UFSM

CIÊNCIA e NATURA

ISSN 2179-460X

ACESSO ABERTO

Ci. e Nat., Santa Maria, v. 47, e85273, 2025 • https://doi.org/10.5902/2179460X85273

Submissão: 01/10/23 • Aprovação: 11/04/25 • Publicação: 14/07/25

Estatística

Modelagem probabilística e inferência do efeito de casa em partidas esportivas

Probabilistic modelling and inference of home effect in sports matches

Giovani Festa Paludo [I], **Eric Batista Ferreira** [I]

[I] Universidade Federal de Alfenas, Alfenas, MG, Brasil

RESUMO

Uma variável aleatória (v.a.) construída para a obtenção do efeito de casa em campeonatos de futebol baseado em pontos, foi recentemente proposta por Paludo, Figueiredo e Ferreira (2023). Trata-se da diferença relativa de pontos entre casa e fora de casa e foi denominada de D. Uma vez proposta, faz-se necessário conhecer a sua distribuição para a realização de inferências, sendo que D é composta pela combinação linear de outras 4 v.a., e foi assumido que o problema surge de duas multinomiais independentes. A partir de definições das distribuições, foram obtidas médias e variâncias e, com isso, foram desenvolvidas duas aproximações: uma pela distribuição normal e; outra pela binomial. Em seguida foi realizado um estudo de simulação para avaliação de qual aproximação poderia ser utilizada e aplicações foram apresentadas para exemplificar novas inferências que foram possibilitadas com o presente estudo. Os dados simulados de D, tanto partindo de vetores hipotéticos quanto observados no Campeonato Brasileiro, apresentaram baixas porcentagens de aderência à binomial, porém tiveram expressiva aderência à normal, sendo que essa foi a distribuição utilizada. As aplicações utilizando a distribuição mostraram diferentes possibilidades de utilização para inferências, uma delas inclusive quando há dados com diferentes números de partidas em casa e fora. Foi possível obter uma representação adequada da v.a. que permitirá realizar inferências acerca de D.

Palavras-chave: Vantagem de casa; Aproximação de distribuição; Futebol; Estatística esportiva

ABSTRACT

A point based random variable (r.v.) built to obtain the home ground effect in soccer leagues was recently proposed by Paludo, Figueireido e Ferreira (2023). It is the points' relative diference between home and away and it was named D. Once it was proposed it's necessary to know its distribution to make inferences. D is composed by the linear combination of 4 r.v. and it was assumed that the problem origins of two independent multionomial r.v.. From the distributions definitions we have obtained expectation and variances and based on this, we have developed two approximations, one by

normal and the other by the binomial distribution. Subsequently it was conducted a simulation study to evaluate the approximations and applications were presented to exemplify new inferences possibilitaded by this paper. The simulated D values from both hypothetical and observed vectors presented low percentages of adherence to the binomial distribution. In the greater part of analyzed situations, the data was expressively adhered to the normal distribution, and this was the selected distribution. The applications using distributions have showed different possibilities of use for inferences, including one when there are different numbers of home and away matches. It was possible to represent properly the r.v. that will permit to make inferences around D.

Keywords: Home win; Distribution approximation; Home cooking; Soccer; Sports statistics

1 INTRODUÇÃO

Esportes profissionais são também importantes atividades de entretenimento a nível global, sendo que geram um grande número de empregos, e por isso são estudados sob vários aspectos. Uma questão que é estudada há bastante tempo, mas que não foi completamente elucidada é a vantagem que um time possui quando está jogando na sua casa (chamada frequentemente de vantagem de casa ou HA). Sendo que o estudo do efeito da casa em esportes é de interesse de várias áreas, desde a educação física (Dawson et al., 2020), psicologia e medicina do esporte (Nevill & Holder, 1999; McCarrick et al., 2021), pesquisa operacional (Goller & Krumer, 2020), estatística (Benz & Lopez, 2021), economia (Van-Ours, 2019; Hegarty, 2021), riscos do mercado de apostas (Marek & Vávra, 2020), entre outras. Ainda, a HA pode ser encontrada em vários esportes (Pollard et al., 2017). No futebol, a vantagem de casa pode ser obtida de duas maneiras principais: utilizando-se pontos (ver Pollard et al. 2008) ou; saldo de gols (ver Marek e Vávra 2020). Como várias das competições mais populares de futebol tem como objetivo a conquista de pontos, isso faz com que estudar a vantagem de casa por pontos seja também relevante.

Um time habilidoso tende a ter um desempenho melhor tanto quando joga em casa quanto quando joga fora de casa, por isso que quando se estuda a vantagem de casa por pontos, frequentemente é necessária a utilização de uma correção pela habilidade do time (Pollard et al., 2008). Porém, foi desenvolvida uma métrica para medir a vantagem de casa que não é inflacionada ou não sofre efeitos diretos relacionados à habilidade do time (Paludo et al., 2023). Tal métrica foi denominada de D, sendo que ela também é uma variável aleatória (v.a.) que ainda não possui estudos em relação à sua distribuição e inferência estatística.

Nas ciências empíricas, a inferência indutiva é importante e é utilizada para encontrar novos conhecimentos, sendo que é uma função da Estatística, providenciar técnicas para fazer inferência indutiva e medir o grau de incerteza das inferências (Mood et al., 1974). Ainda, essa incerteza é medida em termos de probabilidade (Mood et al., 1974). Por isso, faz-se necessário estudar a distribuição de probabilidade de uma variável aleatória para a realização de inferências com controle sob o grau de incerteza. E a distribuição de probabilidade é obtida em estudos de modelagem probabilística. Estes estudos não são raros (exemplos em Usman et al. (2018); Alzubaidi et al. (2022)) e encontrar uma distribuição de probabilidade apropriada leva a obter resultados mais acurados (exemplo em Alzubaidi et al. (2022)).

