Monographic Section

DOI:10.25115/eea.v40i1.7002

Studies on Applied Economics / Estudios de Economía Aplicada

Volume 40-1, January 2022 // ISSN: 1133-3197

An Application of Stochastic Dominances in Sports Analytics

Fernández-Ponce, JM [a]; Rodríguez-Griñolo, MR.[b] and Troncoso-Molina, MA[a]

[a] Departamento de Estadística e Investigación Operativa
Universidad de Sevilla (Spain)
Email: ferpon@us.es

[b] Departamento de Economía, Métodos Cuantitativos e Historia Económica
Universidad Pablo de Olavide (Spain)

Abstract

Stochastic orders or stochastic dominance as they are known in economics, have been widely studied and applied in a variety of scientific fields, from biology to Systems Engineering. However, to the best of our knowledge, there is an application gap in the field of Sports Analytics or Sports Sciences. In this paper, we attempt a first approach to a possible application of stochastic orders to a dataset of LaLiga (Spain) football matches. Our aim is simply to show how a comparison can be extended beyond a simple metric comparison. In particular, we will focus on the first and second dominance stochastic orders as they are the most intuitive and simple to interpret and are also the most widely used in economics.

Keywords: Beta Distribution, Expected Goals, Sports Analytics, Stochastic Orders.

JEL Classification: C81,C88.

Received: 30/11/2021
Accepted: 27/01/2022

Sección Monográfica

DOI:10.25115/eea.v40i1.7002
]

Studies Estudios de Applied Economics Economía Aplicada
Volumen 40-1, Enero 2022 // ISSN: 1133-3197

Una Aplicación de las Dominancias Estocásticas en la Analítica del Deporte

Fernández-Ponce, JM [a]; Rodríguez-Griñolo, MR.[b] and Troncoso-Molina, MA[a]

[a] Departamento de Estadística e Investigación Operativa
Universidad de Sevilla (Spain)
Email: ferpon@us.es

[b] Departamento de Economía, Métodos Cuantitativos e Historia Económica
Universidad Pablo de Olavide (Spain)

Resumen

Los órdenes estocásticos o las dominancias estocásticas tal como se conocen en Economía se han estudiado y aplicado ampliamente en diversidad de campos científicos, desde la Biología hasta la Ingeniería de Sistemas. Sin embargo, hasta donde llega nuestro conocimiento hay una laguna de aplicación dentro del ámbito de las Ciencias del Deporte. En este trabajo, pretendemos una primera aproximación a una posible aplicación de los órdenes estocásticos a un conjunto de datos de partidos de fútbol de LaLiga (España). Nuestro objetivo simplemente es mostrar como se puede ampliar una comparativa más allá de una simple comparación de métricas globales o totales. En particular, nos detendremos en los órdenes de dominancia estocástica de primer y segundo orden por ser los más intuitivos y de más sencilla interpretación y por ser los más usados en Economía.

Palabras clave: Análisis de Datos en Ciencias del Deporte, Distribución Beta, Goles Esperados, Órdenes Estocásticos.

Clasificación JEL: C81,C88.

Recibido: 30/11/2021
Aceptado: 27/01/2022

1 Introduction.

Most of the human activities in the past have been based on the comparisons between things to choose the optimal decision taking into account the basic principle of: 'the more profit, the better decision is'. However, a lot of decisions in real life are selected with uncertainty and under risk since some noncontrolled random phenomenons can influence in the final result. Therefore, the optimal decision in this case has to be justified by using the distribution function of the random phenomenons, if it were known. The simplest way of comparing two distribution functions is by the comparison of the associated means. Such a comparison is based on only two single numbers, and therefore it is often not very informative. For example, assume that we want to compare the effectiveness of field-shot for basketball players. For this comparison, the percentage of success in field-shots is to be used. However, players have not the same effectiveness along a season in different circumstances. That is, the effectiveness has a proper variability that can be modelled by using a probability distribution and, in this way the comparison is more complete. Consequently, there exists the need to establish an approach to compare distributions without the use of their parameters such as means, variances, and so on. These kinds of comparisons between distribution functions are called stochastic orders in the literature. These orders have been widely applied in Economics, Biomedicine, Engineering, and different fields in Sciences (see (Shaked & Shanthikumar, 1994)), but in Exercise and Sports Sciences, so far as we know, we only found one paper (see (Damodaran, 2006)). In (Damodaran, 2006), cricket players are viewed as securities and the team as a portfolio. Stochastic orders are referred to as stochastic dominance in the classical economics literature. They have been widely studied, analysed, and applied to different uses. One of the best known manuals in this regard is the book by (Denuit, Dhane, Goovaerts, & Kaas, 2005), among others. A lof of works dealing with this topic in the literature and a search on Matchscinet result in a total of 383 papers from (Hadar & Russell, 1971) to (Kopa, Kabašinskas, & Šutienė, 2022).

Within Sports Analytics, it is widespread to compare sports performance by means of metrics of a similar nature to the expected goals such as expected attendance or expected threat. There are also numerous contributions on this topic in specialised journals and congresses. We find it very difficult and complex to make a selection of papers in this topic because of the large number of them and, of course, because of their relevance in this field. Nevertheless, and taking into account the historical evolution and the relationship with the subject of this article, we believe that the following works can be representative: (Ensum, Pollard, & Taylor, 2004), (Pollard, Ensum, & Taylor, 2004), (Macdonald, 2012), (Ruiz, Lisboa, Neilson, & Gregson, 2015), (Rathke, 2017), (Spearman, 2018), (Tippett, 2019) and (Singh, 2019), among others. However, for those readers who are interested in more information or further information on this topic, we recommend the literature on this subject in the following website (*Expected Goals Literature*, n.d.).

