DISTRIBUIÇÃO GAUSSIANA DOS RESULTADOS DO CAMPEONATO BRASILEIRO DE FUTEBOL:

UM MODELO PARA ESTIMAR CLASSIFICAÇÕES EM CAMPEONATOS DE MODALIDADES COLETIVAS

M.Sc. ALYSSON RAMOS ARTUSO
Departamento de Engenharia, Centro Universitário Franciscano (Unifae)
E-mail: alysson.artuso@gmail.com

RESUMO

O objetivo deste artigo é formular um modelo para estimar a pontuação necessária para 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 gaussiana de probabilidades e, assim, pode ser usado como parâmetro para os clubes definirem suas metas, dentro de um nível de confiança, 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 mostrou-se eficiente ao ser testado em uma amostra de campeonatos europeus.

PALAVRAS-CHAVE: Futebol; aproveitamento; distribuição gaussiana.

INTRODUÇÃO

A maior e mais importante competição esportiva nacional é o Campeonato Brasileiro de Futebol organizado pela Confederação Brasileira de Futebol (CBF) e dividido em série A, série B e série C. As duas primeiras são, atualmente, disputadas por 20 times e possuem a mesma fórmula de disputa.

Apesar de sua enorme popularidade e de sua importância econômica, o futebol brasileiro tem recorrentes problemas, como dificuldades financeiras e administrativas, além da falta de organização, seja dos clubes, seja das instituições responsáveis pelo gerenciamento do futebol, que acaba submetendo os times a um calendário em que freqüentemente se disputam dois, três ou mais competições numa mesma temporada, por vezes simultaneamente. Como conseqüência pode haver um desgaste acima do desejável dos atletas, ocasionando um alto número de lesões; por isso alguns clubes usualmente priorizam algumas competições em detrimento de outras. Tome-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 mídia esportiva volta seu interesse para esses números de modo que informe os fãs de futebol sobre o desempenho e as chances de seus times (Vendite; Vendite; Moraes, 2005).

Tais estudos são relativamente comuns e difundidos, principalmente nas ligas profissionais americanas, e em especial no caso do beisebol. 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 deste artigo é justamente apresentar uma forma alternativa de fornecer, dentro de um intervalo de confiança, a pontuação necessária para tornar-se campeão, classificar-se para competições continentais e divisões mais altas ou ainda para escapar do rebaixamento a divisões inferiores. Para isso, utiliza-se de resultados passados e conhecimentos estatísticos na proposição do modelo, contribuindo dessa forma para preencher uma lacuna em modelos de predição aplicados ao futebol, em especial, um modelo voltado aos parâmetros do futebol brasileiro.

OBJETIVOS, CARACTERÍSTICAS DO CAMPEONATO BRASILEIRO E PROPOSIÇÃO DO MODELO

Provando-se que o aproveitamento passado em diversos campeonatos brasileiros obedece a uma distribuição gaussiana univariada pode-se, a partir de conceitos de probabilidade e estatística, estimar valores de aproveitamento necessário para os clubes planejarem-se no início da competição.

Antes de 1994, o Campeonato Brasileiro de Futebol atribuía dois pontos para o time vitorioso de uma partida, a partir de então houve uma mudança no sistema de pontuação em caso de vitória para 3 pontos. Para os cálculos deste artigo será considerado 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.

A competição brasileira não segue as mesmas regras ano após ano, sendo freqüente a mudança de regulamento de uma temporada para outra. As fórmulas mais usadas foram a de campeonatos disputados em fase única de turno e returno (série A após 2003 e série B após 2006) e campeonatos disputados em turno único com fase posterior no sistema de *play-offs* (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 por meio de medidas regulatórias da CBF.

Atualmente, a série A e a série B são disputadas cada uma 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 no caso da série A, ou sobem de divisão no caso da série B, e os quatro últimos são rebaixados para a divisão inferior. No total são disputadas 380 partidas em cada divisão.