Por isso, o objetivo do presente estudo foi de: (i) obter a distribuição da v.a. D; (ii) avaliar a distribuição aproximada; (iii) obter os estimadores dos parâmetros da distribuição de D e; (iv) ilustrar com dados reais os resultados obtidos.

2 MATERIAL E MÉTODOS

Serão abordados e descritos conceitos iniciais do desenvolvimento das aproximações para a distribuição de D, a metodologia utilizada para comparação das aproximações e a descrição de métodos utilizados nas aplicações.

2.1 Variável aleatória D e sua respectiva distribuição de probabilidade

Paludo et al. (2023) definiram uma métrica d para obtenção da vantagem de casa. No presente estudo, entende-se que a métrica d é uma variável aleatória, portanto será representada sempre por D. Desta maneira, define-se:

Definição (diferença de pontos relativa): *A diferença de pontos relativa D é a razão entre a diferença de pontos (diferença entre casa e fora) e o total de pontos que o time concorre ao jogar tais partidas, isto é,*

$$D = \frac{Y_c - Y_f}{c_v \times n_c},$$

em que Y_c e Y_f representam a soma de pontos conquistados pelo time em casa e fora de casa, respectivamente, c_v refere-se ao número de pontos atribuídos a cada vitória e n_c é o número total de partidas em casa.

Ressalta-se que qualquer realização da variável aleatória D, isto é, qualquer valor observado, será representado por d. Desta forma, quando um time conquista todos os pontos em casa e nenhum ponto fora de casa, o valor que a diferença de pontos relativa assume é $d = 1{,}0$, enquanto que, quando um time conquista todos os pontos fora de casa e nenhum em casa, $d = -1{,}0$.

Para encontrarmos a distribuição de D, necessitaremos de algumas definições prévias. Inicialmente definiremos o vetor **X**. Seja $\mathbf{X}$ o vetor aleatório tri-dimensional composto pelas variáveis aleatórias *número de vitórias* (X_v), *número empates* (X_e) e *número de derrotas* (X_d), que um dado time obtém ao final de um campeonato, ou seja,

$\mathbf{X}^\top = (X_v, X_e, X_d)$.

Para essas variáveis aleatórias há a restrição de que o número de partidas n, do referido campeonato, obedece a:

$$n = X_v + X_e + X_d. \tag{1}$$

Contudo, estamos interessados em modelar campeonatos que, especificamente, alocam metade das partidas de um time em sua casa e a outra metade fora de casa, onde ele é visitante, isto é,

$n_c = n_f$,

$n = n_c + n_f = 2n_c$,

em que n_c é o número total de partidas em casa e n_f o número total de partidas fora.

Quando consideramos que as vitórias, empates e derrotas podem acontecer em *casa* (c) ou fora de casa (f), temos a seguinte partição

$$\mathbf{X} = \begin{pmatrix} X_v = X_{vc} + X_{vf} \\ X_e = X_{ec} + X_{ef} \\ X_s = X_{dc} + X_{df} \end{pmatrix} = \begin{pmatrix} X_{vc} \\ X_{ec} \\ X_{dc} \end{pmatrix} + \begin{pmatrix} X_{vf} \\ X_{ef} \\ X_{df} \end{pmatrix} = \mathbf{X}_c + \mathbf{X}_f.$$

Assumimos que há independência entre os vetores $\mathbf{X}_c$ e $\mathbf{X}_f$, ou seja, a abordagem escolhida para representar a situação assume que os resultados em casa e fora de casa são independentes.

Analogamente a (1), existe uma restrição para as variáveis aleatórias pertencentes aos vetores $\mathbf{X}_c$ e $\mathbf{X}_f$, que diz que a soma do número de vitórias, empates e derrotas *em casa* é igual ao número total de partidas feitas *em casa* (o mesmo ocorre para fora de casa). Assim temos,

$$n_c = X_{vc} + X_{ec} + X_{dc},$$
$$n_f = X_{vf} + X_{ef} + X_{df}.$$

É razoável assumir que, devido à natureza do problema, $\mathbf{X}_c$ e $\mathbf{X}_f$ seguem uma distribuição multinomial com os seguintes parâmetros e denotada por:

$$\mathbf{X}_c \sim Multi(n_c, p_{vc}, p_{ec}, p_{dc}), \tag{2}$$
$$\mathbf{X}_f \sim Multi(n_f, p_{vf}, p_{ef}, p_{df}), \tag{3}$$

em que p_{vc} é a probabilidade de vitória em casa, p_{ec} a probabilidade de empate em casa, p_{dc} a probabilidade de derrota em casa, p_{vf}, p_{ef} e p_{df} são, respectivamente, as probabilidades de vitória, empate e derrota fora de casa. Observe ainda que, dadas as probabilidades de vitória e empate, a probabilidade de derrota fica automaticamente determinada. Então, decorre que $p_{dc} = 1 - p_{vc} - p_{ec}$ e $p_{df} = 1 - p_{vf} - p_{ef}$.

Como $\mathbf{X}_c$ e $\mathbf{X}_f$ são multinomiais, temos as seguintes propriedades, para $i = \{c,f\}$ e $j = \{v,e,d\}$ e $\{i,j\} \neq \{i',j'\}$,

$$E[X_{ij}] = n_i p_{ij},$$
$$Var[X_{ij}] = n_i p_{ij}(1 - p_{ij}), \tag{4}$$
$$Cov[X_{ij},X_{i'j'}] = -n_i p_{ij} p_{i'j'}. \tag{5}$$

Outra propriedade advinda da distribuição multinomial é que, marginalmente, os números de vitórias e empates (dentro e fora de casa) seguem distribuições binomiais dependentes. Isto é, para os jogos em casa, temos que

$$X_{vc} \sim Bin(n_c,p_{vc}),$$
$$X_{ec} \sim Bin(n_c - x_{vc},p_{ec}),$$

e o mesmo vale para fora de casa.

