

PROBABILIDADES NO FUTEBOL

Bernardo Nunes Borges de Lima, Fábio Enrique Brochero Martínez, Gilcione Nonato Costa, Gustavo Marques Zeferino, Marcelo de Oliveira Terra Cunha, Renato Vidal da Silva Martins

Dep. de Matemática – UFMG

O interesse por probabilidades em assuntos pouco “científicos” é algo que vem de longa data. As chances de ganhar na loteria, de sair a carta que eu quero, de encontrar minha alma gêmea são algumas dentre tantas preocupações comuns, de pessoas comuns, que às vezes esperam dos matemáticos, senão uma resposta cabal, ao menos lampejo de que ainda vale a pena alimentar uma esperança.

E quando o tema é futebol – “a coisa mais importante, dentre as menos importantes” – o interesse aumenta. Aumenta à medida que o campeonato vai chegando a termo, a sina de cada time se desenha, e o torcedor, a própria mídia, vão atrás dos matemáticos como se à procura de um profeta.

Nessas horas perde-se muitas vezes uma ótima oportunidade de divulgação da matemática. Se a postura vai na linha “eis aqui os números que vocês querem”, o matemático segue sendo um ente inatingível, possuidor de uma ciência oculta, acessível talvez a uns poucos agraciados com um dom sabe-se lá de que ordem. Uma resposta no estilo “calcule você as probabilidades de seu time”, além de mais salutar, é muito mais honesta.

A verdade é que há pouco mistério por trás do binômio “probabilidades no futebol”. Os modelos são robustos o bastante para concluirmos que se algum deles não funciona, não é bem porque seu autor não entende de matemática; o que ele não entende é de futebol. De fato, o método de cálculo de probabilidades através de simulações de resultados é de praxe em todo e qualquer modelo. O que varia de modelo para modelo é o modo

como simular um jogo em concreto, e aí contam muito mais nossos conhecimentos futebolísticos do que propriamente matemáticos.

O que fazemos nestas linhas é, primeiro, fundamentar, de forma breve, o método de simulações, que é o principal instrumento de que dispomos na estimativa de probabilidades. E, na sequência, exibimos um modelo de previsão para um campeonato dito “de pontos corridos”. Esperamos que o mesmo enseje o leitor a criar o seu próprio modelo e, assim, estará comprovando o que dizemos acima.

1 A Lei dos Grandes Números à luz do futebol

No ensino médio talvez tenham nos ensinado que para sabermos a probabilidade de um determinado evento, temos de contar, dentre todas as possibilidades em questão, quais são os casos que lhe correspondem. Ou, mais precisamente,

$$\text{probabilidade} = \frac{\text{n}^\circ \text{ de casos favoráveis}}{\text{n}^\circ \text{ de casos possíveis}}. \quad (1)$$

O problema fulcral em uma postura deste tipo é o tanto de hipótese previamente assumida, antes mesmo que qualquer questão probabilística seja de fato colocada. A título de exemplo, contam do professor que perguntou a seu aluno “Qual a probabilidade de sair cara no lançamento de uma moeda não viciada?”. Ao que o aluno replicou “O que é uma moeda não viciada?”. E seu mestre assim o esclareceu: “Aquele cuja probabilidade de sair cara é 1/2”. “Ora, então porque o senhor perguntou?” disse o rapaz com toda razão. De fato, questões probabilísticas não podem nem devem se resumir a meros problemas de contagem posto que entre a “análise combinatória” e a “teoria de probabilidades” vai uma longa distância. No entanto, como bem sabemos, em matemática as coisas não são “verdadeiras” ou “falsas”, mas sim “lógicas” ou “incoerentes” e ninguém nos impede de tomar (1) como ponto de par-

tida e, a partir dela, tentar responder a pergunta “Qual a probabilidade do seu time ser campeão brasileiro?”.