Diferentes modelos para o Campeonato Brasileiro foram propostos por alguns autores. Destacam-se três classes de propostas:

1) *Cálculo probabilístico de resultados*. Pode-se argumentar que o número de gols marcados por um time em uma dada partida obedece a uma distribuição de Poisson, assim essa distribuição é utilizada 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 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. 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 programação linear. Nesse caso, são usados algoritmos que simulam todos os resultados possíveis de todos os jogos restantes do campeonato. Em meados dos anos de 1960, publicações abordavam o problema de saber quando um time está matematicamente eliminado na liga de beisebol americana (Major League Baseball – 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 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 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). 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 alcançar determinadas posições. Nesse 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 assumidas, como a igualdade entre todos os times, a independência dos resultados entre os jogos e a probabilidade sempre igual de um jogo terminar empatado. Dessa forma foi feito um levantamento histórico dos campeonatos de 1996 até 2001 da porcentagem de jogos que acabaram empatados, sendo esse parâmetro estimado por meio de uma distribuição triangular. Ao final são dadas as pontuações necessárias, respeitado um nível de confiabilidade, para 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 atingir determinada colocação perde o sentido quando a competição é composta por até seis clubes a menos e com um número muito maior de jogos, visto que eles são atualmente disputados em dois turnos.

Com a intenção de 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, em vez de trabalhar com a pontuação final, realizar uma análise dos

aproveitamentos, calculados pela porcentagem de pontos conquistados em relação ao total de pontos disputados, necessários para atingir determinada colocação, tratando-os como variáveis independentes que obedecem a uma distribuição gaussiana $N \sim (\mu,\sigma 2)$ 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.

Em Artuso (2007) um modelo similar foi proposto, mas sem rigor na estimação dos parâmetros e com uma base de dados desatualizada, o que ocasionou um resultado não tão eficiente na validação dos resultados. Isso acarretou também imprecisões nas distribuições gaussianas que foram corrigidas e com seus resultados ampliados pelo presente trabalho.

CONSTRUÇÃO DO MODELO

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 interessa somente os campeonatos disputados no sistema de pontos corridos, 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. Assim formou-se a amostra de 26 campeonatos que atendem aos critérios estabelecidos, 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 adéqua ao índice de aproveitamento do primeiro colocado (chamada de X_1), do quarto colocado (X_2) e do quinto último colocado (X_3) dos Campeonatos Brasileiros. Tais posições foram escolhidas por representar, respectivamente, a posição de campeão da competição, o último clube a classificar-se para a Copa Libertadores da América ou para a série A e o último clube a não ser rebaixado para uma divisão inferior.

Já é de conhecimento de estudiosos da área que os pontos feitos por um time num campeonato de futebol obedecem a uma distribuição gaussiana univariada (Emonet, 2000). Como o aproveitamento é uma combinação linear da pontuação, este também obedece a uma distribuição gaussiana em virtude das propriedades da própria distribuição (James, 2006). 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 aceitar ou rejeitar a hipótese de gaussianidade dos dados (Siegel; Castellan, 2006).

Hipóteses testadas:

H_0 = A distribuição de Xi é igual à distribuição gaussiana.

H_1 = A distribuição de Xi não é igual à distribuição gaussiana.

Tabela 1 – Dados da população observada

	Aproveitamento do campeão (X_1)	Aproveitamento do 4º colocado (X_2)	Aproveitamento do 5º último colocado (X_3)
Série A 07	67,5439%	53,5088%	39,4737%
Série A 06	68,4211%	56,1404%	38,5965%
Série A 05	64,2857%	55,5556%	40,4762%
Série A 04	64,4928%	57,2464%	36,9565%
Série A 03	72,4638%	53,6232%	36,2319%
Série A 02	69,3333%	54,6667%	37,3333%
Série A 01	72,8395%	60,4938%	35,8025%
Série A 00	62,5000%	56,9444%	36,1111%
Série A 99	69,8413%	55,5556%	34,9206%
Série A 98	66,6667%	59,4203%	34,7826%
Série A 97	72,0000%	60,0000%	34,6667%
Série A 96	63,7681%	56,5217%	39,1304%
Série A 95	66,6667%	57,9710%	34,7826%
Série A 92	63,1579%	52,6316%	33,3333%
Série A 91	64,9123%	56,1404%	36,8421%
Série A 90	59,6491%	52,6316%	33,3333%
Série A 88	68,1159%	57,9710%	37,6812%
Série A 72	66,6667%	60,0000%	28,0000%
Série A 71	63,1579%	52,6316%	36,8421%
Série B 07	60,5263%	51,7544%	43,8596%
Série B 06	62,2807%	53,5088%	38,5965%
Série B 05	65,0794%	55,5556%	39,6825%
Série B 04	66,6667%	60,8696%	36,2319%
Série B 03	68,1159%	53,6232%	34,7826%
Série B 02	68,0000%	62,6667%	40,0000%
Série B 99	71,4286%	53,9683%	38,0952%
Média	66,4839%	56,2153%	36,7902%
Desvio Padrão	3,5863%	2,9909%	3,0172%