Com base nisso, considere a variável aleatória *número de pontos ganhos* (Y), definida como uma combinação linear do número de vitórias e empates, ponderados pelos números de pontos atribuídos a cada vitória (c_v) e a cada empate (c_e):

$$Y_c = c_v X_{vc} + c_e X_{ec}, \tag{6}$$

$$Y_f = c_v X_{vf} + c_e X_{ef}. \tag{7}$$

Assim, aplicando as propriedades de esperança, temos que as respectivas esperanças de Y_c e Y_f são:

$$E[Y_c] = c_v n_c \left(p_{vc} + \frac{c_e}{c_v} p_{ec} \right), \tag{8}$$

$$E[Y_f] = c_v n_f \left(p_{vf} + \frac{c_e}{c_v} p_{ef} \right). \tag{9}$$

Considere a seguinte reparametrização:

$$a_c = p_{vc} + \frac{c_e}{c_v} p_{ec},$$

$$a_f = p_{vf} + \frac{c_e}{c_v} p_{ef}.$$

A partir disso, temos que (8) e (9) podem ser reescritos como:

$$E[Y_c] = c_v n_c a_c, \tag{10}$$

$$E[Y_f] = c_v n_f a_f. \tag{11}$$

Além disso, utilizando (4), (5), (6) e (7), temos que as variâncias de Y_c e Y_f são

$$\begin{aligned} Var[Y_c] &= Var[c_v X_{vc} + c_e X_{ec}] \\ &= c_v^2 Var[X_{vc}] + c_e^2 Var[X_{ec}] + 2 c_v c_e Cov[X_{vc}, X_{ec}] \\ &= c_v^2 n_c p_{vc}(1 - p_{vc}) + c_e^2 n_c p_{ec}(1 - p_{ec}) - 2 c_v c_e n_c p_{vc} p_{ec} \\ &= n_c c_v \left[c_v p_{vc}(1 - p_{vc}) + \frac{c_e^2}{c_v} p_{ec}(1 - p_{ec}) - 2 c_e p_{vc} p_{ec} \right], \end{aligned} \tag{12}$$

$$Var[Y_f] = n_f c_v \left[c_v p_{vf}(1 - p_{vf}) + \frac{c_e^2}{c_v} p_{ef}(1 - p_{ef}) - 2 c_e p_{vf} p_{ef} \right] \tag{13}$$

Note que a distribuição de probabilidade de Y_c e Y_f não é exatamente binomial, devido à dependência entre as variáveis X. Isso pode ser visto pela própria definição de variável aleatória binomial. Como uma binomial é definida como a soma de variáveis

Bernoulli independentes, ao somar duas variáveis binomiais dependentes, gera-se uma sequência de variáveis Bernoulli que não são mais independentes.

As distribuições exatas de Y_c e Y_f são complexas de se obter e não equivalem a alguma distribuição conhecida (ver Vellaisamy & Punnen (2001) e Butler & Stephens (2017)), sendo que não serão abordadas nesse trabalho. No entanto, alternativamente, utilizaremos duas aproximações para descrever essas distribuições: a primeira aproximação é dada por uma distribuição binomial e a segunda por uma distribuição normal.

2.2 Estudo de simulação

Para avaliar a adequabilidade dos dados às aproximações das distribuições apresentadas, realizou-se um estudo de simulação onde foram utilizados vetores $\mathbf{p}_c = (p_{vc}, p_{ec}, p_{dc})$ e $\mathbf{p}_f = (p_{vf}, p_{ef}, p_{df})$, hipotéticos e observados (Tabela 1 e 2). Os vetores hipotéticos foram escolhidos de maneira que pudessem ser comparadas diferentes situações que estão descritas na Tabela 1, como por exemplo vetores $\mathbf{p}_c = \mathbf{p}_f$, $p_{vc} > p_{vf}$, entre outras situações. No total foram 18 combinações de vetores $\mathbf{p}_c$ e $\mathbf{p}_v$ hipotéticos. Já os vetores observados foram obtidos do Campeonato Brasileiro de Futebol, sendo que foram construídos dois cenários. O primeiro cenário foi com base na seleção de 10 combinações de vetores $\hat{p}_c$ e $\hat{p}_f$ estimados à partir de dados observados obtidos em participações dos times no Campeonato Brasileiro. Esses vetores foram escolhidos com o objetivo que fossem representadas situações distintas que aconteceram no Campeonato Brasileiro (considerando-se as edições de 2006 à 2022). Das 10 combinações, 5 foram escolhidos de maneira que fossem incluídas diferentes combinações em casa e os outros 5 para incluir diferentes situações fora de casa. E, por sua vez, o segundo comparativo consistiu na utilização de todos os parâmetros estimados que foram observados entre 2006 e 2019 no Campeonato Brasileiro. Isto é, foram simuladas 278 (seriam 280 participações, porém duas participações não puderam ser utilizadas por apresentarem um número desbalanceados de partidas entre fora e casa) combinações de vetores e as linhas foram sobrepostas num gráfico.

Cabe ressaltar que o vetor de probabilidades em uma participação foi obtido com base na divisão dos números de vitórias, empates e derrotas pelo número de partidas disputadas. Isto é, se um time venceu 10 partidas em 19 que disputou em casa, ficaria com um $\hat{p}_{vc} = 0{,}526$.