Bem, se tal, há pelo menos dois problemas novos, e sérios, que surgiriam. O primeiro é de ordem computacional, i.e. o total de casos a serem considerados pode ser assustadoramente grande. Imagine o leitor, por exemplo em um campeonato de futebol com 20 times, o que significaria “casos possíveis” a 15 rodadas do final. Levando-se em conta somente o resultado de cada jogo – vitória de um, de outro, ou empate – em uma rodada são 3^{10} os cenários imagináveis. Em 15 rodadas este número sobe para 3^{150} . Bem, em tese, 3^{150} (aproximadamente 3×10^{71}) é até tratável em um computador. De fato, para sistemas criptográficos atuais são usados códigos de tamanho até 2^{1024} , como é o caso da nova codificação do Banco do Brasil. Agora, o que é absolutamente impraticável é gerar em “tempo finito” todos os universos possíveis. Assim, por exemplo, se cada caso fosse gerado em 1 trilhonésimo de segundo (processadores na faixa de Terahertz), o computador gastaria $1,18 \times 10^{50}$ séculos para gerar todos eles. A título de comparação, as estimativas atuais para a idade do Universo apontam para $1,4 \times 10^8$ séculos.

Mas mesmo supondo que nossos computadores fossem de uma geração vindoura, capazes de arrostar grandezas estelares, ainda assim persistiria um outro problema, agora de ordem prática: todos sabemos que nem o primeiro colocado vai perder seus 15 jogos restantes, muito menos o último colocado vencer 15 jogos em sequência. Há muitos cenários que devem ser descartados ou, ao menos, não têm direito de entrar na conta “casos possíveis” com o mesmo peso do que outros cenários bem mais viáveis.

Isto nos leva a refletir sobre o próprio conceito de probabilidade. O fato é que no início do século passado, alguns probabilistas se devotaram a responder precisamente esta mesma pergunta: “afinal de contas, o que é probabilidade?”. O problema se inseria dentro de um contexto mais amplo: a geometria também se perguntava sobre “o que é um ponto”, a teoria dos conjuntos sobre “o que é um conjunto”, e a álgebra sobre “o que

é um número”. O resultado deste esforço é o que se conhece por “axiomática” e a resposta às perguntas acima foi a seguinte: ponto, conjunto, número e também probabilidade, são conceitos que não se definem. A grosso modo, simplesmente “estão aí” e podem ser reconhecidos pelas propriedades que satisfazem.

No caso específico da teoria de probabilidades, o principal nome associado a sua axiomática é o de Kolmogorov. De seus postulados se deduz o modo como eventualmente responderia à pergunta que motivou nosso texto, i.e. quais as chances de um dado time sagrar-se campeão brasileiro. Ante a indagativa, se limitaria a dizer:

- (i) a probabilidade de um time ser campeão é pelo menos 0%;
- (ii) a probabilidade de um time mineiro ser campeão em 2011 é a soma das probabilidades de América, Atlético e Cruzeiro o serem;
- (iii) a probabilidade de um time brasileiro ser campeão é 100%,

e ficaríamos na dúvida ser ele matemático ou político.

O que realmente surpreende é que, na realidade, a “resposta” acima fornece todos os elementos necessários para que possamos de fato estimar a probabilidade de um determinado time ser campeão brasileiro. Ou seja, dos Axiomas de Kolmogorov – dos quais os três itens acima são mera adaptação – se deduz matematicamente a Lei dos Grandes Números, e esta sim fundamenta um método de cálculo de probabilidades bem preciso, i.e. o método de simulações.

Em outras palavras, tomando como ponto de partida o que existe de mais básico capaz de tipificar o verbete “probabilidade”, ou seja, que esta oscila entre 0 e 1 e que para união disjunta, somam-se probabilidades (Axiomas de Kolmogorov) segue um resultado importantíssimo: a probabilidade de um evento é a razão entre o número de ocorrências do mesmo e o número de ensaios quando este tende a infinito (Lei dos Grandes

Números). Ou, mais precisamente,

$$\text{probabilidade} = \lim_{n \rightarrow \infty} \frac{n^\circ \text{ de ocorrências em } n \text{ ensaios}}{n}. \quad (2)$$

Esta afirmação, antes dos Axiomas de Kolmogorov, era a base da *teoria frequencista*, que dava à teoria de probabilidades um caráter, a princípio, mais empírico que matemático. A dedução da Lei dos Grandes Números a partir dos Axiomas de Kolmogorov nos mostra que o “empirismo” em questão já não guarda mais eventuais resquícios pejorativos que o termo possa ter (teste, tentativa, suposição, apriorismo, etc). Ao contrário, o “frequencismo” passa a ser então a nossa grande ferramenta para se estimar probabilidades que (sempre) desconhecemos. Mais ainda, está muito bem fundamentado do ponto de vista teórico, com todo o rigor matemático.