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

Tabela 2 – Testes de gaussianidade

	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistic	df	*p-value*	Statistic	df	*p-value*
Aproveitamento do campeão (X_1)	0,0973	26	0,9574	0,9755	26	0,7669
Aproveitamento do 4º colocado (X_2)	0,1199	26	0,8247	0,9495	26	0,2260
Aproveitamento do 5º último colocado (X_3)	0,1254	26	0,7809	0,9345	26	0,2282

Tendo como base a tabela 2 não se pode rejeitar a hipótese nula de gaussianidade para nenhuma das 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 das séries A e B (caso 1), se há diferença significativa em função do sistema de pontuação (caso 2) e se há diferença significativa por conta da fórmula de disputa do campeonato (caso 3). Como subconjuntos de uma distribuição gaussiana também apresentam distribuição gaussiana (James, 2006), testes de gaussianidade podem ser dispensados.

O teste de hipótese indicado para a igualdade de médias é o teste t de *student* no caso de dados vindos de uma amostra (Marques; Marques, 2005). O teste, tal como será aplicado, exige como pré-requisitos a gaussianidade, 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 pelo teste F de igualdade entre duas variâncias (Mood; Graybill; Boes, 1974). Dessa forma segue-se 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.

Ao nível de significância de 5%, a hipótese de médias iguais terá que ser rejeitada nos três casos para o aproveitamento do quinto último colocado (X_3). Ou seja, devem considerados significativamente diferentes para X_3 os campeonatos da série A e série B, os campeonatos com sistema de pontuação diferente e os campeonatos com fórmulas de disputa diferentes. A estatística F mostra que em

Tabela 3 – Teste t para a igualdade entre duas médias

	Tamanho Amostral	Estatística t	*p-value*	Estatística F	*p-value*
Caso 1 (série A *versus* série B)	Série A: 19 Série B: 7	$t(X_1) = 0,3987$ $t(X_2) = 0,2264$ $t(X_3) = -0,1510$	0,6937 0,8228 0,0418	$F(X_1) = 1,0537$ $F(X_2) = 2,5421$ $F(X_3) = 1,0965$	0,8497 0,1165 0,8037
Caso 2 (vitória 3 pontos *versus* vitória 2 pontos)	3 pontos: 20 2 pontos: 6	$t(X1) = 1,7935$ $t(X_2) = 1,3699$ $t(X_3) = 2,4952$	0,0855 0,1834 0,0199	$F(X_1) = 2,1816$ $F(X_2) = 1,1718$ $F(X_3) = 1,4034$	0,1986 0,7182 0,7569
Caso 3 (turno com *play-offs versus* turno e returno)	Turno e *play-offs*: 19 Turno e returno: 7	$t(X_1) = 0,6548$ $t(X_2) = 1,8893$ $t(X_3) = -2,7409$	0,5188 0,0710 0,0114	$F(X_1) = 1,3586$ $F(X_2) = 2,6689$ $F(X_3) = 1,1878$	0,5671 0,2289 0,8918

nenhum caso a igualdade das variâncias pode ser rejeitado, cumprindo a premissa de homocedasticidade necessária ao teste t.

A explicação para a rejeição da hipótese H_0 para variável X_3 pode ser dada pela sua relação com o rebaixamento. Na maioria dos campeonatos os quatro últimos colocados foram rebaixados, mas houve exceções freqüentes com dois ou seis rebaixados e, até, campeonatos em que não estava previsto o descenso. 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, impossibilitando o uso de campeonatos diferentes para a estimação de parâmetros no que se refere à X_3.

