ANÁLISE DO APROVEITAMENTO DOS TIMES NO CAMPEONATO BRASILEIRO A PARTIR DE UMA DISTRIBUIÇÃO NORMAL

Alysson Ramos ARTUSO[1]

- RESUMO: Os esportes sempre fascinaram a humanidade e nesse contexto cabe colocar o futebol como fonte de estudo. O objetivo deste artigo é formular um modelo para estimar a pontuação necessária para se alcançar certas posições na classificação final do Campeonato Brasileiro de Futebol, Séries A e B. Foram utilizados os dados de campeonatos passados para provar que o índice de aproveitamento obedece a uma distribuição normal (Gaussiana) de probabilidades e, assim, pode ser usado como parâmetro para os clubes definirem suas metas antes do início de competições. O modelo também é válido, com algumas limitações, para campeonatos brasileiros disputados com diferentes fórmulas ou diferentes sistemas de pontuação e se mostrou eficiente ao ser testado em uma amostra de 10 campeonatos europeus. Aplicações do raciocínio apresentado são possíveis em outros campeonatos ao redor do mundo e também em outros esportes com sistema de disputa similar.

- PALAVRAS-CHAVE: Futebol; aproveitamento; distribuição normal.

1 Introdução

O futebol é hoje o esporte mais popular do mundo, presente em mais de 200 países e com mais de 260 milhões de pessoas que o jogam de maneira regular, segundo dados da *Fédération Internationale de Football Association*, FIFA (2007). Nas décadas recentes, esse esporte foi além de uma forma de divertimento ou manifestação cultural, constituindo uma importante atividade econômica mundial, movimentando bilhões de dólares e sendo responsável, em 2000, por 3% do comércio mundial (Szymanski, 2001).

No Brasil, a maior e mais importante competição é o Campeonato Brasileiro organizado pela Confederação Brasileira de Futebol (CBF) e dividido em Série A, Série B e Série C. As duas primeiras são, atualmente, disputadas por vinte times e possuem a mesma fórmula de disputa.

Apesar de sua enorme popularidade e de sua importância econômica, o futebol brasileiro atravessa várias crises. Dificuldades financeiras e administrativas, corrupção e falta de organização por parte dos clubes e das instituições responsáveis pelo gerenciamento do futebol submetem os times a um calendário em que freqüentemente se disputam dois, três ou mais campeonatos numa mesma temporada. Essa situação afeta o comportamento dos jogadores, dado que o espaço entre dois jogos raramente é suficiente para que os jogadores se recuperem fisicamente. Como conseqüência pode haver um

[1] Departamento de Engenharia da Produção, Mecânica e Ambiental, Centro Universitário Franciscano – UNIFAE, Curitiba, PR, Brasil. E-mail: *alysson.artuso@gmail.com*

desgaste acima do desejável dos atletas, ocasionando um alto número de lesões; por isso alguns clubes usualmente priorizam algumas competições em detrimento a outras. Pegue-se como exemplo os times brasileiros mais importantes, que disputam aproximadamente 100 partidas por ano. Considerando que os atletas têm 30 dias de férias, esses times jogam, em média, uma partida a cada 3,3 dias.

Dessa maneira, estimar a pontuação necessária para atingir determinada posição em um campeonato pode dar suporte aos clubes na formulação de suas estratégias, ajudando-os a otimizar seus recursos e a planejar a preparação física de seus atletas. Também a imprensa se interessa por esses números para informar os fãs de futebol sobre o desempenho e as chances de seus times favoritos.

Tais estudos são relativamente comuns e difundidos, principalmente nas ligas profissionais americanas, e em especial no caso do *baseball*. Mas algumas particularidades do futebol, como a possibilidade do empate, e algumas características do Campeonato Brasileiro, como o equilíbrio entre os times e as mudanças freqüentes de regulamento, não permitem uma extensão simples de tais estudos para esse contexto.

Sendo assim, o objetivo desse artigo é justamente apresentar uma forma alternativa de fornecer a pontuação necessária para se tornar campeão, classificar-se para competições continentais e divisões mais altas ou ainda para escapar do rebaixamento a divisões inferiores.

2 Objetivos, características do Campeonato Brasileiro e proposição do modelo

O objetivo principal é propor um modelo baseado nos resultados passados de campeonatos de pontos corridos. Provando-se que o aproveitamento dos clubes obedece a uma distribuição normal (Gaussiana) univariada pode-se, a partir de conceitos de probabilidade e estatística, estimar valores de aproveitamento necessário para os clubes se planejarem no início da competição.

Até 1994, o Campeonato Brasileiro de Futebol atribuía 2 pontos para o time vitorioso de uma partida, 0 pontos para o time derrotado e em caso de empate ambos somavam 1 ponto cada. Recentemente houve uma mudança no sistema de pontuação em caso de vitória, que hoje vale 3 pontos. Para os cálculos desse artigo será considerado, independente do ano da competição, o sistema de pontuação atual, uma vez que se prove a igualdade no tratamento dos dados, sejam eles anteriores ou posteriores ao ano de 1994.

