740**	0.2315	0.3106	0.4218	0.5814	0.8103
43	0.0944	0.1233	0.1410	**0.1614**	**0.1768**	0.2122	0.2803	0.3724	0.4976	0.6669	0.8754
44	0.1152	0.1498	**0.1709**	0.1951	0.2132	0.2548	0.3334	0.4370	0.5726	0.7431	0.9260
45	0.1389	**0.1797**	0.2044	0.2326	0.2535	0.3011	0.3896	0.5029	0.6442	0.8080	0.9549
46	**0.1657**	0.2131	0.2415	0.2736	0.2973	0.3507	0.4480	0.5682	0.7103	0.8609	0.9761
47	0.1957	0.2497	0.2818	0.3178	0.3441	0.4028	0.5071	0.6311	0.7694	0.9023	0.9863
48	0.2286	0.2893	0.3250	0.3646	0.3932	0.4564	0.5658	0.6904	0.8207	0.9335	0.9935
49	0.2643	0.3316	0.3705	0.4133	0.4440	0.5106	0.6228	0.7447	0.8637	0.9560	0.9965
50	0.3027	0.3760	0.4178	0.4633	0.4955	0.5645	0.6769	0.7933	0.8987	0.9718	0.9985
51	0.3432	0.4220	0.4663	0.5137	0.5470	0.6170	0.7274	0.8356	0.9265	0.9824	0.9992

(i.e., $p_2 = 1/3$) and the case for which $p_2 = 0.25$. For this latter case, we can clearly see that $\Pr(C_{i38} \leq 44) = 0.1709$ and $\Pr(C_{i38} \leq 43) = 0.1410$, so that, $C_{i38c} = 44$.

4. Probability of not losing the category

In order to compute the probability of not losing the category for a given team having a final score c, we would have to check the joint distribution of the scores for all twenty teams in the league at the end of the regular season and see in how many cases a team having c points has not lost the category. This implies working with the joint distribution of a 20×1 vector of random variables, in which each of its individual components would have a similar distribution to that described for C_{i38} in Section 3.3. In addition, we have to point out that these individual variables (i.e., $C_{1,38}, \ldots, C_{20,38}$) are not independent random variables, which makes this a complicated theoretical problem to solve. However, it is not difficult to obtain an approximation of this distribution by using a Monte Carlo simulation approximation, such as the one previously described in Section 3.2.

In order to describe this new approach, we define a binary random variable D_1, taking value one if the score c appears as part of the final standings score vector C_{38} and if, in addition, it is larger than the score obtained by the team appearing in the eighteenth final standings ordered position vector C^o_{38}, and zero otherwise. That is,

$$D_1 = \begin{cases} 1 & \text{if } c \in C_{38} \text{ and } c > C^o_{18,38} \\ 0 & \text{otherwise} \end{cases}$$

Therefore, in the simulation process described in step 3 of Section 3.2, each time a simulated final score vector C_{38} is obtained, the random variable D_1 takes on two possible values, one or zero. After a sufficiently large number of replications M has been simulated, and if we let m_c be the number of occasions in which the random variable D_1 has taken value one, and M_c be the total number of occasions in which the value c appeared in the final standings score vector C_{38}, then, for a team obtaining c points at the end of the regular season, m_c/M_c would be an approximation of the probability of not losing the category $p_{nlc}(c)$. We should indicate that in a simulation with a finite number of replications we could have two easy to handle types of indetermination showing up: cases with very low scores or cases with very high scores. More specifically, none of the 50,000 replications performed to obtain the results reported in Figure 2 included scores lower than 15 points or higher than 88 points. The problem is solved by assigning probability zero to the very low values and probability one to the very high values obtained in the simulation process. These simulations were carried out for $M = 50,000$, and values of $c = 0, 1, 2, \ldots, 114$ were considered, with the results reported in Figure 2. In this specific case, the probabilities of not losing the category for a team having 42