Tabela 1 – Vetores $\mathbf{p}_c$ e $\mathbf{p}_c$ hipotéticos utilizados nas simulações para avaliação das duas aproximações

Número	Descrição	$\mathbf{p}_c^\top$	$\mathbf{p}_f^\top$
1	Vetores p_c e p_f iguais	(0,15; 0,15; 0,7)	(0,15; 0,15; 0,7)
2	Vetores p_c e p_f iguais	(0,33; 0,33; 0,34)	(0,33 ;0,33 ;0,34)
3	Vetores p_c e p_f iguais	(0,5; 0,5; 0)	(0,5; 0,5; 0)
4	Vetores p_c e p_f iguais	(0,5; 0,3; 0,2)	(0,5; 0,3; 0,2)
5	Vetores p_c e p_f iguais	(0,6; 0,1; 0,3)	(0,6; 0,1; 0,3)
6	Vetores p_c e p_f iguais	(0,8; 0,1; 0,1)	(0,8; 0,1; 0,1)
7	Vetores p_c e p_f iguais	(0,2; 0,7; 0,1)	(0,2; 0,7; 0,1)
8	mais pontos em casa	(0,5; 0,3; 0,2)	(0,4; 0,4; 0,2)
9	$p_{vc} > p_{vf}$	(0,8; 0,1; 0,1)	(0,5; 0,1; 0,4)
10	mais pontos em casa	(0,8; 0,1; 0,1)	(0,33; 0,33; 0,34)
11	$p_{vc} > p_{vf}$	(0,8; 0,1; 0,1)	(0,1; 0,1; 0,8)
12	mais pontos em casa	(0,2; 0,7; 0,1)	(0,1; 0,8; 0,1)
13	$p_{vc} < p_{vf}$ e $p_{ec} < p_{ef}$	(0,3; 0,1; 0,8)	(0,5; 0,3; 0,2)
14	$p_{vc} < p_{vf}$ e $p_{ec} < p_{ef}$	(0,3; 0,3; 0,4)	(0,5; 0,5; 0)
15	$p_{vc} < p_{vf}$	(0,33; 0,33; 0,34)	(0,7; 0,2; 0,1)
16	$p_{vc} < p_{vf}$	(0,5; 0,3; 0,2)	(0,8; 0,2; 0)
17	$p_{ec} < p_{ef}$	(0,2; 0,3; 0,5)	(0,2; 0,6; 0,2)
18	$p_{ec} < p_{ef}$	(0,2; 0,3; 0,2)	(0,2; 0,8; 0)

Fonte: autores (2023)

A partir de cada combinação dos dois vetores de probabilidade foram calculadas a média e a variância populacional (σ_D^2). Para a aproximação da distribuição de D pela binomial, obtiveram-se os parâmetros n e p também a partir dos vetores de probabilidade $\mathbf{p}_c$ e $\mathbf{p}_f$.

Em seguida, a partir do uso do Software R (R Core Team, 2024), gerou-se uma amostra de dados de tamanho (s) que assumiu os tamanhos de $s = 10, 20, \ldots, 300$ observações a partir da função `rmultinom()`. Para cada tamanho de amostra foram realizadas 1000 iterações, sendo que a cada iteração era gerada uma nova amostra de tamanho s a partir da função `rmultinom()`.

Para verificar se a cada iteração de um tamanho amostral s ou de uma situação (nas 28 situações descritas nas Tabelas 1 e 2), os dados se ajustavam a cada uma das duas aproximações, foi utilizado o teste de aderência de qui-quadrado (χ^2), ambos com

um nível de significância de 5%. Em ambos os casos, testou-se as seguintes hipóteses: $\mathcal{H}_0$: se os dados podiam ser considerados provenientes da distribuição em questão e; $\mathcal{H}_1$: se os dados não podiam ser considerados como provenientes da distribuição em questão.

Tabela 2 – Vetores de probabilidade $\hat{\mathbf{p}}_c$ e $\hat{\mathbf{p}}_c$ estimados a partir de dados observados no Campeonato Brasileiro de Futebol e utilizados nas simulações para avaliação das duas aproximações. Em que * indica o único caso que foi utilizado um dado da edição de 2021 do Campeonato Brasileiro de Futebol

Número	Descrição	$\hat{\mathbf{p}}_c^\top$	$\hat{\mathbf{p}}_f^\top$
1	Atlético-GO em 2012	(5/19; 3/19; 11/19)	(2/19; 6/19; 11/19)
2	Bahia em 2012	(5/19; 9/19; 5/19)	(6/19; 5/19; 8/19)
3	Goiás em 2006	(9/19; 5/19; 5/19)	(6/19; 5/19; 8/19)
4	Cruzeiro em 2006	(10/19; 8/19; 1/19)	(4/19; 3/19; 12/19)
5	Internacional em 2015	(14/19; 3/19; 2/19)	(3/19; 6/19; 10/19)
6	Ceará em 2019	(8/19; 6/19; 5/19)	(2/19; 3/19; 14/19)
7	Cruzeiro em 2019	(5/19; 8/19; 6/19)	(2/19; 7/19; 10/19)
8	Fluminense em 2016	(8/19; 6/19; 5/19)	(5/19; 5/19; 9/19)
9	São Paulo em 2006	(14/19; 4/19; 1/19)	(8/19; 8/19; 3/19)
10	Palmeiras em 2021*	(11/19; 3/19; 5/19)	(9/19; 3/19; 7/19)

Fonte: autores (2023)

Se a amostra gerada na iteração j de tamanho s, seguia a distribuição aproximada testada, registrou-se o valor 1, caso contrário, 0. Em seguida, somou-se a quantidade de iterações nas quais $\mathcal{H}_0$ não foi rejeitada e dividiu-se pelo total de iterações, e assim, foi obtido a variável que aqui foi chamada de “aderência”. Onde 100% de aderência representava que em todas iterações os dados aderiram à distribuição testada e 0% representava que em nenhuma iteração os dados aderiram à distribuição testada.