Lately, the use of expected goals in football is becoming increasingly popular in the media and among non-statisticians. However, there is still a long way to go both from a research point of view and in terms of making it a term in the fan's sporting vocabulary. As discussed at the beginning of this section, comparing two sets of data through their respective means may not be sufficient. This is the case with metrics such as expected goals. This is why in this article we intend to extend the interpretation of this type of metric by using stochastic orders. In essence, an expected goal is intended to measure the quality of a shot, regardless of the sport in which it is applied. This quality is measured through the probability of scoring a goal under the conditions in which the shot is taken. For example, in football, there are many factors that influence this probability, such as the distance to the goal, the angle of the shot, the number of defenders between the goal and the shooter, whether the shooter is right or left footed, whether the shot is taken in open play or in any other type of play, the position of the goalkeeper, etc. Depending on these factors, the most commonly used method is logistic regression. Not all data providers use the same factors to make this estimation, and that is why there is a difference between these values depending on which provider gives the information. How we will expand on this interpretation will be the focus of this article.

The article is organised as follows. Section 2 deals with the mathematical foundations of stochastic orders. In particular, we will focus on the usual or first-order stochastic order and the concave increasing or second-order stochastic order. This section also contains the comparison in

these stochastic senses of beta distributions as well as mixed betas which play an important role in the aim of the article. In Section 3, we explain the analysed dataset from LaLiga (Spain) and the provider who provided us with the data. Section 4 deals with the results obtained and the explanation of the theoretical results by means of two cases extracted from these data. Finally, we conclude the article with a section devoted to a discussion of the results obtained and another section with conclusions and possible future research lines.

2 Mathematical backgrounds

In this section, useful mathematical concepts which will be used later are explained. In short, we recall the notion of stochastic orders that have been widely used in different scientific areas and we also give the beta distribution and its properties which are used in Bayesian Data Analysis (BDA).

2.1 Stochastic Orders

The first and the most intuitive notion of stochastic order is called the usual stochastic order. The basic idea of this order is about what distribution has the higher quantiles. This concept is defined as following.

Definition 2.1. Let X and Y be two random variables with distribution function F and G, respectively. Then, X is said to be smaller than Y in the usual stochastic order (denoted by $X \leq_{st} Y$) if

$$E\left[\phi(X)\right] \leq E\left[\phi(Y)\right] \text{ for all increasing functions } \phi : \mathbb{R} \longrightarrow \mathbb{R}, \tag{1}$$

provided expectations exist.

This order is known in Actuarial Sciences as the first stochastic dominance (see (Denuit et al., 2005)). A simple sufficient condition which implies the usual stochastic order is now given. Let $a(x)$ be defined on I, where I is a subset of the real line. The number of sign changes of a in I is defined by

$$S^{-}(a) = \sup S^{-}\left[a(x_1), a(x_2), \ldots, a(x_m)\right], \tag{2}$$

where $S^{-}(y_1, y_2, \ldots, y_m)$ is the number of sign changes of the indicated sequence, zero terms being discarded, ant the supremum in (2) is extended over all sets $x_1 < x_2 < \cdots < x_m$ such that $x_i \in I$ and $m < \infty$. Then, the following theorem can be found page 10 of (Shaked & Shanthikumar, 1994).

Theorem 2.2. *Let X and Y be two random variables with (discrete or continuous) density functions f and g, respectively. If*

$$S^{-}(g - f) = 1 \text{ and the sign sequence is } -, +,$$

then $X \leq_{st} Y$.

Another important stochastic order is the well-known as *the increasing concave [convex] order*. The definition is as follows.

Definition 2.3. Let X and Y be two random variables. Then, X is said to be smaller than Y in the increasing concave [convex] order (denoted by $X \leq_{icv[icx]} Y$) if

$$E\left[\phi(X)\right] \leq E\left[\phi(Y)\right] \text{ for all increasing concave [convex] functions } \phi : \mathbb{R} \longrightarrow \mathbb{R}, \tag{3}$$

provided expectations exist.

For more details about this order see (Shaked & Shanthikumar, 1994). Similarly to Theorem 2.2, it can be proved that if X and Y are two random variables with distribution functions F and G, respectively, and with finite means such that $E(X) \leq E(Y)$, and $S^{-}(G - F) \leq 1$ and the sign sequences is $-, +$ when equality holds, then $X \leq_{icv} Y$, (see Theorem 4.A.22 in (Shaked & Shanthikumar, 1994)). Furthermore, assume that the random variables X and Y are defined on a bounded support and they are nonnegative. Hence, it can be shown that if f and g are the density functions of X and Y, respectively, then $S^{-}(f - g) = 2$ and the sign sequences is $+, -, +$

implies that $S^-(G-F)=1$ and the sign sequences is $-,+$. The proof of this result is trivial by using Theorem 3.A.44 and Theorem 4.A.22 in (Shaked & Shanthikumar, 1994). It is well-known that the *icv* ordering is a partial ordering. To avoid this fact, we define an ordering based on an inequality measure widely connected with the Gini Index. Let CV_X be the coefficient of variation for the random variable X, that is $CV_X=\sqrt{var(X)}/E(X)$, and similarly for CV_Y.

Definition 2.4. Let X and Y be two random variables with finite variance. It is said that Y is better than X in variation sense (denoted by $X\leq_{cv}Y$)if $CV_Y^2\leq CV_X^2$.

Remark 1. The variation ordering has interesting properties.

1. $X\leq_{cv}Y$ if, and only if $cX\leq_{cv}dY$ for all c,d in $\mathbb{R}$.
2. If $a>0$ and $a\neq E(X)$ then $X\leq_{cv}a+X$.
3. If $E(X)\neq 0$ then $X=_{cv}\frac{X}{\sigma_X}$, where $var(X)=\sigma_X^2$.