Ao contrário dos campeonatos europeus, a competição brasileira não segue, ano após ano, as mesmas regras, sendo freqüente a mudança de regulamento de um ano para outro. As fórmulas mais usadas foram a de campeonatos disputados em fase única de turno e returno (Série A após 2003 e Série B 2006) e campeonatos disputados em turno único com fase posterior no sistema de *playoffs* (Série A anterior a 2003 e série B anterior a 2006, com exceções). Em alguns anos houve a exclusão ou inserção de clubes que inicialmente fariam ou não parte do campeonato através de medidas regulatórias da CBF, responsável pelo campeonato brasileiro. Infelizmente também tem se mostrado comum a perda de pontos de alguns times ou a remarcação de jogos devido aos escândalos de arbitragem e à falta de organização de clubes e instituições.

Felizmente, com ações governamentais, nova legislação e gestão mais profissional dos clubes, esse panorama tem mudado. As leis atuais garantem por pelo menos dois anos

as mesmas regras para os campeonatos. Assim, a Série A é disputada por 20 clubes que se enfrentam em turno e returno em um campeonato de pontos corridos. Ao final dos 38 jogos que cada clube faz, o primeiro colocado na classificação sagra-se campeão, os quatro primeiros classificam-se para a Copa Libertadores da América e os quatro últimos são rebaixados para a divisão inferior. A Série B obedece a um regulamento semelhante, com a única diferença dos quatro primeiros, que ao invés de irem para competições intercontinentais se classificam para a Série A do ano seguinte. No total são disputadas 380 partidas em cada divisão.

Dessa forma é coerente pensar num modelo que se ajuste a diferentes tipos de campeonatos ainda que a tendência seja permanecer o atual regulamento, vigente desde 2003 na Série A.

Outros modelos para o Campeonato Brasileiro foram propostos por alguns autores. Destacam-se três diferentes propostas:

1) Cálculo probabilístico de resultados. Apoiando-se na literatura existente, pode-se argumentar que o número de gols marcados por um time em uma dada partida obedece a uma distribuição de Poisson, assim são estudadas as distribuições de Poisson Bivariadas, com destaque para a classe de Holgate, como adequadas para a modelagem conjunta do placar final de uma partida de futebol. Conhecidos os resultados recentes dos times cujo confronto se queira modelar o resultado, há diversos métodos propostos para a estimação dos parâmetros. De posse de um desses métodos que forneçam as probabilidades de ocorrência de placares, pode-se aplicar essas estimativas para calcular, por exemplo, a probabilidade de um determinado time derrotar outro, de uma seleção ser campeã de um torneio ou a quantidade necessária de pontos que uma equipe deve conquistar para passar à próxima fase do campeonato. Tal abordagem é utilizada por Arruda (2000) para suas previsões em relação às competições brasileiras e à Copa do Mundo, o autor é responsável pelo *site* chance de gol (http://www.chancedegol.com.br). Algumas limitações desse modelo são descritas pelo próprio autor (Arruda, 2000, p. 47):

> É importante considerar, contudo, que todos os resultados abordados e alcançados nesta tese (das probabilidades previstas às medidas de calibração) dependem fortemente dos critérios utilizados para a formação dos bancos de dados (...) e dependem também de vários outros fatores, muitos deles essencialmente subjetivos:
>
> - Escolha de jogos: quais competições devem e quais não devem ser consideradas na composição do banco de dados;
> - Inclusão de times: restrição ou não aos jogos que envolvam um ou mais dos times participantes do campeonato, cujos jogos se quer prever;
> - Escolha do sistema de pesos;
> - Determinação da "idade máxima" dos jogos;
> - A pessoa que anunciará as probabilidades (métodos Implícitos);
> - Critérios de "empate técnico" para a "comparação tríplice", além do próprio critério de "comparação tríplice";
> - Discretização escolhida para os valores de p nas curvas de calibração.

2) Programação Linear. Uma segunda possibilidade segue uma abordagem vinda da pesquisa operacional, mais especificamente como um problema de programação linear. Neste caso, são usados algoritmos que simulam todos os resultados possíveis de todos os jogos restantes do campeonato. Dessa forma obtêm-se informações detalhadas sobre as condições exatas de classificação e o desempenho de cada equipe participante. Em

meados dos anos de 1960, publicações abordavam o problema de saber quando um time está matematicamente eliminado na liga de baseball americana (MLB), utilizando algoritmos de fluxo em redes. Na década de 1990, começou-se a utilizar modelos de programação linear inteira e a provar-se alguns resultados teóricos sobre o problema da eliminação. De maneira análoga, pode se trabalhar com o problema de classificação garantida, que consiste em determinar a quantidade mínima de pontos que um time precisa fazer para garantir a sua classificação em um campeonato esportivo. Nesse modelo é calculado o número de pontos necessários a conquistar para garantir a classificação independentemente de quaisquer outros resultados dos adversários. Há maneiras, ainda, de se calcular o número mínimo de pontos a conquistar para ainda se manter chances de classificação, dependendo de resultados de outros times. De diferencial tem-se que os números dados por esse método são precisos, dado que o modelo considera exaustivamente todas as possibilidades de combinação de resultados e que é resolvido através da técnica de programação linear inteira, sendo mais consistente do que informações baseadas em estimativas de probabilidade de vitória. Porém com a desvantagem de superestimar a pontuação necessária, dado que não leva em conta o aproveitamento passado ou a qualidade dos times que se enfrentam. Esse método é utilizado no Campeonato Brasileiro por Ribeiro e Urrutia (2005) que desenvolveram o *site* FutMax (http://www.futmax.org/) para apresentar seus resultados. Outra limitação é que a pontuação necessária dada antes do início do campeonato não é muito esclarecedora e não tem muita serventia como objetivo a ser atingido pelas equipes envolvidas. Em recente entrevista, os próprios autores reconhecem essa limitação (Gouveia, 2006): "No início do campeonato há poucas informações que auxiliem os cálculos. Os dados fornecidos pelo Futmax passam a se tornar interessantes a partir do segundo turno".