2.3 Aplicações

Foram construídas três aplicações, todas utilizando dados de campeonatos de futebol. E como as aplicações tiveram como ideia central representar diferentes possibilidades de uso da distribuição de D, portanto, os dados utilizados variaram entre cada uma das aplicações.

Os dados da primeira divisão da *La Liga* da Espanha, *Premier League* da Inglaterra e da *Serie A* da Itália de 2006/2007 à 2019/2020 foram obtidos através do pacote do Software R `engsoccerdata` (Curley, 2020) e os dados de 2020/2021 à 2022/2023 foram obtidos no website `www.soccerway.com`. Os dados de todos os times que participaram

do Campeonato Brasileiro de 2006 à 2022 e os dados da participação de um time da liga nacional da Argentina de 2018/2019 foram obtidos também em `www.soccerway.com`, website já utilizado nos estudos de Pollard et al. (2008) e Silva et al. (2018). Em todas as aplicações, os dados obtidos foram: o nome do time, o número de vitórias, empates e derrotas em casa e fora de casa.

A aplicação 1 se baseou na comparação da média amostral $\bar{d}$ entre dois times que participaram de um mesmo campeonato e foi baseada em dados do Campeonato Brasileiro de 2006 a 2019. A aplicação 2 foi baseada em dados da primeira divisão das 4 ligas de futebol, de 2006 à 2022 para o caso do Campeonato Brasileiro e de 2006/2007 à 2022/2023 no caso da *La Liga* da Espanha, *Premier League* da Inglaterra e da *Serie A* da Itália. Neste caso foi realizada uma comparação das médias de d das quatro ligas pelo teste t de comparação de médias. Também nessa aplicação foi apresentada uma representação longitudinal dos valores de d ao longo das edições. E por último, a aplicação 3 consistiu na construção de um exemplo para obtenção da média amostral $\bar{d}$ para campeonatos desbalanceados, isto é, onde o número de partidas em casa e fora de casa não é igual. Nessa aplicação foram utilizados os resultados da participação do River Plate na liga nacional da Argentina em 2018/2019.

Nas aplicações 1 e 2, o seguinte procedimento foi utilizado para realização do teste de hipóteses. Considere $D_1 \dot{\sim} \mathrm{N}(\mu_1, \sigma_1^2)$ e $D_2 \dot{\sim} \mathrm{N}(\mu_2, \sigma_2^2)$ como sendo a v.a. D aplicada à duas populações (1 e 2), isto é, D_1 representa a diferença de pontos obtida por um time (população 1) em todas as edições da competição e D_2 representa a diferença de pontos obtida por outro time (população 2) em todas as edições da competição sob consideração. Queremos testar se a média de D é igual para as duas populações, isto é, realizaremos o seguinte teste de hipóteses, $\mathcal{H}_0 : \mu_1 = \mu_2$ e $\mathcal{H}_1 : \mu_1 \neq \mu_2$. O procedimento para a realização do teste de hipóteses envolveu as seguintes etapas: (i) obtenção das médias amostrais ($\bar{D}$) e o desvio padrão amostral (S^2); (ii) Aplicação do teste F para verificar se as variâncias são homogêneas (Bolfarine & Sandoval, 2001); (iii) se as variâncias são homogêneas aplica-se o teste t para variâncias homogêneas Zar (2010), ou para variâncias heterogêneas Zar (2010) (neste caso aproximam-se os graus de liberdade para permitir a utilização da estatística t) e; (iv) realização da decisão e conclusão sobre o resultado do teste de hipóteses.

A aplicação 3 consistiu na obtenção da média amostral ($\bar{d}$) para um campeonato que o número de partidas em casa e fora são diferentes, inicialmente deve-se obter as estimativas para os vetores de probabilidade $\mathbf{p}_c$ e $\mathbf{p}_f$. Para a obtenção das estimativas, utilizam-se as frequências relativas, isto é, $\hat{p}_{vc} = X_{vc}/n_c$. De posse das estimativas dos vetores $\mathbf{p}_c$ e $\mathbf{p}_f$, calcula-se:

$$a_c = \hat{p}_{vc} + \frac{c_e}{c_v}\hat{p}_{ec}$$

e

$$a_f = \hat{p}_{vf} + \frac{c_e}{c_v}\hat{p}_{ef}.$$

Desta maneira, pode-se obter uma média amostral de D, isto é, $\bar{d}$ com base na diferença entre a_c e a_f, isto é,

$$\bar{d} = a_c - a_f.$$

Com isso obteve-se $\bar{d}$ para um time quando o campeonato é desbalanceado.

O estudo de simulação, os testes estatísticos e as ilustrações foram todas construídas utilizando o Software R (R Core Team, 2024). Ressalta-se que: D é a v.a. do efeito de casa chamada de diferença relativa de pontos; μ_{fla} é a média populacional do efeito de casa para o time do Flamengo por exemplo; d é uma realização da v.a., isto é, uma observação; $\bar{d}$ é a média amostral do efeito de casa e; $\hat{\mathbf{p}}$ é uma estimativa de um vetor de parâmetros populacionais $\mathbf{p}$.

3 RESULTADOS E DISCUSSÃO

Inicialmente são apresentados as aproximações, seguido pelo estudo de simulação e aplicações.