From now on, we use the notation $\leq_*$ when the $\leq_{st}$ or the $\leq_{icv[icx]}$ orderings will be referred. Most of real-life random variables might have been generated from a mixture of several distributions and not a single distribution. The mixture distribution is a weighted summation of n distributions $\{g_1(x;\theta_1),\cdots,g_n(x;\theta_n)\}$ where $\{\omega_1,\cdots,\omega_n\}$ are the corresponding weights . As is obvious, every distribution in the mixture has its own parameter $\theta_i\subset\mathbb{R}^{k_i}$. The mixture distribution is formulated as:

$$f(x;\theta_1,\cdots,\theta_n)=\sum_{i=1}^{n}\omega_i g(x;\theta_i),\ \text{subject to}\ \sum_{i=1}^{n}\omega_i=1.$$

For commodity, we also denote a mixture distribution by using random variables: $X=\sum_{i=1}^{n}\omega_i X_i$. The distributions can be from different families, for example from beta and normal distributions. However, this makes the problem very complex and sometimes useless; therefore, mostly the distributions in a mixture are from one family (e.g., all beta distributions) but with different parameters.

It is immediately obtained the following results:

1. Assume that X and Y are two mixture distributions with the same weighting parameter and with $n=2$. Let $X_1(Y_1)$ and $X_2(Y_2)$be the corresponding components for $X(Y)$, respectively, in such weighted summation whose parameters have the same dimension. By using Theorem 1.A.3 and Theorem 4.A.8 in(Shaked & Shanthikumar, 1994), if $X_i\leq_* Y_i$ for $i=1,2$ then $X\leq_* Y$.
2. Assume that $X=p_1X_1+(1-p_1)X_2$ and $Y=p_2X_1+(1-p_2)X_2$ two mixture distributions. If $p_1\leq p_2$ and $X_2\leq_* X_1$ then $X\leq_* Y$. This result is shown by using definitions 2.1 and 2.3.

2.2 The Beta distribution

The beta distribution plays an important role in BDA (see (Ma & Leijon, 2011)). For example, the beta distribution can be used to describe the initial knowledge concerning the probability of success, such as the probability that a player gets a successful field shot in basketball. The beta distribution is a suitable model for the random behavior of percentages and proportions. It has also been used as applications such as time allocation in project management and control systems (see (Hahn, 2008)). The definition is as follows.

Definition 2.5. Let X be an univariate random variable with density function f. Then, X is said to have a beta distribution with parameters a and b (denoted as $X\sim Be(a,b)$), if

$$f(x;a,b)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}\ \text{for}\ x\in[0,1],$$

where $\Gamma(\cdot)$ is the gamma function.

Now, let $X \sim Be(a_1, b_1)$ and $Y \sim Be(a_2, b_2)$ be two beta distributions with $a_1 \leq a_2$ and $b_1 \geq b_2$, and a_i and b_i are non-negative for all $i = 1, 2$. Consequently, the function $\phi(x) = x^{a_2-a_1}(1-x)^{b_2-b_1}$ is strictly increasing with $\phi(0) = 0$ and $\phi(1) = +\infty$. Hence, the equation

$$\frac{B_1}{B_2}\phi(x) = 1, \tag{4}$$

where $B_i = \frac{\Gamma(a_i)\Gamma(b_i)}{\Gamma(a_i+b_i)}$ for $i = 1, 2$, has only one solution. It is easily shown that

$$S^-(g - f) = S^-\left(\frac{B_1}{B_2}\phi(x) - 1\right).$$

Then, it is obtained that $S^-(g - f) = 1$ and the sign sequences is $-, +$. Therefore, by applying Theorem 2.2, $X \leq_{st} Y$ holds. Using the same reasoning, in the case in which $a_1 < a_2$ and $b_1 < b_2$, and if it is fulfilled that

$$\frac{B_1}{B_2}\phi(x_0) \leq 1, \text{ where } x_0 = \frac{a_2 - a_1}{(a_2 - a_1) + (b_2 - b_1)}, \tag{5}$$

you would also have the usual stochastic order. That is, $X \leq_{st} Y$.

Now, assume that $X \sim Be(a_1, b_1)$ and $Y \sim Be(a_2, b_2)$ with $a_1 < a_2$ and $b_1 < b_2$, but equation (5) is not verified. Consequently, the function $\phi(x)$ is a concave function on $[0, 1]$ and $\phi(0) = \phi(1) = 0$. Hence, there exist at most two solutions for the equation (5) with $x \in (0, 1)$ since the function $\phi(\cdot)$ depends on two density functions, f and g.

Therefore, $S^-(g - f) = 2$ and the sign sequences is $-, +, -$. Hence, $S^-(G - F) = 1$ and the sign sequences is $-, +$ if the equality holds (see the proof of Theorem 3.A.44 in (Shaked & Shanthikumar, 1994)). Furthermore, if $a_1/b_1 < a_2/b_2$ (that is, $E(X) < E(Y)$) then $X \leq_{icv} Y$ (see Theorem 4.A. 22 in (Shaked & Shanthikumar, 1994)). In the case that $a_1/b_1 > a_2/b_2$, we similarly obtain that $Y \leq_{icx} X$. Furthermore, the random variables can be ordered in terms of the coefficient of variation defined as $CV(X) = \sqrt{var(X)}/E(X)$. That is, if $a_1/b_1 < a_2/b_2$, and $a_1 < a_2$, and $b_1 < b_2$, hold then $CV(X) > CV(Y)$. Consequently, this fact can be interpreted as Y is better than X in variation sense since the mean of Y is greater than the mean of X, and Y has less relative variability than X.

Nevertheless, beta distributions are not as usual in Sports Analytics as we can believe. The most realistic common situation is given by a mixture of distributions due to a unimodal density that does not model too well real circumstances. For this reason, the definition of a mixture of beta distributions are now given.

Definition 2.6. It is said to be that the random variable X follow a mixture of beta distributions (denoted by $X \sim MB(n, \mathbf{p}; \mathbf{a}, \mathbf{b})$) if the corresponding density function is given by

$$f(x; \mathbf{p}, \mathbf{a}, \mathbf{b}) = \sum_{i=1}^{n} p_i Be(a_i, b_i), \tag{6}$$

where $\mathbf{p} = (p_1, \cdots, p_n)$ with $\sum_{i=1}^{n} p_i = 1$ and $p_i \in (0, 1)$; $\mathbf{a} = (a_1, \cdots, a_n)$ with $a_i > 0 \quad \forall i$, and $\mathbf{b} = (b_1, \cdots, b_n)$ with $b_i > 0 \quad \forall i$; and the density function of a beta distribution with the corresponding parameters is denoted by $Be(a_i, b_i)$.