Portanto, se quisermos realmente saber qual a probabilidade de um dado time ser campeão, o que temos de fazer é repetir inúmeros campeonatos e ver em quantos deles o clube leva a melhor. Quanto mais edições, mais preciso o cálculo. Um total de 3 títulos em 4 campeonatos jogados é bem menos significativo do que, digamos, 612 conquistas em 1000 disputas. A probabilidade real da equipe erguer a taça não é então de 75%, como a princípio se cogitava, mas deve girar em torno de 60%.

Ora, o leitor arrazoado dirá que é impossível repetir mil campeonatos e com razão. Mas não iremos “realizar” mil campeonatos e sim “simular” não só mil, mas milhares e até milhões deles se for necessário. O computador faz isto em segundos ou, quando muito, em minutos. Tudo o que temos de fazer é ensiná-lo a “sortear” de forma judiciosa o resultado de cada jogo. Na próxima seção, propomos um modelo de cálculo que esperamos elucidar o leitor.

Gostaríamos de enfatizar, que o modelo que descreveremos abaixo é apenas um dentre os inúmeros que poderiam ser propostos e que o objetivo principal é o de utilizá-lo como veículo de divulgação científica, não havendo qualquer estudo sobre a sua validação estatística.

2 Um modelo de cálculo de probabilidades

Como dissemos antes, o modo de se estimar probabilidades em futebol é simples e de certa forma comum aos diferentes modelos. A ideia é partir da situação atual do campeonato e simular os jogos restantes. Ao final, o computador gera a classificação dos times e a registra. Depois, repete o procedimento um grande número de vezes e no término do processo já está apto a divulgar números que aproximam muito bem as probabilidades geradas pelo modelo idealizado. Se, por exemplo, queremos saber a probabilidade de um dado time ser campeão, basta ver a proporção entre as simulações em que terminou líder sobre o total de simulações.

Para determinar o resultado de um jogo $A \times B$, o computador recebe dois vetores (de entradas não negativas que somam 1) que caracterizam, naquele momento, os times A e B: (PV_A, PE_A, PD_A) e (PV_B, PE_B, PD_B) . Chamamos estes de *vetores de força* de cada time. A partir destes dados, forma-se o vetor probabilidade do jogo, que é

$$P_{A \times B} = \left(\frac{PV_A + PD_B}{2}, \frac{PE_A + PE_B}{2}, \frac{PD_A + PV_B}{2} \right),$$

em que as três coordenadas são, na ordem, as probabilidades de vitória de A, empate, e vitória de B. Supondo, e.g., que $P_{A \times B} = (0.5, 0.2, 0.3)$, o que o computador faz é dividir o intervalo $[0,1]$ em três partes: $[0,0.5]$, $(0.5,0.7)$ e $[0.7, 1]$, e, então, sortear um número aleatório entre 0 e 1 (ele sabe como fazer isto). Se, por exemplo, o número sorteado foi 0.4579, então o resultado do jogo é vitória de A. Depois de simular uma rodada, atualizam-se todos os vetores de força de todos os times, de acordo com os resultados da rodada, e simula-se a rodada seguinte. Repare o leitor que na próxima simulação do campeonato, o resultado de $A \times B$ pode ser diferente, seja porque o computador sorteou um número diferente, seja porque os próprios vetores de força dos times estão diferentes em cada rodada de cada simulação.

Gostaríamos de salientar que o caráter aleatório é dado unicamente pelo sorteio desses números aleató-

rios no intervalo $[0,1]$ (ou, no linguajar dos probabílistas, variáveis aleatórias com distribuição uniforme no intervalo $[0,1]$). Sorteados esses números, todo o resto é determinístico. Observe que sorteamos um desses números para cada simulação de um jogo, portanto para uma simulação do atual Campeonato Brasileiro de Futebol a 20 rodadas do final seriam necessários 200 desses números (o campeonato tem 10 jogos por rodada). Como precisamos repetir essa simulação um número muito grande de vezes, normalmente em torno de 500.000, precisaremos de uns 100 milhões desses números para dar nossos “palpites” sobre o desfecho do campeonato ao final da décima oitava rodada (o campeonato tem 38 rodadas).