3.1 Resultados metodológicos

São apresentadas na sequência, as aproximações binomiais e normais obtidas para Y e para D.

3.1.1 Aproximações binomiais para as distribuições de Y_c e Y_f

Uma primeira aproximação possível é feita pela própria binomial. Note, porém, que devido à dependência entre as vitórias e empates num dado local, as variâncias

(12) e (13) não se comportam como "npq". Por exemplo, para dentro de casa, a variância tipicamente binomial é diferente da variância apresentada pela variável Y_c:

$$n_c c_v \left[c_v p_{vc}(1-p_{vc}) + c_e^2/c_v p_{ec}(1-p_{ec}) - 2c_e p_{vc} p_{ec}\right] \neq n_c c_v a_c(1-a_c).$$

Vamos assumir as esperanças exatas (dadas em 10 e 11) e variâncias aproximadas, para garantir a exigência

$$Var[\cdot] = E[\cdot](1-p) \tag{14}$$

da binomial seja atendida. Para isso, escrevermos a aproximação:

$$Var[Y_c] = n_c c_v a_c(1-a_c) \tag{15}$$

$$Var[Y_f] = n_f c_v a_f(1-a_f). \tag{16}$$

Agora, precisamos estabelecer o número máximo de sucessos para as variáveis Y_c e Y_f. Notamos que os pontos ganhos provém de duas fontes: vitórias e empates. No entanto esses pontos ganhos são ponderados por c_v e c_e, respectivamente. Sendo assim, o valor máximo de sucessos da variável Y_c é $c_v n_c$ (e, analogamente, para Y_f é $c_v n_f$).

A probabilidade de sucesso de uma binomial que descreve aproximadamente o comportamento de Y_c deve se compor pela probabilidade de vencer mais a probabilidade de empatar, ponderada pelo incremento percentual que essa segunda probabilidade traz. Sendo assim, temos que,

$$Y_c \dot{\sim} Bin\left(c_v n_c,\ a_c\right),$$
$$Y_f \dot{\sim} Bin\left(c_v n_f,\ a_f\right).$$

3.1.2 Aproximações normais para as distribuições de Y_c e Y_f

Uma segunda aproximação que se pode considerar é a aproximação normal. Como é usual, variáveis binomiais são aproximadas por normais com média $\mu = np$ e $\sigma^2 = npq$. No nosso contexto, podemos seguir um de dois caminhos. No primeiro, utilizamos esperanças exatas (10 e 11) e variâncias aproximadas (15 e 16). Sendo assim, temos que:

$$Y_c \dot{\sim} N\left(n_c c_v a_c,\ n_c c_v a_c(1-a_c)\right),$$
$$Y_f \dot{\sim} N\left(n_f c_v a_f,\ n_f c_v a_f(1-a_f)\right).$$

E no segundo, utilizamos esperanças exatas (10 e 11) e variâncias exatas (12 e 13). Sendo assim, temos que:

$$Y_c \dot{\sim} N\left(n_c c_v a_c,\ n_c c_v\left[c_v p_{vc}(1-p_{vc}) + \frac{c_e^2}{c_v} p_{ec}(1-p_{ec}) - 2c_e p_{vc} p_{ec}\right]\right),$$
$$Y_f \dot{\sim} N\left(n_f c_v a_f,\ n_f c_v\left[c_v p_{vf}(1-p_{vf}) + \frac{c_e^2}{c_v} p_{ef}(1-p_{ef}) - 2c_e p_{vf} p_{ef}\right]\right).$$

3.1.3 Aproximação binomial para a distribuição de D

Finalmente, considere D a variável que denota a diferença entre pontos ganhos em casa e fora de casa, relativa ao total de pontos distribuídos no campeonato

$$D = \frac{Y_c - Y_f}{c_v n_c}.$$

Assim, a variável aleatória D tem esperança

$$\begin{aligned}
E[D] &= E\left[\frac{Y_c - Y_f}{c_v n_c}\right] \\
&= \frac{1}{c_v n_c}\left(E[Y_c] - E[Y_f]\right) \\
&= \frac{1}{c_v n_c}\left(c_v E[X_{vc}] + c_e E[X_{ec}] - c_v E[X_{vf}] - c_e E[X_{ef}]\right) \\
&= \frac{1}{c_v n_c}\left(c_v n_c p_{vc} + c_e n_c p_{ec} - c_v n_f p_{vf} - c_e n_f p_{ef}\right) \\
&\overset{n_c = n_f}{=} \frac{1}{c_v n_c}\left(c_v n_c p_{vc} + c_e n_c p_{ec} - c_v n_c p_{vf} - c_e n_c p_{ef}\right) \\
&= p_{vc} + \frac{c_e}{c_v} p_{ec} - p_{vf} - \frac{c_e}{c_v} p_{ef} \\
&= a_c - a_f.
\end{aligned} \tag{17}$$

E a variância de D fica

$$\begin{aligned}
\sigma_D^2 &= Var[D] \\
&= Var\left[\frac{Y_c - Y_f}{c_v n_c}\right] \\
&= \frac{Var\left[Y_c\right] + Var\left[Y_f\right] - 20Cov[Y_c, Y_f]}{(c_v n_c)^2} \\
&= \frac{Var\left[c_v X_{vc} + c_e X_{ec}\right] + Var\left[c_v X_{vf} + c_e X_{ef}\right]}{(c_v n_c)^2} \\
&= \frac{c_v^2 Var[X_{vc}] + c_e^2 Var[X_{ec}] + 2Cov[X_{vc}, X_{ec}] + c_v^2 Var[X_{vf}] + c_e^2 Var[X_{ef}] + 2Cov[X_{vf}, X_{ef}]}{(c_v n_c)^2}
\end{aligned}$$

$$= \frac{c_v^2 n_c p_{vc}(1-p_{vc}) + c_e^2 n_c p_{ec}(1-p_{ec}) + c_v^2 n_c p_{vf}(1-p_{vf}) + c_e^2 n_c p_{ef}(1-p_{ef})}{(c_v n_c)^2}$$

$$= \frac{1}{c_v n_c}\left[c_v p_{vc}(1-p_{vc}) + \frac{c_e^2}{c_v} p_{ec}(1-p_{ec}) + c_v p_{vf}(1-p_{vf}) + \frac{c_e^2}{c_v} p_{ef}(1-p_{ef})\right]. \quad (18)$$