Several estimation procedures based on EM algorithms have been proposed in the literature (see (Ghojogh, Ghojogh, Crowley, & Karray, 2020) and (Schröeder & Rahmann, 2017), among others). It is not our purpose here to obtain new methods for parameter estimation of the parameters in a mixture of beta distributions, but it does establish at least a sufficient condition to be able to compare them in a stochastic sense.

Corollary 2.7. *Let X and Y be two random variables with distribution $MB(2, \mathbf{p}_1; \mathbf{a}, \mathbf{b})$ and $MB(2, \mathbf{p}'; \mathbf{a}', \mathbf{b}')$, respectively. If $p_1 \leq p_1'$, and $X_2 \leq_* X_1$, and $X_i \leq_* Y_i$ for $i = 1, 2$, then $X \leq_* Y$.*

Proof. The proof is immediately obtained by using the results in page 3. Furthermore, for more clarity in the proof, we assume that $X = p_1X_1 + (1 - p_1)X_2$, and similarly for Y. □

3 Material and Methods

3.1 Dataset and variables

In this study, match data were collected during the 2018-2019 and 2020-2021 LaLiga (the first division in Spain) seasons from the database provided by `https://understat.com/`. We have decided not to analyse the data from the 2019-2020 season due to the special circumstances caused by the Covid19 pandemic. This meant that only 27 matchdays were played up to 8 March. Subsequently, it was not resumed until 11 June, with all that this entailed in terms of types of training. Many players had to carry out physical training sessions in their own homes and when they returned to training together, they did so in groups due to security measures. Therefore, we decided not to include this season's data in this study. Consequently, we only scrapped the expected goal metric for each shot and for the team during both seasons. Particularly, we analyze the following team: `Atlético de Madrid (ATM), FC Barcelona (FCB), Real Betis (RBB), Getafe CF (GCF), Real Madrid (RM), Real Sociedad (RS), Sevilla FC (SFC), Valencia CF (VCF) and Villarreal CF (VIL).` For the scrapping process we use scripts from `https://github.com/ewenme/understatr.`

Initially, the scrapped file from `understatr` has 20 variables, but we select the following for our analysis:

- `xG,` is the value of the expected goal metric for the corresponding shot.
- `situation,` game action causing the shot to goal. It could be one of the following: Direct Free kick, From Corner, Open Play, Penalty, and Set Piece.
- `hteam, ateam,` names of the home and away teams, respectively.
- `ha,` is a factor variable indicating if the shot was made by the home team (`h`) or the away team (`a`).
- `year,` is a binary factor variable indicating the the year of the season when the shot was made. If the value is 2018, then the corresponding season is 2018-19 and if the value is 2020, then the season is 2020-21.
- `result,` the circumstance after the shot. It could be one of the following: Blocked Shot, Goal, Missed Shots, Own Goal, Saved Shot, and Shot On Post.
- `lastAction,` action before the shot.

We will ony analyze this dataset for shots which are made from an Open Play and remove those shots from Own Goal and Rebounds. This last modality belongs to a variable which is named `lastAction` in the dataset.

3.2 Data

Each one of the teams analyzed played 38 games for each season. All shots, verifying the restrictions before explained, are considered. It is obtained that 3140 and 2489 shots were made in the 2018-19 and in the 2020-21 seasons, respectively. It should be noted that the 2020-21 season is the season that has been played with the stadium attendance restriction. It is not the intention here to make a comparison to see if the effect of playing at home has diminished due to the support of the fans and the extra motivation, as due to the two-year difference between the coaches change and even the players of the teams themselves. However, it is curious to see the possible effect of playing under very special and new conditions due to the global pandemic by COVID-19. In any case, we can see these differences in total shots and goal percentages during both seasons in Table 1.

Team	Home		Away	
	2018-19	2020-21	2018-19	2020-21
ATM	157(13.4)	187(17.7)	133(8.3)	149(10)
FCB	248(13.3)	232(13.4)	162(17.3)	203(14.3)
GCF	131(15.3)	116(10.3)	114(12.3)	101(6)
RBB	152(9.9)	137(9.5)	104(13.5)	149(8.7)
RM	223(10.8)	222(11.3)	161(8)	159(13.8)
RS	141(9.9)	159(13.2)	101(13.9)	132(9)
SFC	225(12)	186(10.2)	166(9.6)	143(11.9)
VCF	187(8)	117(15.4)	140(12.1)	97(11.3)
VIL	182(9.9)	146(11.7)	158(6.7)	117(18.8)

Table 1: Total shots and goal percentage in open play for each team during 2018-19 and 2020-21 seasons.

3.3 Statistical Analysis

First, we analyze a mixed model for this data assuming that `year` and `ha` are fixed effects, and `team` is a random effect. However, this model does not fit well due to the asymmetry of the residuals. This asymmetry is originated from the proper asymmetry of the expected goal (`xG`) variable. Furthermore, and by using diagnostic plots, the FCB team provokes that assumptions about constant variance of random effects and normality of random effects are not verified. We are not interested to remove FCB team from our data looking for a better mixed model since we think that FCB is one of the best team in the LaLiga and their game style gave them several European soccer championships in the last years. Another possible solution could be to use a Box-Cox transformation to get normality in the residuals, but asymetry is so strong that it is not corrected. Consequently, we will analyze this data under a nonparametric paradigm. We apply the Kruskal-Wallis test to detect differences in the `xG` variable by taking into account the season depending on if a team plays at home or away. These results can be viewed in Table 2. In terms of these differences, the `xG` variable is grouped or not.