3) Técnicas de simulação. O método de simulação Monte Carlo é a terceira alternativa encontrada na literatura científica aplicada à previsão de pontos necessários para se alcançar determinadas posições. Neste caso, a idéia principal é criar um modelo que gere aleatoriamente o número de pontos obtidos por cada time, em cada partida, dado alguns parâmetros prévios. O campeonato inteiro é simulado e os times são classificados de acordo com seu número acumulado de pontos. Silva, Garcia e Saliby (2002) aplicaram essa abordagem ao Campeonato Brasileiro de Futebol. Para a construção do modelo algumas pressuposições foram necessárias, como a igualdade entre todos os times, a independência dos resultados entre os jogos e a probabilidade sempre igual de um jogo terminar empatado. O parâmetro fundamental da simulação é a chance de uma partida terminar em empate. Dessa forma foi feito um levantamento histórico dos campeonatos de 1996 até 2001 da porcentagem de jogos que acabaram dessa forma, sendo esse parâmetro estimado através de uma distribuição triangular. Ao final são dadas as pontuações necessárias, respeitado um nível de confiança, para se atingir determinadas posições dentro do campeonato. Porém, as simulações ficam restritas ao número de competidores dos anos estudados, que são diferentes dos números atuais, e há o fato de não se poder prever nada a respeito das chances de um determinado time atingir determinada posição.

Como dito anteriormente, mudanças no regulamento do Campeonato Brasileiro foram freqüentes, de forma que os três modelos apresentados foram desenvolvidos numa época em que a competição era disputada de maneira bastante diferente, entre 24 ou mais clubes que jogavam em turno único, com os oito primeiros se classificando para fases

eliminatórias e os quatro últimos sendo rebaixados. Assim a pontuação necessária para se atingir determinada colocação perde o sentido quando a competição é composta por até 6 clubes a menos e com um número muito maior de jogos, visto que eles são atualmente disputados em dois turnos.

Visando propor um modelo para o atual campeonato, mas também aplicável para campeonatos com outras fórmulas de disputa e expandindo algumas limitações dos métodos apresentados, inclusive no que se refere às prerrogativas de que necessitam (como independência entre jogos sucessivos), propõe-se, ao invés de se trabalhar com a pontuação final, realizar uma análise dos aproveitamentos[2] necessários para se atingir determinada colocação, tratando-os como variáveis independentes que obedecem a uma distribuição normal $N \sim (\mu,\sigma^2)$ univariada. Assume-se, também, que o campeonato de um ano é independente do campeonato do ano anterior. Com isso, busca-se um modelo simples e confiável em suas predições, com aplicação possível a outros campeonatos e outros esportes.

3 Construção do modelo

Todos os campeonatos brasileiros de futebol em todas as suas divisões foram considerados, desde seu início em 1971, porém várias fórmulas de disputa foram usadas nesse período. O sistema de pontuação de campeonatos anteriores a 1994 foi atualizado para que seguissem o mesmo padrão dos dias de hoje, com empate valendo um ponto e vitória valendo três. Como somente os campeonatos disputados no sistema de pontos corridos são de interesse desse modelo, foram levadas em conta as competições das quais participaram pelo menos 20 times com todos jogando contra todos, num mínimo de 19 jogos, ignorando-se fases posteriores quando existentes. A justificativa é manter uma baixa dependência entre as variáveis, entre um campeonato e outro e evitar comparações entre grupos ou campeonatos que não necessariamente sejam equilibrados (Emonet, 2000). Assim formou-se a seguinte população de campeonatos que atendem aos critérios estabelecidos, de acordo com o ano em que foram disputados: Séries A de 1971, 1972, 1988, 1990, 1991, 1992 e de 1995 até 2006; e Séries B de 1999 e de 2002 até 2006 – totalizando uma população de 24 observações, com os respectivos aproveitamentos mostrados na Tabela 1.

Com o objetivo de utilizar os valores históricos de aproveitamento para fornecer a pontuação necessária de um próximo campeonato, cabe, num primeiro momento, analisar a distribuição de probabilidades que melhor se adeqüa ao índice de aproveitamento do primeiro colocado (chamada de X_1), do quarto colocado (X_2) e do quarto último colocado (X_3) dos campeonatos brasileiros. Tais posições foram escolhidas por representar, respectivamente, a posição de campeão da competição, o último clube a se classificar para a Copa Libertadores da América ou para a Série A e o último clube a ser rebaixado para uma divisão inferior.

Na literatura se observa que os pontos feitos por um time num campeonato de futebol obedecem a uma distribuição normal univariada (Emonet, 2000). Como o

[2]O aproveitamento é calculado pela porcentagem de pontos conquistados em relação ao total de pontos disputados.

aproveitamento é uma combinação linear da pontuação, este também obedece a uma distribuição normal devido às propriedades da própria distribuição (James, 2006).