Por isso, para os cálculos posteriores, foram desconsiderados na análise da variável X_3 todos campeonatos da série A anteriores a 2003 e quaisquer campeonatos da série B. Para as demais variáveis a hipótese nula mostrou-se válida e, portanto, serão usados todos os dados disponíveis.

Isso causa um problema adicional na estimação dos parâmetros. O método de estimação utilizado, por ponto, não garante um percentual de validade e uma extensão simples para a população toda, assim uma estimação por intervalo para a média populacional faz-se necessária.

RESULTADOS

Após se verificar os critérios necessários para a formulação do modelo e a exclusão de observações em relação a X_3 que não se adequavam às premissas apresentadas, chega-se a um modelo das variáveis estudadas que obedecem a distribuições normais univariadas com parâmetros pontualmente estimados $X_1 \sim N(0,6648 ; 0,0359^2)$, $X_2 \sim N(0,5622 ; 0,0299^2)$ e $X_3 \sim N(0,3835 ; 0,0175^2)$.

A estimação por ponto de um parâmetro não possui uma medida do possível erro cometido na estimação, sendo conveniente conhecer a precisão dessa estimação. Então, o problema é o de determinar os limites superior e inferior, entre os quais esteja compreendido o verdadeiro valor do parâmetro, ou seja, fazer uma estimação do intervalo. No caso da estimação intervalar da média é usada a distribuição t de *student* para o cálculo, para o caso da estimação intervalar do desvio-padrão a distribuição usada é a qui-quadrado (MARQUES; MARQUES, 2005). O nível de significância utilizado foi de 5%.

Assim, para o caso das 26 observações das variáveis X_1 e X_2 o intervalo no qual o valor da média populacional está contido é bastante pequeno, sendo a correção que a média poderia sofrer é pequena, de 0,85% e 0,84%, respectivamente. Já para a variável X_3 esse intervalo está situando entre 0,3640 e 0,4029, uma variação de 5,07% para cima ou para baixo. Como o interesse é planejar a pontuação do campeonato com segurança convém superestimar o parâmetro μ, trabalhando com a média das variáveis no limite superior do intervalo.

A estimação por intervalo do desvio-padrão também é necessária e seu cálculo resulta nos intervalos de [0,0346; 0,0363], [0,0291; 0,0302] e [0,0162; 0,0176] para X_1, X_2 e X_3, respectivamente. Novamente utilizando os limites superiores, pode-se descrever as variáveis com as distribuições normais univariadas $X_1 \sim N(0{,}6705; 0{,}0363^2)$, $X_2 \sim N(0{,}5669; 0{,}0302^2)$ e $X_3 \sim N(0{,}4029; 0{,}0176^2)$.

Dessa forma, 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 refere-se ao sistema de disputa do campeonato de 2007, com 20 times enfrentando-se 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_1 e X_2, da fórmula de disputa do campeonato, sendo esse um fator muito positivo do modelo. Os resultados preditos para a variável X_3 são válidos somente para campeonatos disputados no sistema de pontos corridos da série A com vitória valendo 3 pontos. A pontuação refere-se a um campeonato disputado em turno e returno por 20 participantes. Para outras fórmulas de disputa basta utilizar o aproveitamento para o cálculo da pontuação: *pontuação* = *aproveitamento* x *número de jogos* x *pontuação da vitória*.

Com o auxílio da tabela 4, se um time almeja ser campeão da competição é coerente, com chance de 90% de acerto, ele colocar como meta um aproveitamento de 71,70%, ou 82 pontos. Se o objetivo é subir da série B para a série A é preciso, dentro da mesma probabilidade anterior, atingir um aproveitamento de 60,56% ou 69 pontos.