Afterwards, we estimate the density function in every case by using mixed beta distributions. To do so, we will explain the procedure we have carried out to estimate the corresponding parameters of the mixed distribution of betas. First, a cluster analysis has been performed on the variable `xG` using the k-means algorithm. Specifically, we have used the R package `factoextra` by means of the command `fviz_nbclust` . We have also checked the results using two methods available as an option in this command: the wss method and the silouhette method. In both cases, the optimal number of clusters in all cases that we will present later in the results section has been two clusters. Once we had an estimate of the number of clusters, we proceeded with the estimation of the beta parameters as well as the corresponding weights. This has been done using the R package `maxLik` as can be seen in the web `https://rpubs.com/MatthewPalmeri/646676`.

4 Results

The different results obtained after data analysis are shown below. First, we now comment on the results shown in Table 2. This table contains the p-values corresponding to the Kruskal-Wallis test. In such a way that we confront whether the team plays home or away versus season. That is, the first column of Table 2 compares the `xG`s of each team when they play at home in the 2018-19 and 2020-2021 season. For example, the teams that have a significant difference are ATM, GCF, SFC, and RS which has a p-value equals to 0.04. The second column deals with the comparison of the seasons when the teams play away. The third and fourth column is the comparison between whether they play as home and away, leaving the season factor fixed. Thus, for example, in the 2020-2021 season, we can see that RM and VIL are the only teams that have had significantly different results.

Taking into account the results of Table 2, the distributions to be compared for each team are

decided. For example, for FCB there are no significant differences between the crossovers, then they are considered as a single distribution grouping all the values of the `xG`s in a single variable.

To simplify the interpretation, we will focus only on the comparison between RM and FCB and the comparison of RBB when playing away in both seasons. The corresponding estimates can be seen in Table 3. For example, the first column refers to the type of crossover between factors. Thus, the type denoted by `H`, means when the team plays as a home team grouping all seasons. In this case, we would have a two-component beta mixture, so that the parameter values of each beta appear in the columns labeled `shape1` and `shape2`. That is, the first component of the beta mixture for the RBB when playing at home is a beta whose parameters are 2.272 and 45.599. The sixth column of the table refers to the weight in the mixture of each component. At this point, it is important to clarify that this weight does not have to coincide with the percentage of data in the corresponding range of the mixture component, as it can and does happen that the intersection of the ranges is not empty. However, we can state that this weight is a very approximate value to the percentage of observations in each range. It is obvious that in all tables the component that has the greatest weight by far outweighing the rest of the components is the first one, i.e. the one corresponding to the lowest value of the `xG`s, which makes sense since during a match very few shots on the goal are taken with a value of `xG` greater than 0.2.

Team	Home	Away	2018-19	2020-21
ATM	0.0055	0.0424	0.5977	0.2981
FCB	0.651	0.5856	0.737	0.1407
GCF	0.007397	0.1912	0.6397	0.3524
RBB	0.5642	0.04446	0.1846	0.8575
RM	0.1042	0.1386	0.2922	0.044
RS	0.04	0.086	0.3769	0.1631
SFC	0.021	0.8641	0.0974	0.1992
VCF	0.962	0.01092	0.4235	0.1126
VIL	0.9318	<0.001	0.4649	<0.001

Table 2: p-values for Kruskal-Wallis test.

FCB	Type	Components	Shape 1	Shape 2	Weight
		1	2.6682	41.4349	0.775
		2	4.0694	5.4425	0.225
RM		1	2.535	45.243	0.8
		2	2.152	3.188	0.2
RBB	H	1	2.2720	45.599	0.8226
		2	4.668	6.972	0.1774
	A-2018	1	2.3142	38.8693	0.8171
		2	12.1056	14.4257	0.1829
	A-2020	1	2.2892	45.7960	0.8586
		2	7.1633	9.1513	0.1434

Table 3: Estimates of mixed beta distributions of the RBB `xG`s.

4.1 Case 1

In this example, we are going to perform a stochastic comparison of the `xG` of the RM and those of the FCB. This type of comparison, as we have already seen and interpreted in Section 2, allows us to broaden the understanding of the data without being at all a complete ordination. In this specific case, the first comparison that is usually made in media and social networks is through

averages or absolute values. Let us take as an example that RM in the 2018-2019 season got a total of 49 goals in open play and its `xG` was 50.2, while FCB got an `xG` of 65.6 and got 64 goals. As described in Section 1, we cannot forget that the expected goals correspond to the mean of the Poisson binomial distribution, so comparing two `xG` in global terms would be equivalent to comparing two means with the risk thet they are nonrobust measures. In any case, for both teams, the total number of goals scored in open play is practically equal to their expected goal value. However, it appears from this overall comparison that FCB's performance was somewhat better than RM's in the 2018-2019 season. Recall that in that season, FCB was LaLiga champion with a total of 90 goals and RM was the third with a total of 63 goals. However, in the 2020-2021 season, RM came second with a total of 67 goals and FCB came third with 85 goals, of which in open play were 35 and 57, respectively. All this means that currently the media and social networks often use average metrics or total measures to make team performance comparisons with the danger that this entails as they are not robust and to some extent uninformative measures of reality. For this reason, a stochastic comparison between the corresponding distributions is proposed, the interpretation of which can be seen in Section 2. However, it should be clarified that stochastic comparisons are not complete orders but partial orders. That is, it may happen that we have two distributions that are not comparable in some stochastic sense.

From Table 3, we can express the mixture of the beta distributions of the FCB as follows according to the notation given in Section 2, page 3:

$$X = 0.775Be(2.6682, 41.4349) + 0.225Be(4.0694, 5.4425).$$

Similarly, we obtain the mixture of beta distributions for the RM

$$Y = 0.8Be(2.535, 45.243) + 0.2Be(2.152, 3.188).$$

If we plot the density functions of both mixtures (see Figure 1), it can be seen how these functions cross at least four times which implies that they cannot be compared stochastically until at least up to fourth-order stochastic order. In this paper, and because of its interpretations in economics, we only propose up to order 2, i.e. the usual stochastic order ($\leq_{st}$) and the increasing concave order ($\leq_{icv}$), also called stochastic dominance of the first and second order, respectively. This result was expected because of the shape of the curvature of a mixture that has two modes can hardly be compared stochastically in a global way.