Tabela 1 - Dados da amostra

	Aproveitamento do campeão (X_1)	Aproveitamento do 4º colocado (X_2)	Aproveitamento do 4º último colocado (X_3)
Série A – 1971	68,421%	56,140%	34,211%
Série A – 1972	64,286%	55,556%	38,889%
Série A – 1988	64,493%	57,246%	36,232%
Série A – 1990	72,464%	53,623%	36,232%
Série A – 1991	69,333%	54,667%	36,000%
Série A – 1992	72,840%	60,494%	33,333%
Série A – 1995	62,500%	56,944%	30,556%
Série A – 1996	69,841%	55,556%	34,921%
Série A – 1997	66,667%	59,420%	33,333%
Série A – 1998	72,000%	60,000%	34,667%
Série A – 1999	63,768%	56,522%	33,333%
Série A – 2000	66,667%	57,971%	34,783%
Série A – 2001	63,158%	52,632%	33,333%
Série A – 2002	64,912%	56,140%	35,088%
Série A – 2003	59,649%	52,632%	33,333%
Série A – 2004	68,116%	57,971%	31,884%
Série A – 2005	66,667%	60,000%	28,000%
Série A – 2006	63,158%	52,632%	35,088%
Série B – 1999	62,281%	53,509%	38,596%
Série B – 2002	65,079%	55,556%	39,683%
Série B – 2003	66,667%	60,870%	33,333%
Série B – 2004	68,116%	53,623%	34,783%
Série B – 2005	68,000%	62,667%	33,333%
Série B – 2006	71,429%	53,968%	38,095%
Média	**66,6879%**	**56,5140%**	**34,6279%**
Desvio Padrão	**3,4940%**	**2,9053%**	**2,6279%**

Fonte: Rec.Sport.Soccer Statistics Foundation (RSSSF) e autor.

Com o intuito de verificar essa afirmação foram aplicados o teste de *Kolmogorov-Smirnov*, o mais utilizado em trabalhos similares, e o teste de *Shapiro-Wilk*, mais adequado para um número pequeno de observações, com a finalidade de se aceitar ou rejeitar a hipótese de normalidade dos dados (Siegel e Castellan, 2006). Os resultados dos testes são apresentados na Tabela 2. Todos os cálculos foram realizados com o auxílio do *software* MATLAB.

Hipóteses testadas:
H_0 = A distribuição de X_i é igual à distribuição normal
H_1 = A distribuição de X_i não é igual à distribuição normal

Tabela 2 - Testes de Normalidade

	Kolmogorov-Smirnov			*Shapiro-Wilk*		
	Statistic	df	***p-value***	Statistic	df	***p-value***
Aproveitamento do campeão (X_1)	0,0931	24	**0,9202**	0,9719	24	**0,7148**
Aproveitamento do 4º colocado (X_2)	0,1011	24	**0,9573**	0,9459	24	**0,2209**
Aproveitamento do 4º último colocado (X_3)	0,1864	24	**0,3402**	0,9399	24	**0,2749**

Com base na Tabela 2 não se pode rejeitar a hipótese nula de normalidade para nenhuma das as variáveis, pois ao se observar o *p-value,* em todos os casos ele está acima do nível de significância de 0,05. Assim as três variáveis serão tratadas como distribuições normais.

Há ainda mais alguns requisitos a serem confirmados, se há diferença significativa entre campeonatos brasileiros da Série A e B (caso 1), se há diferença significativa em função do sistema de pontuação (caso 2), e se há diferença significativa devido à fórmula de disputa do campeonato (caso 3). Como subconjuntos de uma distribuição normal também apresentam distribuição normal (James, 2006), testes de normalidade nesses subconjuntos não foram efetuados.

Um teste comum para a igualdade de tratamentos (médias) é o teste t de *student* (Marques e Marques, 2005), o teste, tal como será aplicado, exige como pré-requisitos a normalidade, independência e homocedasticidade dos dados. O primeiro pré-requisito já foi cumprido. O segundo também, por hipótese inicial do trabalho que assume independência entre campeonatos de anos diferentes, uma suposição bem mais branda do que o de outros modelos que tratam jogos sucessivos como independentes. A igualdade entre as variâncias (homocedasticidade) pode ser testada através do teste F de igualdade entre duas variâncias (Mood, Graybill e Boes, 1974).

A Tabela 3 resume a aplicação do teste F de igualdade entre duas variâncias para os três casos descritos anteriormente.

Hipóteses testadas:

H_0 = As variâncias são iguais.

H_1 = As variâncias não são iguais.

Tabela 3 - Teste bilateral F de igualdade entre duas variâncias

	Tamanho Amostral	Estatística F	*F crítico* ($\alpha = 0,05$)
Caso 1 (Série A *versus* Série B)	Série A: 18 Série B: 6	$F(X_1) = 0,6913$ $F(X_2) = 2,4583$ $F(X_3) = 1,4015$	$F_{17,5}(0,05) = 0,3559$ $F_{17,5}(0,95) = 4,5904$
Caso 2 (vitória 3 pontos *versus* vitória 2 pontos)	3 pontos: 18 2 pontos: 6	$F(X_1) = 1,2655$ $F(X_2) = 0,7562$ $F(X_3) = 0,8175$	$F_{17,5}(0,05) = 0,3559$ $F_{17,5}(0,95) = 4,5904$
Caso 3 (turno com *playoffs versus* turno e returno)	Turno e *playoffs*: 19 Turno e returno: 5	$F(X_1) = 0,8745$ $F(X_2) = 2,9854$ $F(X_3) = 1,1918$	$F_{18,4}(0,05) = 0,3416$ $F_{18,4}(0,95) = 5,8211$