Tabela 4 – Probabilidades, obtidas a partir da função densidade de probabilidade (f.d.p.) de cada variável

Probabilidade	1º colocado (X_1)		4º colocado (X_2)		5º último colocado (X_3)	
	Aprov.	Pontos	Aprov.	Pontos	Aprov.	Pontos
50%	67,05%	76	56,69%	65	40,29%	46
55%	67,51%	77	57,07%	65	40,51%	46
60%	67,97%	77	57,46%	66	40,74%	46
65%	68,45%	78	57,85%	66	40,97%	47
70%	68,95%	79	58,27%	66	41,21%	47
75%	69,50%	79	58,73%	67	41,48%	47
80%	70,11%	80	59,23%	68	41,77%	48
85%	70,81%	81	59,82%	68	42,11%	48
90%	71,70%	82	60,56%	69	42,55%	49
95%	73,02%	83	61,66%	70	43,18%	49
97,5%	74,16%	85	62,61%	71	43,74%	50
99,0%	75,49%	86	63,72%	73	44,38%	51
99,9%	78,27%	89	66,02%	75	45,73%	52

Como foi usada a população dos próprios campeonatos passados, a distribuição gaussiana garante que os resultados antigos estarão dentro de suas predições, sendo ilógico proceder a uma validação do modelo utilizando-se os dados passados ou o resultado de simulações feitas com parâmetros estimados a partir desses dados. 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 4.

VALIDAÇÃO E DISCUSSÃO DOS RESULTADOS

Outros campeonatos disputados em termos parecidos ao brasileiro são o francês, o espanhol, o inglês, o italiano e o alemão, em suas divisões principais. A maioria dos campeonatos europeus possui a mesma fórmula de disputa há mais de cinco décadas, sendo excelentes objetos de pesquisas futuras. Para a construção dessa amostra foram sorteados, pela geração de números randômicos do *software* Matlab, 15 campeonatos, desde 1994, quando o sistema de pontuação do futebol foi modificado, entre os cinco países citados. E assim montou-se a tabela 5.

Foi aplicado o teste t para a diferença entre duas médias μ com variância desconhecida (Marques; Marques, 2005) para detectar se há diferença significativa entre as médias da amostra sorteada e a média da distribuição gaussiana definida para o Campeonato Brasileiro:

Hipóteses testadas:
H_0 = As médias são iguais.
H_1 = As médias não são iguais.

Ao nível de significância de 5%, o pressuposto de homocedasticidade é satisfeito para a aplicação do teste t e portanto a hipótese H_0 não pode ser rejeitada

Tabela 5 – Amostra de Campeonatos Europeus

Campeonato – Ano de Início	Aproveitamento do 1° colocado (X_1)	Aproveitamento do 4° colocado (X_2)	Aproveitamento do 5° último colocado (X_3)
Italiano 01/02	61,40%	48,25%	39,22%
Inglês 98/99	69,30%	58,77%	36,84%
Francês 03/04	69,30%	57,02%	34,21%
Alemão 06/07	68,63%	58,82%	39,21%
Inglês 00/01	70,17%	59,65%	36,84%
Francês 99/00	63,72%	52,94%	42,16%
Alemão 04/05	75,49%	56,86%	37,25%
Espanhol 95/96	69,05%	58,73%	36,51%
Italiano 05/06	66,67%	47,37%	28,07%
Espanhol 02/03	68,42%	53,51%	38,60%
Espanhol 00/01	70,18%	55,26%	36,84%
Inglês 96/97	65,79%	59,65%	35,96%
Francês 06/07	71,05%	50,00%	37,72%
Francês 00/01	66,67%	55,88%	39,21%
Espanhol 97/98	64,91%	55,26%	39,47%
Média Amostral	68,0500%	55,1980%	37,2073%
Desvio Padrão Amostral	3,3778%	4,0450%	3,1492%

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

Tabela 6 – Teste t para a igualdade entre duas médias

	Tamanho Amostral	Estatística t	*p-value*	Estatística F	*p-value*
Aproveitamento do 1° colocado (X_1)	Brasileiro: 26 Europeu: 15	-1,37502	0,17697	1,829032	0,18221
Aproveitamento do 4° colocado (X_2)	Brasileiro: 26 Europeu: 15	0,920999	0,36271	3,232513	0,26612
Aproveitamento do 4° último colocado (X_3)	Brasileiro: 26 Europeu: 5	0,761649	0,45613	1,127295	0,83813

em nenhum caso, assumindo não existir diferença significativa entre o Campeonato Brasileiro e a amostra de campeonatos europeus.