This circumstance, which will occur very frequently, forces us to take a position on the matter to be able to carry out a stochastic comparison. Applying the results of Section 2, it is very simple to obtain that $Y_1 = Be(2.535, 45.24) \leq_{ST} X_1 = Be(2.66, 41.43)$. This implies from the outset that the distribution of the low value RM `xG`s is smaller than FCB `xG`s in some stochastic sense. And furthermore, as a consequence of the usual stochastic order, any quantile of the distribution of the RM's `xG` will be less than or equal to the corresponding quantile of the distribution of the FCB's `xG`.

Similarly, checking the conditions in Section 2, we obtain that $Y_2 = Be(2.152, 3.188) \leq_{icv} X_2 = Be(4.06, 5.44)$. This means that for high RM's `xG` has lower and more variable values than those corresponding to the FCB in some stochastic sense.

Due to the impossibility of a comparison for the mixtures of the `xG` of both teams, we will make a comparison in terms of the coefficient of variation according to Definition 2.4. By a simple calculation and taking into account the weights, we obtain that the CV of RM is 4.637 and the corresponding one of FCB is 5.428. Therefore, we can interpret that the `xG` of RM is better than that of FCB in the sense of stochastic variation, which does not mean that RM has higher mean and lower variance as it is the case. However, it is verified that the variance of the RM's xG is about half the variance of the FCB's xG.

From now on, we will distinguish between these two clusters in the following way: low xG for those clusters below 0.15 and high xG for those above 0.15. This division coincides with the two components of the mixture of the beta distribution. In short, RM took a total of 139 shots on goal in the two seasons with an `xG` greater than 0.15, and of these 53 were goals. This means an effectiveness rate of approximately 38.13%. FCB, however, took a total of 189 shots on goal with an `xG` above 0.15 and scored 70 goals, i.e. 37.03% effectiveness. Admittedly, this is not a significant difference. And, as you select the shots with less `xG` and calculate the goal percentage,

RM is almost as effective as FCB (7.77% and 5.59%, respectively). This fact is not a contradiction with the orderings analysed since these orderings only take into account the value of the expected goal and not the result of the shot. The information provided by the stochastic order can be interpreted in some way as a pattern of play, both in terms of the team's offensive play and the defensive system employed by the opponents.

4.2 Case 2

Below is a comparison for RBB when they played away in the 2018-19 and 2020-21 seasons. One of the questions that may arise in this analysis is whether there are significant differences between the two squads and the tactics and strategies employed by two different coaches when playing at home. As far as the `xG` values between the two seasons are concerned, there are no significant differences (p-value=0.5642 in the Kruskal-Wallis test, see Table 2). And as for the players' squads, the website Transfermarket, which specialises in valuing teams, we can say that both valuations are very similar. We think that extrapolating beyond the `xG` values to interpret the context of both seasons can lead to misinterpretations. That is why for a more in-depth analysis of the game it is necessary to use the video analysis that is already widely used by technical staff in football today.

Having clarified the context of the RBB during those two seasons, let us now analyse the `xG`s produced as visitors from the point of view of stochastic orders. To do so, looking at Table 3, we see how the `xG`s of the RBB can be modelled as a mixture of two-component beta distributions for both seasons. From Table 3, we can express the mixture of beta distributions for the RBB when playing at home jointly during 2018-19 and 2020-21 seasons as follows according to the notation given in Section 2, page 3:

$$X = 0.8226Be(2.272, 45.599) + 0.1774Be(4.668, 6.972)$$

Similarly, we obtain the mixture of beta distributions for the RBB when playing away for each season separately,

$$Y = 0.8171Be(2.3142, 38.8693) + 0.1829Be(12.1056, 14.4257) \text{ and}$$
$$Z = 0.8586Be(2.2892, 45.796) + 0.1434Be(7.1633, 9.1513),$$

where Y is for the 2018-19 sesason and Z for the 2020-21 season.

It is easy to obtain the following stochastic comparisons between the components of the mixtures:

$$Be(2.3142, 38.8693) \leq_{ST} Be(12.1056, 14.4257) \text{ and } Be(2.2892, 45.796) \leq_{ST} Be(7.1633, 9.1513),$$

$$Be(2.2892, 45.796) \leq_{ST} \text{ and } Be(7.1633, 9.1513) \leq_{icv} Be(12.1056, 14.4257).$$

These partial orders between the components may lead us to believe that the icv order between the mixtures will be verified. However, this circumstance does not occur due to the number of crossings between the respective density functions as can be seen in Figure 2.

Moreover, in Figure 3a and 3b it is clear how the two mixed distributions cannot be compared in the sense of first and second-order stochastic dominance since the respective density functions cross at least four times. This forces us to compare by components. In fact, in Figure 3a, it can be clearly observed how the components with the lowest `xG` values of the 2018-19 season stochastically dominate in first order to the corresponding ones in the 2020-21 season. This fact is easily demonstrated by comparing the corresponding parameters of the beta distribution in the first component of the mixture, since they satisfy the criterion given in Theorem 2.2. This implies not only a higher mean but also that the quantiles of any order are higher for Quique Setién's team. Nevertheless, and even though they were seasons without significant differences in the low `xG`s.