Assim como feito anteriormente, a notação $F(X_1)$, $F(X_2)$ e $F(X_3)$ se refere, respectivamente, à distribuição de dados do aproveitamento do primeiro colocado, do quarto colocado e do quarto último colocado dos campeonatos brasileiros. Como a região de aceitação da hipótese H_0 está entre os F críticos, todas as estatísticas F estão dentro da região de aceitação, não se podendo afirmar que as variâncias possuem diferenças significativas. Apresenta-se na Tabela 4 o teste t para a igualdade entre duas médias assumindo como iguais as variâncias populacionais.

Hipóteses testadas:

H_0 = As médias são iguais.

H_1 = As médias não são iguais.

Tabela 4 - Teste t para a igualdade entre duas médias

	Tamanho Amostral	Estatística t	*p-value*
Caso 1 (Série A *versus* Série B)	Série A: 18 Série B: 6	$t(X_1) = -0,1896$ $t(X_2) = -0,1758$ $t(X_3) = -1,9046$	0,8514 0,8620 0,0700
Caso 2 (vitória 3 pontos *versus* vitória 2 pontos)	3 pontos: 18 2 pontos: 6	$t(X_1) = 2,0761$ $t(X_2) = 1,1565$ $t(X_3) = 2,1237$	0,0498 0,2599 0,0452
Caso 3 (turno com *playoffs versus* turno e returno)	Turno e *playoffs*: 19 Turno e returno: 5	$t(X_1) = 0,2095$ $t(X_2) = 1,1302$ $t(X_3) = -2,2951$	0,8360 0,2706 0,0316

A notação $t(X_1)$, $t(X_2)$ e $t(X_3)$ se refere, respectivamente, à estatística t do aproveitamento do primeiro colocado, do quarto colocado e do quarto último colocado dos campeonatos brasileiros. Ao nível de significância de 0,0500, a hipótese de médias iguais terá que ser rejeitada no caso 2 para o aproveitamento do primeiro colocado (X_1) e do quarto último colocado (X_3). E também devem ser considerados significativamente diferentes os campeonatos que sejam disputados no sistema de turno com fase posterior dos disputados em sistema de pontos corridos, mas isso somente para o caso do quarto último colocado (X_3).

Uma explicação para a rejeição da hipótese H_0 no que se refere à variável X_3 é a sua relação com o rebaixamento. Na maioria dos campeonatos os quatro últimos colocados foram rebaixados, mas houve exceções e, até, campeonatos em que não estava previsto o rebaixamento. Tal situação ocorreu nos anos de 1971, 1972, 1992 e 2000 e pode ter influenciado, ao longo do campeonato, o rendimento dos times localizados nas últimas colocações da tabela.

Independente disso, para os cálculos posteriores, foram desconsiderados na análise da variável X_3 os campeonatos disputados no sistema de vitória valendo 2 pontos e os campeonatos que não obedeciam a fórmula atual de turno e returno. Para a variável X_1 foram desconsiderados os campeonatos disputados no sistema de vitória valendo 2 pontos.

4 Resultados

Com os critérios necessários verificados e a exclusão de observações que não se adequavam às premissas do modelo, foi feita a Tabela 5, que atualiza os dados da Tabela 1.

Tabela 5 - Observações utilizadas no modelo

	Aproveitamento do campeão (X_1)	Aproveitamento do 4º colocado (X_2)	Aproveitamento do 4º último colocado (X_3)
Série A – 1971		56,140%	
Série A – 1972		55,556%	
Série A – 1988		57,246%	
Série A – 1990		53,623%	
Série A – 1991		54,667%	
Série A – 1992		60,494%	
Série A – 1995	62,500%	56,944%	
Série A – 1996	69,841%	55,556%	
Série A – 1997	66,667%	59,420%	
Série A – 1998	72,000%	60,000%	
Série A – 1999	63,768%	56,522%	
Série A – 2000	66,667%	57,971%	
Série A – 2001	63,158%	52,632%	
Série A – 2002	64,912%	56,140%	
Série A – 2003	59,649%	52,632%	33,333%
Série A – 2004	68,116%	57,971%	31,884%
Série A – 2005	66,667%	60,000%	28,000%
Série A – 2006	63,158%	52,632%	35,088%
Série B – 1999	62,281%	53,509%	
Série B – 2002	65,079%	55,556%	
Série B – 2003	66,667%	60,870%	
Série B – 2004	68,116%	53,623%	
Série B – 2005	68,000%	62,667%	
Série B – 2006	71,429%	53,968%	38,095%
Média	**66,0375%**	**56,5141%**	**33,2800%**
Desvio Padrão	**3,2017%**	**2,9052%**	**3,3551%**

Fonte: Rec.Sport.Soccer Statistics Foundation (RSSSF) e autor.