Ressalta-se que a $Cov[X_{vc}, X_{ec}]$ é 0, pois assumiu-se que as variáveis Y_c e Y_v, são independentes, dado que a abordagem inicial utilizada para representar o modelo foi de que a distribuição dos resultados em cada $\mathbf{X}_c$ e $\mathbf{X}_f$ seguem, cada um, uma multinomial independente (2 e 3).

Para D, as mesmas aproximações (binomial e normais) podem ser estabelecidas. Primeiramente, a aproximação binomial é construída utilizando-se a esperança exata e a variância aproximada de maneira análoga. Reconhecemos que o número de ensaios (número máximo de sucessos) é $c_v n$, utilizamos a esperança exata (17) e aproximamos a variância (18), de tal forma que a exigência (14) da binomial seja satisfeita. Assim encontramos a probabilidade de sucesso

$$p_1^* = \frac{E[D]}{c_v n} = \frac{a_c - a_f}{c_v n} \quad (19)$$

e

$$Var[D]_1^* = c_v n \left(\frac{a_c - a_f}{c_v n}\right)\left(1 - \frac{a_c - a_f}{c_v n}\right)$$
$$= (a_c - a_f)\left(1 - \frac{a_c - a_f}{c_v n}\right). \quad (20)$$

No entanto, note que (19) precisa variar entre 0 e 1, o que não acontece quando $a_c < a_f$. Sendo assim, sugere-se a reparametrização a seguir para garantir essa condição:

$$p_2^* = 0{,}5 + c_v n_c p$$
$$= 0{,}5 + \frac{a_c - a_f}{2}$$
$$= \frac{1 + a_c - a_f}{2}.$$

Isso implica a seguinte mudança na variância (20):

$$Var[D]_2^* = c_v n p^*(1-p^*)$$
$$= c_v n \frac{1 + a_c - a_f}{2}\left(1 - \frac{1 + a_c - a_f}{2}\right)$$

$$= c_v n \left(\frac{1 + a_c - a_f}{2}\right)\left(\frac{1 - a_c + a_f}{2}\right)$$
$$= \frac{c_v n}{4}(1 + a_c - a_f)(1 - a_c + a_f).$$

Finalmente, podemos escrever a aproximação binomial, da seguinte maneira:

$$D \dot{\sim} Bin\left(c_v n,\ \frac{1 + a_c - a_f}{2}\right),$$

sendo que está será denominada como aproximação pela binomial para D.

3.1.4 Aproximação normal para a distribuição de D

Apresenta-se uma maneira de aproximar a distribuição de D pela normal. Sendo que nessa aproximação utiliza-se a esperança exata (17) e a variância exata (18). Dessa forma,

$$D \dot{\sim} N\left(a_c - a_f,\ \sigma_D^2\right).$$

Assim, esta será denominada de aproximação pela normal.

3.2 Resultados do estudo de simulação

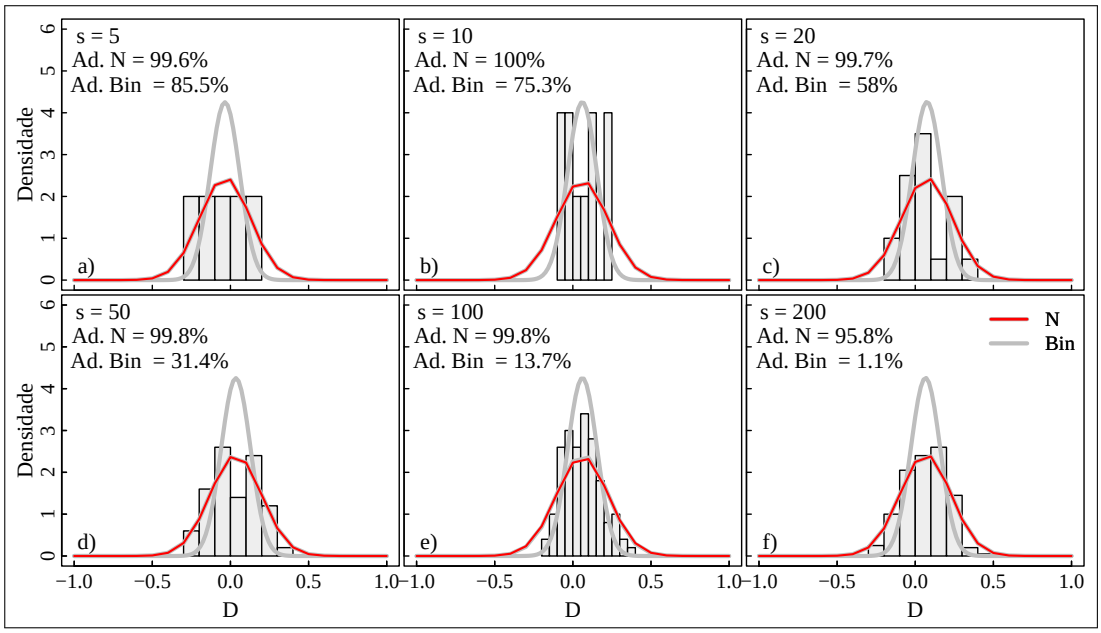
O primeiro resultado do estudo de simulação, confeccionado para permitir uma melhor visualização dos valores gerados e das distribuições aproximadas, consistiu em 6 histogramas com valores de D observados em uma única simulação e as respectivas curvas das distribuições aproximadas (Figura 1). Sendo que o primeiro gráfico (Figura 1a) corresponde ao tamanho amostral de $s = 5$, indo até $s = 200$ no sexto gráfico (Figura 1f), isto é, quando a amostra foi composta por 200 valores observados. Todos os 6 gráficos foram baseados em dados simulados a partir do mesmo parâmetro hipotético (item de número 1 da Tabela 1).