For the second component, the one for the 2018-19 season and the one for the 2020-21 season, it is obtained in a simple way by applying the results of section 3 that $Be(7.1633, 9.1513) \leq_{icv} Be(12.1056, 14.4257)$. That is, the values of the `xG`s between 0.15 and 0.8 for the 2020-21 season are lower than the corresponding values for the 2018-19 season and more variable in some stochastic

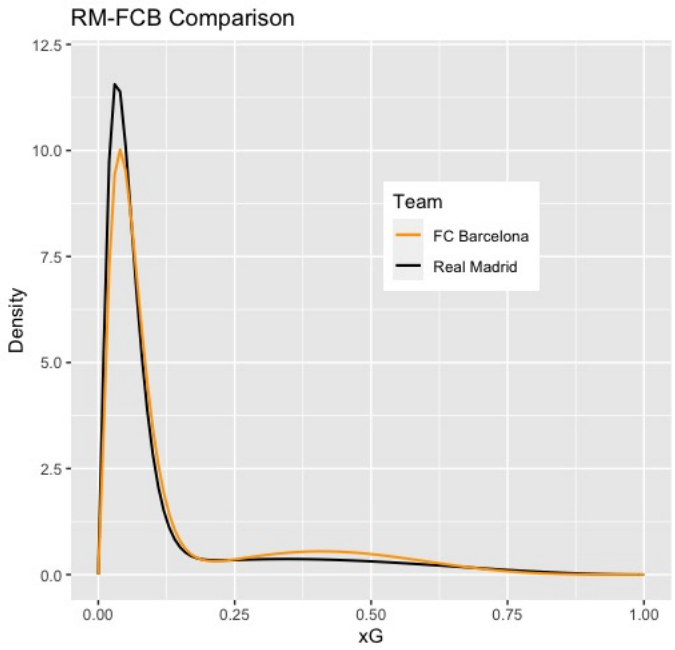


Figure 1: A comparison between FCB and RM

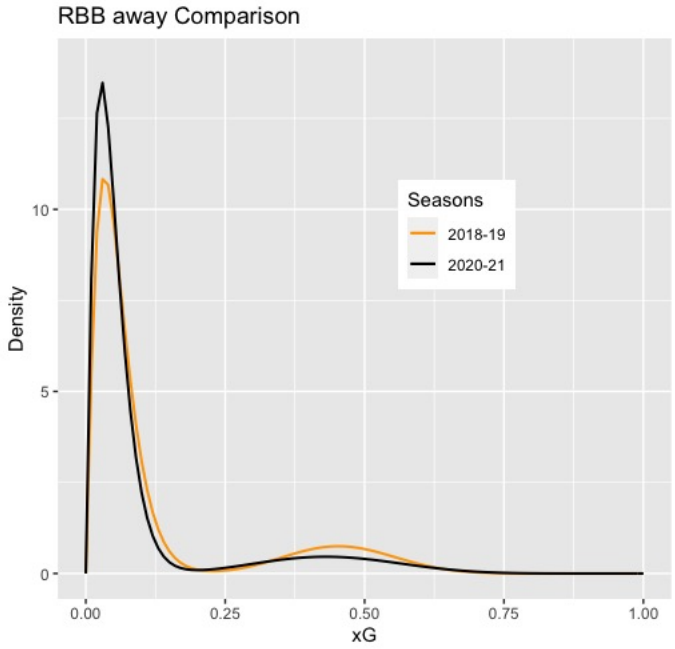
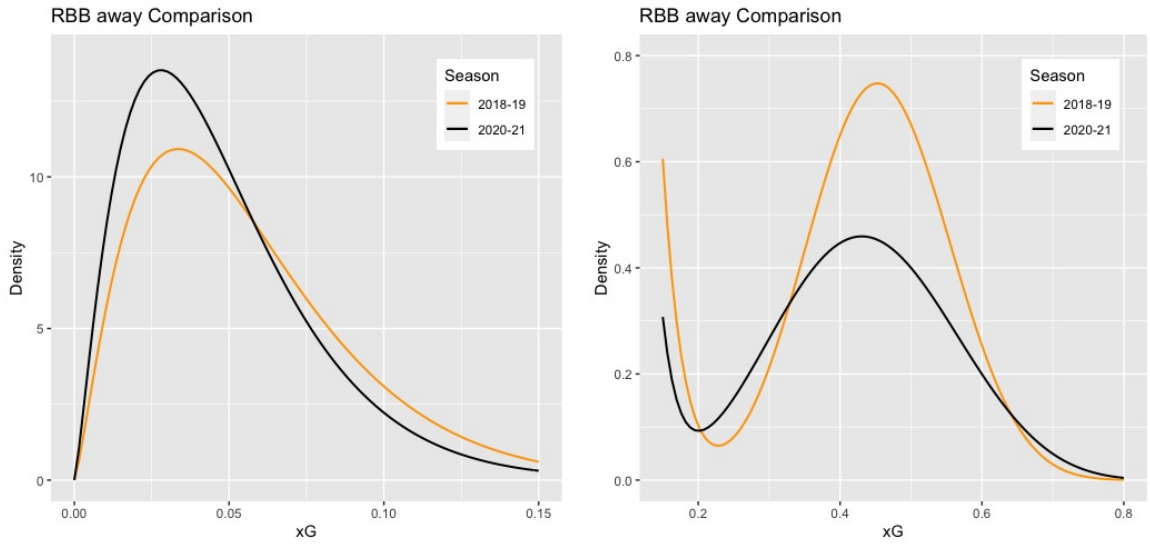


Figure 2: RBB away comparison during 18-19 and 20-21 seasons



(a) Density Functions for `xG` <0.15

(b) Density Functions for 0.15< `xG`< 0.8

Figure 3: RBB comparison as an away team in Seasons 2018-19 and 2020-21

sense. However, if we analyse the effectiveness (understanding effectiveness as up to now, the goals scored) between these `xG` values, we obtain that in Setién's season RBB shot at goal in open play a total of 20 times and scored 9 goals (45%), while in Pellegrini's season, RBB shot at goal in open play a total of 21 times and scored 6 goals (approximately 28.57%).

For the case of the first component, we would have that the one for the 2018-19 season dominates the one for the 2020-21 season in the usual stochastic sense. That is, it has a higher mean in some stochastic sense. This first component is for `xG` values lower than 0.15. For the Setien's team, there were 84 shots in open play scoring 5 goals (approximately 5.9% effectiveness) and for Pellegrini's team there were a total of 128 shots scoring 7 goals (approximately 5.4% effectiveness).