O modelo proposto pelos autores destas linhas – cujos resultados efetivos encontram-se disponíveis em www.mat.ufmg.br/futebol e que certamente é apenas um dentre a infinidade de bons modelos para este fim – consiste em atribuir a cada time participante do campeonato, na verdade, dois vetores de força

$$PC = (PVC, PEC, PDC) \text{ e } PF = (PVF, PEF, PDF),$$

em que PVC, PEC e PDC serão utilizados, respectivamente, no cálculo das probabilidades de vitória, empate e derrota do clube em uma partida jogando como mandante (em “casa”), e PVF, PEF e PDF seguem a mesma lógica, desta feita sendo a equipe em questão visitante (joga “fora de casa”).

As condições impostas aos vetores acima para que gerem, de fato, vetores de probabilidade é que suas coordenadas devem ser números não negativos que somem 1. Assim, nenhum time terá chance “negativa” de vencer, ou empatar, ou perder um jogo e, além disso, é claro, vitória, empate ou derrota são as únicas possibilidades aceitáveis.

A cada rodada, os vetores de força são realimentados. Se o time jogou como mandante, muda-se o seu vetor mandante, e o mesmo vale para seu vetor visitante, caso o jogo tenha sido fora de casa. A ideia de se tratar um mesmo time como dois diferentes – um em casa, e outro fora dela – não é nenhum tipo de esquizo-

frenia futebolística. Ao contrário, reflete apenas o modo como a torcida, ou o próprio campo onde a equipe habitualmente joga, podem influenciar no desempenho do clube. Em suma, jogos iguais em campos distintos são distintos. Além disso, permite que surja naturalmente no modelo a dificuldade de enfrentar certos adversários em certos estádios.

A questão chave é como atualizar os vetores de força de duas equipes que se enfrentam, após a realização do jogo. Pode-se dizer que isto é o que distingue os diferentes modelos de previsão. A primeira premissa da qual partimos é a de que, se um time vence uma partida, aumenta sua probabilidade de vencer a partida seguinte. A premissa reflete algo quase intuitivo no futebol: tem mais chance de vencer quem vence mais. Além disso, este aumento na probabilidade de vitória é tanto maior quanto melhor for o time derrotado. Analogamente, se o time perde uma partida, aumenta sua probabilidade de perder o próximo jogo e o aumento é tanto maior quanto pior for seu adversário. Claramente, existem inúmeras maneiras objetivas para medir a qualidade de um time, como a colocação na tabela de classificação, o número de vitórias etc. O parâmetro que adotamos é o “rendimento” dado por

$$r = \frac{\text{n}^\circ \text{ de pontos conquistados}}{\text{n}^\circ \text{ de pontos disputados}}.$$

Podemos entender o modo como modificamos os tais vetores de probabilidade por meio de um exemplo. Vamos supor que houve o jogo São Paulo x Cruzeiro, disputado no Morumbi, e que o São Paulo saiu-se vencedor do confronto. Então o novo vetor de força do São Paulo como mandante, digamos PC_{Spo}^1 , é obtido a partir do vetor anterior, digamos PC_{Spo}^0 , da seguinte forma:

$$PC_{\text{Spo}}^1 = \frac{p \cdot PC_{\text{Spo}}^0 + r_{\text{Cru}} \cdot (1, 0, 0)}{p + r_{\text{Cru}}},$$

em que r_{Cru} é o rendimento do Cruzeiro (que também deve ser modificado para o próximo jogo) e p é o “peso” que damos ao passado. Por exemplo, se $p = 0$ então o que passou não conta para nada, e o fato de que o São Paulo venceu o Cruzeiro no Morumbi daria ao São

Paulo “força máxima” para a próxima partida no Morumbi. Por outro lado, se damos a p um valor muito alto, os vetores de força seguem quase que inalterados jogo a jogo, e perdemos o “fator moral” trazido pelas vitórias. Ou seja, p indica a sensibilidade dos times em relação ao resultado de uma partida. A escolha de um valor ótimo de p é algo que carece de um estudo estatístico mais adequado, mas para isto precisaríamos de uma série mais longa de Campeonatos Brasileiros com o mesmo regulamento (apenas a título de curiosidade futebolística, o Campeonato Brasileiro de pontos corridos e com 20 clubes, que é o formato atual, se iniciou apenas em 2006).