Cabe, então, comparar as predições teóricas feitas pelo modelo, presentes na tabela 4, com os resultados reais da amostra dos campeonatos europeus. A comparação será feita ao nível de confiança de 90%.

Tabela 7 – Comparação entre o aproveitamento teórico necessário e os valores amostrais

	Aproveitamento do 1º colocado (X_1)	Aproveitamento do 4º colocado (X_2)	Aproveitamento do 5º último colocado (X_3)
Teórico (Tabela 4) nc = 0.90	71,70%	60,56%	42,55%
Italiano 01/02	61,40%	48,25%	39,22%
Inglês 98/99	69,30%	58,77%	36,84%
Francês 03/04	69,30%	57,02%	34,21%
Alemão 06/07	68,63%	58,82%	39,21%
Inglês 00/01	70,17%	59,65%	36,84%
Francês 99/00	63,72%	52,94%	42,16%
Alemão 04/05	75,49%	56,86%	37,25%
Espanhol 95/96	69,05%	58,73%	36,51%
Italiano 05/06	66,67%	47,37%	28,07%
Espanhol 02/03	68,42%	53,51%	38,60%
Espanhol 00/01	70,18%	55,26%	36,84%
Inglês 96/97	65,79%	59,65%	35,96%
Francês 06/07	71,05%	50,00%	37,72%
Francês 00/01	66,67%	55,88%	39,21%
Espanhol 97/98	64,91%	55,26%	39,47%
Resultados acima do valor teórico	1	0	0

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

Praticamente todos os campeonatos estão dentro do intervalo fornecido pelo modelo teórico, com exceção do campeonato alemão de 2004/2005 no que se refere ao seu campeão. Em 45 resultados testados apenas um escapou do intervalo teórico, uma margem de erro de 2,22%, dentro do esperado de 10%. Um resultado bastante satisfatório visto que o modelo não foi elaborado para fornecer probabilidades sobre os campeonatos europeus, mas sim sobre os campeonatos brasileiros.

Portanto, apresenta-se um modelo extremamente simples, mas confiável, dado que o aproveitamento obedece a uma distribuição gaussiana, para predizer, no

início da competição, a pontuação que um clube necessita para atingir seu objetivo dentro do Campeonato Brasileiro. Além de se mostrar uma proposta muito mais flexível e operacionável do que os outros modelos discutidos. 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 estimação dos parâmetros da distribuição gaussiana deu-se pelos estimadores da média e do desvio-padrão amplamente conhecidos e utilizados (MOOD; GRAYBILL; BOES, 1974) calculados de forma automatizada pelo *software* Matlab. O procedimento de estimação por um intervalo de confiança foi especialmente interessante para o caso da variável X_3, que possui poucas observações, oferecendo uma maior consistência aos resultados. Para os outros casos a contribuição é ínfima e poderia ser descartada em nome de um modelo mais simples e igualmente robusto.

CONCLUSÕES

Pelos resultados passados e pelos conceitos de estatística chegou-se a um modelo gaussiano para o índice de aproveitamento do primeiro colocado (variável X_1), do quarto colocado (X_2) e do quinto último colocado (X_3) dos Campeonatos Brasileiros com parâmetros $X_1 \sim N(0{,}6705; 0{,}0363^2)$, $X_2 \sim N(0{,}5669; 0{,}0302^2)$ e $X_3 \sim N(0{,}4029; 0{,}0176^2)$ e assim, a partir de conceitos simples de probabilidade, é possível calcular o aproveitamento e a pontuação desejada dentro de um intervalo de confiança. Os parâmetros foram estimados selecionando o limite superior de um intervalo de confiança com $\alpha = 0{,}05$, oferecendo uma segurança maior em relação a parâmetros estimados pontualmente. O modelo apresentou resultados satisfatórios ao predizer a pontuação necessária para uma amostra de 15 campeonatos europeus, mesmo não tendo sua formulação pensada para tais campeonatos. Vale ressaltar que a função densidade de probabilidade para X_3 é teoricamente válida apenas para competições da série A. Com mais observações dessa variável na série B, ou a inclusão de dados de outros campeonatos estatisticamente equivalentes, os parâmetros podem modificar-se e a conclusão ficar mais robusta.