Therefore, we can generally conclude that RBB when away in the 2018-19 season created more quality scoring opportunities than in 2020-21 and with more effectiveness, which is not a necessary basis for a better final ranking in LaLiga. In the 2018-19 season, the RBB was in tenth place with only 50 points and in the 2020-21 season it was sixth with 61 points. The explanation for this can be found in the variance of the number of goals per game, which indeed implies that the variance of the number of goals per game in the 2020-21 season should be higher than in the 2018-19 season. Regardless of this fact, and without going into subjective assessments of both seasons, it is a known fact that Pellegrini's team plays a more attractive game for the spectator and with better results in general.

5 Discussions

In this paper, we have presented a stochastic comparison applied to sports analytics. It is well known that stochastic orders have many applications in the field of economics. However, these types of orders are of a partial nature, which may imply that the given two probability distributions are not comparable in some sense. Added to all this is the fact that the expected goals metric is widely discussed and questioned in terms of the information it provides, since it can be considered out of context of the game. However, it is no less true that we wanted to give another view of the classical interpretation provided by these metrics of the average type, trying to get a little more information. It should also be taken into account that the providers of soccer match data use different methods to estimate expected goals, so the results provided here may differ if we use data from other companies. It is not our aim here to question the quality of the data we have used, let alone claim that it is the most accurate on the market. We simply want to show one more

application that can be easily extrapolated to any data set with the same characteristics. Having clarified this point, we would like to emphasize the fact that some assumptions that we have used in the study can, of course, be debated and criticized. We refer to the assumption of independence of the shots and that due to the nature of the expected goals, the highest percentage will always occur in low values. Looking at the corresponding tables, in the first component, in almost all teams they are higher than 75%. This may raise the question of analyzing high values of the `xG` as possible outliers in the total set and even eliminating them from the study if they exist. We have avoided such a situation simply for simplicity of analysis since the objective of the study was stochastic comparison. Clearly, we could have kept only the first component and from that we could have performed the stochastic comparisons, but we considered it more interesting to see the complexity involved in stochastic dominances in the case of mixed distributions as in this case.

We also want to clarify that although we have provided data corresponding to 9 teams, we have only analyzed two cases, one corresponding to the comparison between FCB and RM and the comparison of RBB when they have played away, simply due to the lack of space and not to be too tiring for the reader.

6 Conclusions and future research

The main conclusion that can be drawn from this paper is that the use of stochastic dominances can be useful to obtain information beyond simple means or variances for a Sports Sciences data set. However, it should be clear that it does not completely solve the problem because, as we have already seen, it may happen that we have the non-comparability of variables in some stochastic sense. This gives rise to an open a new field of research in Sports Analytics in the sense of increasing the order of stochastic orders and trying to interpret them in the field of data analysis in sports, even proposing hypothesis tests of stochastic orders that improve the existing ones in the literature.

Acknowledgement(s)

We would like to give our sincere thanks to the support team of `understat.com` for allowing us to use the data from their website.

Additionally, the authors would like to thank to Prof. Martí Casals from University of Vic for his valuable comments on mixed models in Section 3.

Notes on contributor(s)

Miguel Alejandro Troncoso-Molina collaborated in this work while he was doing his final degree project in Mathematics, under the supervision of Professor Fernández-Ponce, at the Universidad de Sevilla.

References

Damodaran, U. (2006). Stohcastic dominance and analysis of odd batting performance: the indian cricket team, 1989-2005. *Journal of Sports Science and Medicine*, *5*, 503-508.

Denuit, M., Dhane, J., Goovaerts, M., & Kaas, R. (2005). *Actuarial theory for dependent risks. measures, orders and models.* John Wiley and Sons.

Ensum, J., Pollard, R., & Taylor, S. (2004). Applications of logistic regression to shots at goal in association football: calculation of shot probabilities, quantification of factors and player/team. *Journal of Sports Sciences*, *22*(6), 504.

Expected goals literature. (n.d.). `https://docs.google.com/document/d/1OY0dxqXIBgncj0UDgb97zOtczC-b6JUknPFWgD77ng4/edit`.

Ghojogh, B., Ghojogh, A., Crowley, M., & Karray, F. (2020). Fitting a mixture distribution to data: tutorial. `arXiv:`*1901.06708v2 [stat.OT]*..

Hadar, J., & Russell, W. (1971). Stochastic dominance and diversification. *JJournal fo Economic Theory*, *3*, 288-305.

Hahn, E. (2008). Mixture densities for project management activity times: A robust approach to pert. *European Journal of Operational Research*, *188*(2), 450-459.

Kopa, M., Kabašinskas, A., & Šutienė, K. (2022). A stochastic dominance approach to pension-fun selection. *IMA Journal of Management Mathematics*, *33*(1), 139-160.

Ma, Z., & Leijon, A. (2011). Bayesian estimation of beta mixture models with variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(11).

Macdonald, B. (2012). An expected goals model for evaluating nhl teams and players. In *Mit sloan sports analytics conference 2012.* Boston, USA..

Pollard, R., Ensum, J., & Taylor, S. (2004). Estimating the probability of a shot resulting in a goal: The effects of distance, angle and space. *International Journal of Soccer and Science,*, *2*(1), 50-55.

Rathke, A. (2017). An examination of expected goals and shot efficiency in soccer. *Journal of Human Sport and Exercise,*, *12*, 514-529.

Ruiz, H., Lisboa, P., Neilson, P., & Gregson, W. (2015). Measuring scoring efficiency through goal expectancy estimation. In *European symposium on artificial neural networks, computation intelligence and machine learning.*

Schröeder, C., & Rahmann, S. (2017). A hybrid parameter estimation algortihm for beta mixtures and applications to methylation state classification. *Algorithms for Molecular Biology,*, *18*, 12-21.

Shaked, M., & Shanthikumar, G. (1994). *Stochastic orders and their applications.* Academic Press, Boston.

Singh, K. (2019). *Introducing expected threat (xt).* (`https://karun.in/blog/expected-threat.html`)

Spearman, W. (2018). Beyond expected goals. In *Mit sloan sports analytics conference 2018.* Boston, USA..

Tippett, J. (2019). *The expected goals philosophy: A game-changing way of analysing football.* Independently published.