Como qualquer subconjunto de uma distribuição normal é também normalmente distribuído (James, 2006), os dados da Tabela 5 permitem construir um modelo das variáveis estudadas, que obedecem a distribuições normais univariadas com parâmetros $X_1 \sim N(0{,}6604 ; 0{,}0320^2)$, $X_2 \sim N(0{,}5651 ; 0{,}0291^2)$ e $X_3 \sim N(0{,}3328 ; 0{,}0336^2)$. Lembra-se que a função densidade de probabilidade da curva normal é dada pela Equação 1.

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (1)$$

A probabilidade da variável X assumir um valor entre a e b (a < b) é dada pela Equação 2.

$$P(a < X < b) = \int_a^b f(x)\,dx \qquad (2)$$

Através das equações anteriores e com o auxílio computacional do *software* MATLAB foi calculado o limite superior **b** para cada probabilidade **P (a < X < b)**, esta está apresentada na primeira coluna da Tabela 6. O limite inferior **a** foi selecionado como zero e o limite superior **b** tem seu valor apresentado na coluna aproveitamento de cada variável.

Tabela 6 - Probabilidades, obtidas a partir da função densidade de probabilidade de cada variável, para o aproveitamento e pontuação necessária para se alcançar determinada colocação em um campeonato disputado por 20 times em turno e returno

Probabilidade	1º colocado (X_1)		4º colocado (X_2)		4º último colocado (X_3)	
50%	66,04%	75	56,51%	64	33,28%	38
55%	66,44%	76	56,88%	65	33,70%	38
60%	66,85%	76	57,25%	65	34,13%	39
65%	67,27%	77	57,63%	66	34,57%	39
70%	67,72%	77	58,04%	66	35,04%	40
75%	68,20%	78	58,47%	67	35,54%	41
80%	68,73%	78	58,96%	67	36,10%	41
85%	69,36%	79	59,53%	68	36,76%	42
90%	70,14%	80	60,24%	69	37,58%	43
95%	71,30%	81	61,29%	70	38,80%	44
97,5%	72,31%	82	62,21%	71	39,86%	45
99,0%	73,49%	84	63,27%	72	41,09%	47
99,9%	75,93%	87	65,49%	75	43,65%	50

Assim, pode-se estabelecer a probabilidade de um time ser campeão, classificar-se para competições internacionais ou ser rebaixado de divisão de acordo com o aproveitamento necessário. A pontuação exemplificada se refere ao sistema de disputa do campeonato de 2007, com 20 times se enfrentando em sistema de turno e returno e vitória valendo três pontos, porém cabe salientar que o índice de aproveitamento apresentado independe, para a variável X_2, da fórmula de disputa do campeonato, sendo esse um fator muito positivo do modelo.

Com o auxílio da Tabela 6, se um time almeja ser campeão da competição é coerente, com chance de 90% de acerto, ele colocar como meta um aproveitamento de 70,14%, ou 80 pontos. Se o objetivo é subir da Série B para a Série A é preciso, dentro da mesma probabilidade anterior, atingir um aproveitamento de 60,24% ou 69 pontos. A

terceira variável guarda uma sutileza: para se escapar do rebaixamento é preciso estar na frente do quarto último colocado pelo menos nos critérios de desempate, mais seguro é acrescentar um ponto à pontuação do quarto último colocado, assim um clube que deseja permanecer em sua divisão deve almejar uma pontuação de 44 pontos dentro de um nível de confiança de 90%. Obviamente que quanto maior o nível de confiança com o qual se trabalha maior é a segurança do time de conseguir seu objetivo ao alcançar a pontuação fornecida pelo modelo. Assim, um clube que alcançar, por exemplo, 80 pontos será campeão em 90% dos campeonatos disputados. É lógico que também é possível ser campeão com 75 pontos, mas as chances de isso ocorrer são bem menores, de 50%.

Para proceder a uma validação do modelo podem-se usar dados futuros, quando estiverem disponíveis, ou usar uma amostra de outros campeonatos similares ao redor do mundo, testando a hipótese de não haver diferença significativa entre esses campeonatos e o campeonato brasileiro, para então verificar a consistência das predições dadas pela Tabela 6.

5 Validação dos resultados

Outros campeonatos nacionais disputados em termos parecidos ao brasileiro são o francês, espanhol, inglês, italiano e alemão, em suas divisões principais. A maioria dos grandes campeonatos nacionais europeus possui a mesma fórmula de disputa há mais de cinco décadas, sendo excelentes objetos de pesquisa futura. Para a construção dessa amostra foram sorteados, através da geração de números randômicos do *software* MATLAB, dez campeonatos, dentre os últimos 12 anos entre os cinco países citados. Vale lembrar que somente nos últimos 12 anos o sistema de pontuação do futebol foi modificado, com a vitória valendo 3 pontos, fato que permanece até hoje. E assim montou-se a Tabela 7 a seguir:

Tabela 7 - Amostra de Campeonatos Europeus

Campeonato – Ano de Início	Aproveitamento do 1º colocado (X_1)	Aproveitamento do 4º colocado (X_2)	Aproveitamento do 4º último colocado (X_3)
Italiano – 2001	61,404%	48,246%	34,211%
Inglês – 1998	69,298%	58,772%	35,965%
Francês – 2003	69,298%	57,018%	34,211%
Alemão – 2006	68,627%	58,824%	36,275%
Inglês – 2000	70,175%	59,649%	36,842%
Francês – 1999	63,725%	52,941%	41,176%
Alemão – 2004	75,490%	56,863%	35,294%
Espanhol – 1995	69,048%	58,730%	34,921%
Italiano – 2005	66,667%	47,368%	27,193%
Espanhol – 2002	68,421%	53,509%	37,719%
Média Amostral	**68,215%**	**55,192%**	**35,381%**
Desvio Padrão Amostral	**3,786%**	**4,495%**	**3,533%**