Além disso, outras conclusões foram possíveis no que se refere a três casos: 1) os Campeonatos Brasileiros das séries A e B podem ser aceitos como iguais no que diz respeito ao aproveitamento do primeiro e do quarto colocado da competição, mas não para o caso do quinto último. 2) O sistema de pontuação das vitórias não influencia no aproveitamento do 1º colocado e do 4º último colocado de um campeonato, mas mostrou-se significativo para o caso 5º último colocado. 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, mais uma vez, provoca diferenças significativas para o caso do rebaixamento. Possíveis explicações para esse fato foram levantadas, como a ausência de rebaixamento em diversos campeonatos, mas necessitam de posteriores pesquisas para serem conclusivas.

Gaussian distribution of the results in Brazilian Football Championship: a model to estimate classifications in terms collective championships

ABSTRACT: The objective of this paper is to formulate a model to estimate necessary scores to achieve certain places at the final classification 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 gaussian distribution of probabilities and can be used as a parameter to define objectives form each team, with a reliable level, 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 European's football championship.
KEY WORDS: Soccer; football; performance; gaussian distribution.

Distribución gaussiana de los resultados del Campeonato Brasileño del Fútbol: un modelo para estimar las posiciones em los campeonatos de términos coletivas

RESUMÉN: El objetivo deste trabajo es formular un modelo para estimar las puntuaciones necesarias para alcanzar ciertos lugares en la posición final de la clasificación de Campeonato Brasileño del Fútbol, de las divisiones A y B. Los datos de antiguos campeonato se utilizó para demostrar que el rendimiento de la equipo obedece disctribución gaussiana de probabilidades y se puede utilizar como parámetro para definir objetivos de cada equipo, con un nivel confiable, antes de comenzar las competiciones. El modelo también es válido, con algunas limitaciones, para los campeonatos brasileños que se disputó con normas diferentes o diferentes sistemas de punto y parece eficiente cuando se analizaron en una muestra con el campeonato europeo de fútbol.
PALABRAS CLAVES: Fútbol; rendimiento; disctribucíon gaussiana.

REFERÊNCIAS

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

ARTUSO, A. R. Análise do aproveitamento dos times no campeonato brasileiro a partir de uma distribuição normal. *Revista Brasileira de Biometria*, v. 25, n. 4, p. 49-63, 2007.

EMONET, B. *Revisiting Statistical Applications in Soccer*. Lausanne: Swiss Federal Institute of Technology, 2000.

GOUVEIA, F. Os números da paixão. *Revista Eletrônica de Jornalismo Científico*, Sociedade Brasileira para o Progresso da Ciência, n. 79, nov. 2006. Disponível em: <http://www.comciencia.br>. Acesso em: 30 ago. 2007.

JAMES, B. R. *Probabilidade*: um curso em nível intermediário. 3. 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.

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

REC. SPORT. SOCCER STATISTICS FOUNDATION (RSSSF). *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. *International Transactions in Operational Research*, v. 12, n. 4, p. 375-386, 2005.

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

SILVA, C. F.; GARCIA, E. S.; SALIBY, E. Soccer Championship Analysis using Monte Carlo Simulation. In: WINTER SIMULATION CONFERENCE. *Proceedings of the 2002 Winter Simulation Conference*, San Diego, v. 1, p. 2.011-2.016, 2002.

SZYMANSKI, S. Economics of Sport: introduction. *The Economic Journal*, Oxford, v. 111, n. 469, p. 1-3, fev. 2001.

VENDITE, C. C.; VENDITE, L. L.; MORAES, A. C. de. Scout no futebol: uma ferramenta para a imprensa esportiva. In: CONGRESSO BRASILEIRO DE CIÊNCIAS DA COMUNICAÇÃO, 28., Rio de Janeiro, 2005. *Anais*... Rio de Janeiro: Intercom, 2005.

Recebido: 21 jan. 2008
Aprovado: 7 maio 2008

Endereço para correspondência
Alysson Ramos Artuso
Rua Padre Isaías de Andrade, 414 – Parolin
Curitiba-PR
CEP 80220-140