O outro fator que consideramos nessa fórmula é o rendimento do adversário. O que a fórmula acima diz é que, se o Cruzeiro tivesse, por exemplo, um rendimento muito baixo, então a vitória do São Paulo não seria tão significativa a ponto de alterar substancialmente suas chances de vitória.

De modo semelhante, após a derrota no Morumbi, o vetor visitante do Cruzeiro é modificado da seguinte forma:

$$PF_{Cru}^1 = \frac{p \cdot PF_{Cru}^0 + (1 - r_{Spo}) \cdot (0, 0, 1)}{p + (1 - r_{Spo})}.$$

Na fórmula acima aparece $1 - r_{Spo}$ (e não r_{Spo} , como na anterior) pois o raciocínio se inverte: perder, por exemplo, para um time hipotético que sempre vence ($r = 1$) não significa absolutamente nada ($1 - r = 0$ e o vetor força fica intacto).

Como se viu pelo cálculo acima, em caso de vitória de um dos times, reforça-se a ideia de que vencer uma partida aumenta a probabilidade de vencer a próxima, uma vez que a componente PV aumenta e as demais componentes diminuem. Simetricamente, a probabilidade do time derrotado ser novamente derrotado aumenta.

Contudo, o mesmo não se aplica em caso de empate entre as equipes. Ou seja, empatar uma partida não implica apenas que aumentem as chances de empatar a próxima e diminuam as chances de vencer ou perder. De fato, suponhamos que um time, jogando como vi-

sitante, empatasse com o líder do campeonato. É razoável supor que a sua probabilidade de vencer uma partida, como visitante, também aumente. Da mesma forma, se o empate é com o lanterna do campeonato, a probabilidade de perder a próxima partida é que deve ser aumentada. A grosso modo, empate com time bom é “meia vitória”, empate com time ruim é “meia derrota”.

O modo como pomos isto em números é o seguinte. Voltemos a São Paulo x Cruzeiro no Morumbi, e suponha que houve empate. Se $r_{Cru} \leq 1/2$ (esse número é um pouco arbitrário, mas vamos considerá-lo por simplicidade), então o Cruzeiro não é considerado um time tão forte, e atualizamos o vetor força do São Paulo como mandante da seguinte forma:

$$PC_{Spo}^1 = \frac{p \cdot PC_{Spo}^0 + (1 - 2r_{Cru}) \cdot (0, \frac{1}{2}, \frac{1}{2}) + 2r_{Cru} \cdot (0, 1, 0)}{p + 1}.$$

Ou seja, $(0, \frac{1}{2}, \frac{1}{2})$ é a tal “meia derrota” que dissemos acima e $(0, 1, 0)$ é o empate puro. Entendemos melhor a fórmula considerando os casos extremos. Se o Cruzeiro tiver rendimento $r_{Cru} = 1/2$, então o segundo termo do numerador se anula e o terceiro é $(0, 1, 0)$, ou seja, apenas a tendência ao empate do São Paulo em casa será aumentada. Mas supondo, hipoteticamente, que o Cruzeiro sequer pontuou no campeonato, seu rendimento é nulo e o terceiro termo do numerador desaparece, sendo o segundo igual a $(0, \frac{1}{2}, \frac{1}{2})$. Ou seja, neste caso o São Paulo teve uma “meia derrota” para o Cruzeiro apesar de ter, na prática, empatado o jogo, o que deve fazer crescer tanto sua tendência a futuros empates quanto derrotas. Os casos intermediários $0 < r_{Cru} < 1/2$ correspondem à média ponderada entre os vetores $(0, \frac{1}{2}, \frac{1}{2})$ e $(0, 1, 0)$, dependendo do rendimento do Cruzeiro.