Fonte: Rec.Sport.Soccer Statistics Foundation (RSSSF)

Foi aplicado o teste t para a média populacional μ com variância desconhecida (Marques e Marques, 2005) para detectar se há diferença significativa entre as médias da amostra sorteada e a média da distribuição normal definida para o campeonato brasileiro (Tabela 8):

Hipóteses testadas:

H_0 = A média μ é igual a μ_0 (X_i)

H_1 = A média μ não é igual a μ_0 (X_i)

Tabela 8 - Teste t para a média populacional com variância desconhecida

	Valor de μ_0	Estatística t	*p-value*
Aproveitamento do 1º colocado (X_1)	μ_0 (X_1) = 66,0375%	1,8193	0,0761
Aproveitamento do 4º colocado (X_2)	μ_0 (X_2) = 56,5141%	-0,9301	0,2133
Aproveitamento do 4º último colocado (X_3)	μ_0 (X_3) = 33,2800%	1,8802	0,0714

Ao nível de significância de 5%, a hipótese H_0 não pode ser rejeitada em nenhum caso, portanto assume-se como não existindo diferença significativa entre o campeonato brasileiro e a amostra de campeonatos europeus.

Cabe, então, comparar as predições teóricas feitas pelo modelo, presentes na Tabela 9, com os resultados reais da amostra dos campeonatos europeus. A comparação será feita ao nível de confiança de 95%.

Tabela 9 - Comparação entre o aproveitamento teórico necessário e os valores amostrais

	Aproveitamento do 1º colocado (X_1)	Aproveitamento do 4º colocado (X_2)	Aproveitamento do 4º último colocado (X_3)
Teórico (Tabela 6) b = 0.95	**71,30%**	**61,29%**	**38,80%**
Italiano – 2001	61,404%	48,246%	34,211%
Inglês – 1998	69,298%	58,772%	35,965%
Francês – 2003	69,298%	57,018%	34,211%
Alemão – 2006	68,627%	58,824%	36,275%
Inglês – 2000	70,175%	59,649%	36,842%
Francês – 1999	63,725%	52,941%	41,176%
Alemão – 2004	75,490%	56,863%	35,294%
Espanhol – 1995	69,048%	58,730%	34,921%
Italiano – 2005	66,667%	47,368%	27,193%
Espanhol – 2002	68,421%	53,509%	37,719%
Resultados acima do valor teórico	**1**	**0**	**1**

Apenas dois resultados não estão dentro do intervalo fornecido pelo modelo teórico, o campeonato Alemão de 2004 no que se refere ao seu campeão e o campeonato Francês de 1999 em relação ao 4º último colocado. Em 30 resultados testados, 2 escaparam do intervalo teórico, uma margem de erro de 6,67%, enquanto o esperado era de 5%. O pequeno tamanho da amostra justifica essa diferença, porém estudos mais aprofundados devem ser feitos antes de se fazer a extensão do modelo para qualquer campeonato. Ainda assim considera-se o resultado satisfatório, visto que o modelo não foi elaborado para fornecer probabilidades sobre os campeonatos europeus, mas sim sobre os campeonatos brasileiros.

Intuitivamente o campeonato brasileiro se mostra mais competitivo do que os europeus, nos quais poucos times disputam o primeiro lugar mesmo com o passar de décadas. A competição brasileira também é caracterizada pela grande troca de treinadores e jogadores ao final das temporadas e mesmo durante as competições, o que traz força para a premissa de independência entre campeonatos consecutivos, o mesmo não se podendo dizer sobre as competições européias. Em alguns casos, há também diferenças entre os campeonatos em relação no número de participantes e posições que provocam o rebaixamento ou a classificação para competições internacionais.

Portanto, apresenta-se um modelo extremamente simples, mas confiável, dado que o aproveitamento obedece a uma distribuição normal, para predizer a pontuação que um clube necessita para atingir seu objetivo dentro do campeonato brasileiro. O desenvolvimento do modelo também pode ser aplicável a outros campeonatos e outros esportes que envolvam a disputa de muitas partidas, mesmo que não haja a possibilidade do empate, como basquete, vôlei ou futsal.

A construção de um modelo dentro das mesmas premissas para outros campeonatos ao redor do mundo, como o francês, espanhol, inglês, italiano e alemão também é possível. Confirmando-se a semelhança entre esses campeonatos pode-se, inclusive, formular um modelo ainda mais robusto e confiável. Podendo se prosseguir com outros testes de igualdade de médias, como a ANOVA, inclusive para observar se há diferença entre diversos campeonatos de paises diferentes, entre campeonatos com número diversos de clubes rebaixados etc, de forma a refinar o modelo.

A análise multivariada dos dados também é de interesse, uma vez que não há uma independência exata entre as variáveis, fato exemplificado por um campeonato em que há um índice grande de empates, o que faz diminuir o aproveitamento X_1 e X_2 e aumentar o X_3. O que não é um problema neste presente trabalho, dado que essa é uma análise anterior ao início do campeonato, para servir como parâmetro de pontuação necessária para os clubes atingirem seus objetivos, mas que não cabe, dentro da mesma linha de raciocínio, para um campeonato já em andamento.