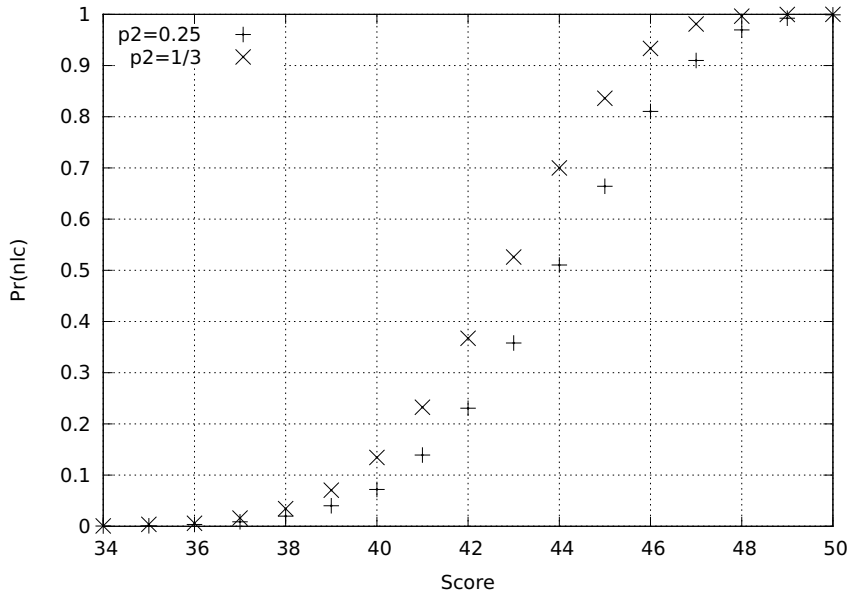

Figure 2: *Probability of not losing the category, p_{nlc} for each final score. Probabilities were computed by simulations and with M =50,000 replications, with $p_1 = p_3$ and two different values for the probability of a draw, $p_2 = 1/3$ (i.e., equiprobability assumption) and $p_2 = 0.25$ (i.e., equal strength assumption).*

and 43 points are 0.3673 and 0.5259, respectively, under the equiprobability assumption, and 0.2307 and 0.3581, respectively, under the *equal strength* assumption with $p_2 = 0.25$. If one wishes to compute instead the probability of playing the European Champions League or the Europa League tournament, the binary variable should be defined accordingly.

5. Dynamic computation of probabilities during the regular season

From a practitioners' point of view, it would be very interesting to have the possibility of computing, after a given number of rounds of games have been played (say k) and conditioned on the current score vector (say C_k), the probability that a given team wins the league, plays the European Champions League or ends in a position that will make that team not to lose its category at the end of the regular season. These probabilities can then be used by the teams to make strategic decisions during the regular season and not at the end of it when things cannot be changed. For example, a team whose main objective at the end of the current regular season is to stay in the first division, would be able to determine that, if the probability of not losing the category, say at mid-season, is smaller than 0.10, the team will change its coach at mid-season. However, another team whose main objective at the end of the regular season is to win the league

could decide that it would change its coach if the probability of winning the league at mid-season is lower that 0.75. That is, conditions and decision are highly linked to the team's objectives during the regular season. We should also mention that, depending on the specific conditions of the league under study, it is very likely that our equiprobability or equal strength assumptions provide a solution that may not be too realistic. In fact, if there are reasons that lead us to believe that these hypotheses do not hold, we should probably propose a more complex or general probability model that allows us to improve the reported results. In any case, we do believe that for any soccer league it would be interesting and useful to have the reference values that can be easily obtained from the model under the aforementioned assumptions.

Therefore, our aim is to be able to compute, for a given team A_i, a given date k for which games were scheduled during the regular season ($k < 38$), and conditioned on the current score vector C_k, the dynamic conditioned probability that at the end of the regular season the team A_i loses its category (plays the European Champions League, plays the Europa League tournament or wins the league).

The method proposed in Section 3.3, in which we now have $n = 38 - k$ can be used to compute the exact probability distribution for each team's score at the end of the regular season, conditioned on the score each team has at the k-th round of games, C_{ik}. That is, we would be able to find the marginal distribution of the random variable C_{i38} (team's A_i score at the end of the regular season), conditioned on the current information we have for the k-th round of games. However, in order to compute the probability that, at the end of the regular season, a given team does not lose its category, conditioned on the current information we have (say C_k), it is necessary to take into account the complete structure the score vector C_k has; that is, the score all twenty teams have at that specific date. From this information, the computation of the probability of a team not losing its category means, as we saw in Section 4 above, working with the joint probability distribution of the scores for all twenty teams. Moreover, if we consider that those scores are not independent we will soon arrive at the conclusion that the analytical computation of this probability is a complicated probability problem, just as we had in Section 4.

As one can see, this is also a very simple problem if we decide to use Monte Carlo simulation techniques to solve it. There are differences, however, with the solution we proposed in Section 4, which will be described in detail below. In this case, the simulation process would start by taking the scores in the k-th date as given or known (i.e., C_k is assumed to be known) and, thus, we would only need to simulate the results for the remaining $38 - k$ dates for which games are scheduled. The whole process can be easily summarized as follows:

- We assume that the 20×1 current score vector for the k-th round of games, C_k, is known.