Agora, se o Cruzeiro vem forte no campeonato, digamos com rendimento $r_{Cru} > 1/2$, então o empate do São Paulo está mais para “meia vitória”. Desse modo, o vetor probabilidade do São Paulo é atualizado assim:

$$PC_{Spo}^1 = \frac{p \cdot PC_{Spo}^0 + (2r_{Cru} - 1) \cdot (\frac{1}{2}, \frac{1}{2}, 0) + 2(1 - r_{Cru}) \cdot (0, 1, 0)}{p + 1}.$$

Após o empate, o vetor probabilidade do Cruzeiro como visitante é atualizado de forma absolutamente análoga, levando-se em conta o rendimento r_{Spo} do São Paulo.

Resumindo, suponha que o campeonato teve sua vigésima rodada realizada e que os vetores de força foram atualizados segundo o algoritmo acima, utilizando-se os resultados reais ocorridos até a vigésima rodada. Então o programa simula a vigésima primeira rodada, atualiza os vetores de força e a tabela de classificação (somando 1 ponto por empate e 3 por vitória); em seguida com esta nova tabela e os novos vetores de força simula a próxima rodada, e assim sucessivamente, até simular a última rodada. Terminada a simulação de um campeonato, neste momento fazemos as perguntas pertinentes, por exemplo, o Flamengo será rebaixado? Todas as respostas devem ser da forma sim ou não. Em seguida, repetimos este procedimento um número grande de vezes (digamos 1.000.000 de vezes), de modo que o valor considerado para a probabilidade do Flamengo ser rebaixado no fim do campeonato concluída a vigésima rodada seja o número de simulações em que o Flamengo foi rebaixado dividido pelo número de simulações realizadas.

Uma pergunta natural que o leitor já deve ter se questionado é sobre quais seriam as condições iniciais dos vetores de força. Leva-se em consideração o campeonato anterior? Começa-se com $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$? Nossa experiência mostra que isto faz pouca diferença quando simulamos um campeonato de pontos corridos, quando já se passaram 10, 15 rodadas.

É lógico que o modelo acima é passível de crítica. Será que $p = 5$ é um bom peso? Sequência muito longa de vitórias não começa a aumentar a chance do time perder? O histórico de jogos entre dois times não mereceria também ser levado em conta? Bem, este é apenas um modelo e o leitor aguçado pode e deve criar algum outro de seu agrado (nós já criamos vários). Se com ele responder melhor às perguntas acima e tantas outras que possam vir à tona, não será pelos seus conhecimentos matemáticos, muito menos por eventual

vínculo acadêmico. Ao contrário, as questões mais relevantes em probabilidades no futebol são questões de futebol.

Nota. Este trabalho é parte do Projeto Difundindo Probabilidades via Campeonatos de Futebol, que conta com apoio da FAPEMIG por meio do Edital de Popularização da Ciência e Tecnologia.

Referências

- [1] JAMES, J. *Probabilidade: um curso em nível intermediário*. 3. ed. Rio de Janeiro: IMPA, 2008. (Projeto Euclides)
- [2] BROCHERO, F.; COSTA, G. N.; TERRA CUNHA, M.; DE LIMA, B. N. B.; MARTINS, R. V. Futebol: uma caixinha de... sorteios. *Ciência Hoje*, v. 254, p.24–29, 2008.
- [3] RICHARD, I. *The pleasures of probability*. New York: Springer, 1995. (Undergraduate Texts in Mathematics; Readings in Mathematics)

Bernardo Nunes Borges de Lima

bnblima@mat.ufmg.br

Fábio Enrique Brochero Martínez

fbrocher@mat.ufmg.br

Gilcione Nonato Costa

gilcione@mat.ufmg.br

Gustavo Marques Zeferino

gumaze@gmail.com

Marcelo de Oliveira Terra Cunha

tcunha@mat.ufmg.br

Renato Vidal da Silva Martins

renato@mat.ufmg.br

Departamento de Matemática

Instituto de Ciências Exatas

Universidade Federal de Minas Gerais