Conclusões

O foco principal desse artigo foi formular um modelo simples, eficiente e de fácil utilização para predizer a pontuação necessária que um clube deve ter para alcançar seus objetivos. O resultado foi alcançado com a conclusão de que o índice de aproveitamento do primeiro colocado (variável X_1), do quarto colocado (X_2) e do quarto último colocado (X_3) dos campeonatos brasileiros obedece a uma distribuição normal com parâmetros $X_1 \sim N(0{,}6604 ; 0{,}0320^2)$, $X_2 \sim N(0{,}5651 ; 0{,}0291^2)$ e $X_3 \sim N(0{,}3328 ; 0{,}0336^2)$ e assim, a

partir de conceitos simples de probabilidade, é possível calcular o aproveitamento e a pontuação desejada dentro de um intervalo de confiança. O modelo apresentou resultados satisfatórios ao predizer a pontuação necessária para uma amostra de 10 campeonatos europeus, mesmo não tendo sua formulação pensada para tais campeonatos.

Outras conclusões foram possíveis no que se refere a três casos: 1) Os campeonatos brasileiros da Série A e B podem ser aceitos como iguais no que diz respeito às três variáveis estudadas dentro de um nível de significância de 5%. 2) Já o sistema de pontuação das vitórias influencia no aproveitamento do 1º colocado e do 4º último colocado de um campeonato, mas não se mostrou significativo para o caso 4º colocado geral. 3) Por fim o sistema de disputa do campeonato em pontos corridos ou em turno único com fase posterior não modifica o aproveitamento do 1º e do 4º colocado geral, mas provoca diferenças significativas para o caso do rebaixamento. Possíveis explicações para esse fato foram levantadas, mas necessitam de posteriores pesquisas para serem conclusivas.

Na comparação com outros modelos, essa proposta se mostra mais simples e flexível, tanto em suas premissas como em suas aplicações, podendo ter seu uso estendido para outros campeonatos ao redor do mundo e outros esportes com sistema de disputa similar.

ARTUSO, A. R. Soccer team's performance analysis in Brazilian championship from a normal distribution. *Rev. Bras. Biom.*, São Paulo, v.25, n.4, p.49-63, 2007.

- *ABSTRACT: Sports had always fascinated humanity, in this context, soccer was taken as a study source. The objective of this paper is formulate a model to estimate necessary scores to get some position at the final ranking of the Brazilian National Soccer Championship, Division A and Division B. The data from old championship was used to prove that the performance's team obeys a normal (Gaussian) distribution of probabilities and can be used as a parameter to define objectives form each team before beginning the competitions. The model is also valid, with some limitations for Brazilian championships that was disputed with different rules or different point systems and it appears efficient when tested in a sample with ten European's football championship. Applications of the presented reasoning are possible in other championships around the world and also in other sports with similar dispute system.*

- *KEYWORDS: Soccer; football; performance; normal distribution.*

Referências

ARRUDA, M. L. *Poisson, Bayes, Futebol e DeFinetti*. 2000. 123f. Dissertação (Mestrado em Estatística) – Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2000.

EMONET, B. *Revisiting statistical applications in soccer*. Lausanne: Swiss Federal Institute of Technology, 2000.

KUNS, M. *265 million playing football*. FIFA Magazine, Zurique, p.10-15, jul. 2007.

GOUVEIA, F. Os números da paixão. *Ver. Eletrôn. Jornal. Cient.*, Sociedade Brasileira para o progresso da Ciência, n.79, 2006. Disponível em: <http://www.comciencia.br>. Acesso em: 30 ago. 2007.

JAMES, B. R. *Probabilidade:* um curso em nível intermediário. 3rd. ed. Rio de Janeiro: IMPA, 2006.

MARQUES, J. M.; MARQUES, M. A. M. *Estatística básica para os cursos de engenharia*. Curitiba: Domínio do Saber, 2005.

The MathWorks Inc. *MATLAB*. versão.7 R14, 2004.

MOOD, A. M.; GRAYBILL, F. A.; BOES, D. C. *Introduction to the theory of statistics***.** 3rd. ed. New York: McGraw Hill, 1974.

RSSSF – Rec. Sport. Soccer Statistics Foundation. *Historical Domestic Results*. Disponível em: <http://www.rsssf.com/>. Acesso em: 30 ago. 2007.

RIBEIRO, C. C.; URRUTIA, S. An application of integer programming to playoff elimination in football championships. *Int. Trans. Oper. Res.*, Oxford, v.12, n 4, p.375-386, 2005.

SIEGEL, S.; CASTELLAN, N. J. *Estatística não-paramétrica para ciências do comportamento*. 2nd. ed. Porto Alegre: Artmed, 2006.

SZYMANSKI, S. Economics of sport: introduction. *Econ. J.***,** Oxford, v.111, n.469, p.1-3, 2001.

SILVA, C. F.; SALIBY, E. S.; SALIBY, E. Soccer championship analysis using Monte Carlo simulation. In: WINTER SIMULATION CONFERENCE, 2002, San Diego. *Proceedings...,* v. 1, p 2011-2016.

Recebido em 05.09.2007.

Aprovado após revisão em 18.02.2008.