- For a given team A_i, we define a binary random variable D_2, taking value one if, at the end of the regular season, the team's position in the final standings ordered

score vector C^o_{38} is in one of its first seventeenth places, and zero otherwise. That is, D_2 will take value one if team A_i's score value $c = C_{i38}$ at the end of the regular season is larger that the score obtained by the team at the eighteenth position in the final standings ordered score vector C^o_{38} (i.e., $C^o_{18,38}$), and zero otherwise. We should point out that we are not taking into account any additional criteria such as, for example, goal differences that would decide the final position of two teams (i.e., the ones in positions seventeenth and eighteenth) in case of two teams having the same score, mainly because after thirty-eight games this is not so likely to occur. That is,

$$D_2 = \begin{cases} 1 & \text{if } C_{i38} > C^o_{18,38} \\ 0 & \text{if } C_{i38} \leq C^o_{18,38} \end{cases}$$

which can be easily done simultaneously for all twenty teams, so that we would now have a 20×1 vector of binary indicator variables.

- For the remaining $38-k$ rounds of games, repeat step 3 in the simulation process described in Section 3.2, so that we obtain the final standings ordered score vector C^o_{38} at the end of the regular season. The binary variable D_2 will then take on values one or zero.

- Repeat the whole process of generating the remaining $38-k$ dates for which games are scheduled for a sufficiently large number of replications M. In each replication, the binary variable will take on values one or zero. If we let m be the number of occasions in which the binary variable D_2 has taken value one, then m/M would be an approximation of the probability of team's A_i not losing its category $p_{nlc}(c)$, conditioned on the current score vector C_k.

We now apply this to the soccer league motivating our proposals (see Table 4). The third column in Table 4 (labelled as C^o_{19} in the left-hand side of the table), includes the standings for the First Division Spanish Soccer League after the $k = 19$-th round of games (January 24, 2010). Using the method just described in this section and $M = 10,000$ replications, we have computed the probabilities, conditioned on the scores at $k = 19$, of not losing the category, playing at least the "Europa League" (formerly UEFA tournament), playing the European Champions League, and winning the league for all twenty teams in the 2009-2010 regular season. These results are listed on the right-hand side of Table 4. In order to compare this prediction, based on the information available when about 50% of the regular games were played, with the actual final standings for the last regular season, it is probably worth noting that Barcelona won the league and that, in addition, Real Madrid, Valencia, and Sevilla classified to play the European Champions League. Furthermore, Mallorca and Getafe classified to play the "UEFA Europa League", and Valladolid, Tenerife and Xerez lost their category. There were

Table 4: *Teams' classification in the First Division Spanish Soccer League and probabilities computed with the Monte Carlo simulations approximation. In this case, we were in the $k = 19$-th date for which games were scheduled-January 24, 2010 (for Pr(Win), if two or more teams have the same number of points at the end of the regular season, a tie-breaking mechanism that uses a uniform random variable has been applied).*

	Team	C^o_{19}	p_{nlc}	Pr(Europa)	Pr(Champ)	Pr(Win)
1	Barcelona	49	1.0000	0.9983	0.9899	0.6812
2	Real Madrid	44	1.0000	0.9870	0.9452	0.2318
3	Valencia	39	1.0000	0.9190	0.7570	0.0591
4	Mallorca	34	0.9989	0.7043	0.3838	0.0088
5	Deportivo	34	0.9992	0.6916	0.3758	0.0093
6	Sevilla	33	0.9976	0.6370	0.3068	0.0055
7	Getafe	30	0.9900	0.3876	0.1471	0.0022
8	Athletic	30	0.9900	0.4016	0.1525	0.0022
9	Villarreal	26	0.9399	0.1580	0.0425	0.0001
10	Sporting	24	0.9037	0.0829	0.0175	0.0001
11	Atlético	23	0.8894	0.0671	0.0160	0.0002
12	Osasuna	23	0.8867	0.0653	0.0212	0.0001
13	Rácing	23	0.8825	0.0738	0.0244	0.0001
14	Espanyol	20	0.7389	0.0317	0.0062	0.0000
15	Almería	18	0.5951	0.0148	0.0018	0.0000
16	Málaga	17	0.5120	0.0084	0.0011	0.0000
17	Valladolid	17	0.5234	0.0083	0.0012	0.0000
18	Tenerife	17	0.5089	0.0068	0.0006	0.0000
19	Zaragoza	14	0.3021	0.0009	0.0001	0.0000
20	Xerez	8	0.0689	0.0001	0.0001	0.0000

two relevant issues that provided not expected results for the 2009-2010 regular season: Zaragoza did not lose its category and Deportivo did not play the Europa League. Zaragoza's performance during the second half of the regular season was quite better than that in the first half of the regular season (obtaining 27 points out of 57 possible points), a fact that allowed the team to stay in the first division. Deportivo's performance during the second half of the regular season was quite unexpectedly bad (obtaining only 13 points out of the possible 57 points, while in the first half it had obtained 34 out of 57 points). As can be clearly seen, this is a fact the proposed method clearly did not take into account because its prediction was based on past data. Of course, it is clear that dynamic predictions would be better as we approach the end of the regular season.

6. Conclusions and practical recommendations

We have proposed an approximate method to compute the probability that a team having 42 points has of losing its right to play in the first division the next regular

season. Under the assumption that all possible classifications are equally likely, this method allows us to obtain an estimated value of 0.3673 for this probability, and, an estimated value of 0.2307 under the *equal strength* assumption with probability of a draw of $p_2 = 0.25$.

We have described the normal and Monte Carlo simulated approximations, as well as the exact method, to estimate what would be the objective score a team should aim for in order to stay in the first division of the Spanish soccer League. All three methods have concluded that the objective score for such a team should be of at least 43 points.

Finally, we have also proposed a simulation-based method that allows us to compute, in a dynamic form and after the k-th round of games has ended, the probability, conditioned on the scores it has up to and including that k-th date, of a team not losing its category (or winning the league, of playing the European Champions League or the Europa League tournament)

As we have already mentioned in previous sections, the equiprobability and equal strength assumptions, even after being considered too simplistic or not too realistic hypotheses, have two fundamental and very relevant advantages: under these assumptions, computations are quite simple because of their underlying independence assumption between games, and, in addition, they do not require of any additional a priori information to be able to compute the probabilities of interest. In practice, if one wishes to study the problem of a "real" league, just like the First Division Spanish Soccer League in which there are real reasons to believe that the probability of winning a game, losing a game or that the result of the game is a draw for each team is different (i.e., large or even extreme differences in the budgets for the different teams), then the results reported here can be only considered as upper or lower bounds for the probabilities of interest. For example, it is quite reasonable to believe that a team in the first (or last) position in the league will have a probability larger (or smaller) than $1/3$ of winning most of its games and this can clearly result in the fact that the probability values reported here for winning the league or winning a place to compete in the Champions League (or losing its category) for this specific team can be then considered as lower (or upper) bounds for the real probability of interest.

Future research includes the possibility of not having equally likely classifications or adding some additional information, such as some differential characteristics the different teams in the league have. For example, teams having a larger budget (i.e., richer teams) have more possibilities of bringing better players to their teams. One way of approaching this new problem could be, for example, to establish an *a priori* probability of winning each game that somehow depends on the team's budget. An additional possibility would be to establish this probability taking previous results as the basis for it. In any case, this is out of the scope of this paper and it will be the objective of future research.

Acknowledgements

This research was supported by grants SEJ2007-61362/ECON, MTM2007-60112, ECO 2010-15332 and MTM2010-14913 (Ministerio Español de Ciencia e Innovación and FEDER), and IT-334-07 (Departamento de Educación del Gobierno Vasco - UPV/EHU Econometrics Research Group). The authors would also like to thank two reviewers for providing thoughtful comments and suggestions which led to substantial improvement of the paper.

References

Agresti, A. (1990). *Categorical Data Analysis*. New York: Wiley.

Brillinger, D. R. (2008). Modelling game outcome of the Brazilian 2006 Series A Championship as ordinal-valued. *Brazilian Journal of Probability and Statistics*, 22, 89-104.

Cottrell, A. and Lucchetti, R. (2009) Gretl User's Guide. Gnu Regression, Econometrics and Time Series. http://sourceforge.net/projects/gretl/files/manual/ [Online; November, 2009 version].

Cryer, J. B. and Miller, R. B. (1991). *Statistics for Business: Data Analysis and Modelling*. Boston: PWS-KENT publishing Company.

Díaz-Emparanza, I. (2002). Is a small Monte Carlo analysis a good analysis? Checking the size, power and consistency of a simulation-based test. *Statistical Papers*, 43(4), 567-577.

Hogg, R. W. and Tanis, E. A. (1988). *Probability and Statistical Inference*. New York: Macmillan Publishing Company.

Karlis D. and Ntzoufras J. (2000). On modelling soccer data. *Student*, 3, 229-245.

Karlis, D. and Ntzoufras, I. (2009). Bayesian modelling of football outcomes: using the Skellam's distribution for the goal difference. *IMA Journal of Management Mathematics*, 20, 133-145

Kleijnen, J. P. C. (1987). *Statistical Tools for Simulation Practitioners*. New York: Marcel Dekker, Inc.

Lee, A. J. (1997). Modeling scores in the premier league: is Manchester United *really* the best? *Chance*, 10, 15-19.

Morris, C. (1975). Central limit theorems for multinomial sums. *The Annals of Statistics*, 14(1), 165-188.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications*. New York: Wiley.

Rue, H. and Salvesen, O. (2000). Prediction and retrospective analysis of soccer matches in a league. *Journal of the Royal Statistical Society-Series D (The Statistician)*, 49, 399-418.