Ainda, para a obtenção dos valores de aderências em cada gráfico, foram utilizadas 1000 simulações, sendo que uma aderência de $85{,}5\%$ (Figura 1a) indicou que a $\mathcal{H}_0$ não foi rejeitada em $85{,}5\%$ das simulações, isto é, que em $85{,}5\%$ das simulações não existiram indícios suficientes para afirmar que as amostras de tamanho $s = 5$ não vieram da distribuição binomial. Na medida que o tamanho amostral foi aumentando, a aderência à aproximação binomial foi diminuindo até chegar em $1{,}1\%$ no tamanho amostral de $s = 200$ (Figura 1f). Este comportamento foi diferente do observado para a

aproximação normal, que, de uma maneira geral, manteve-se com valores de aderência próximos a 100% nos 6 tamanhos apresentados (Figura 1).

Destaca-se que em todos os gráficos da Figura 1, a aproximação binomial ficou mais alongada, enquanto que a aproximação normal ficou mais achatada, ou seja, a aproximação binomial tendeu a apresentar probabilidade nula para valores mais afastados da média, diferentemente da aproximação normal. Essa característica de maior achatamento da aproximação normal provavelmente foi o que fez com que ela consiga descrever melhor as frequências mais afastadas da média (extremidades) e, provavelmente essa foi uma importante característica que ajudou a explicar uma maior aderência da aproximação normal quando comparada com a binomial.

Figura 1 – Histograma dos valores simulados de D e as curvas das distribuições (linhas cinzas e vermelhas). Em que s é o tamanho de amostra utilizado, que foi diferente em cada gráfico e o termo "Ad." significa aderência, isto é, a porcentagem de iterações (de um total de 1000 iterações) em que os dados aderiram à distribuição para aquele tamanho de amostra. Todos os gráficos utilizaram o vetor de probabilidade $\mathbf{p}_c = (0{,}5; 0{,}3; 0{,}2)^\top$ e $\mathbf{p}_f = (0{,}4; 0{,}4; 0{,}2)^\top$



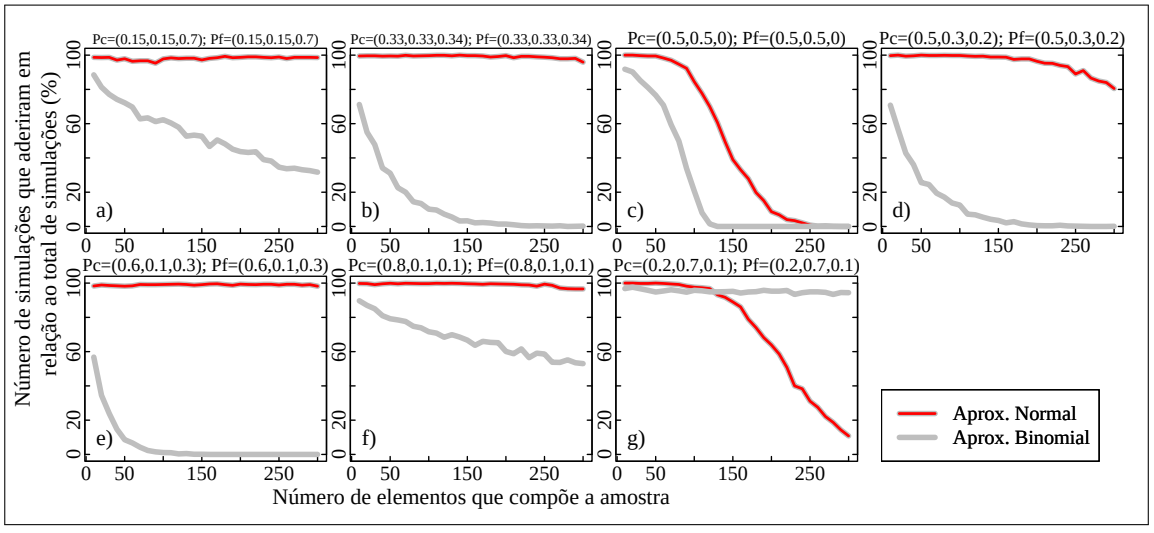
Fonte: autores (2023)

Enquanto que a Figura 1 trouxe resultados para uma única combinação de vetores $\mathbf{p}_c$ e $\mathbf{p}_f$, nas Figuras 2, 3 e 4, cada gráfico corresponde a uma combinação diferente dos $\mathbf{p}_c$ e $\mathbf{p}_f$. Ainda, as Figuras 2, 3 e 4 foram feitas para uma sequência de tamanhos amostrais s, ou seja, $s = 10, 20, ..., 300$. Enquanto que as Figuras 4 e 5 foram baseadas em vetores $\hat{\mathbf{p}}_c$ e $\hat{\mathbf{p}}_f$ estimados a partir de dados observados no Campeonato

Brasileiro, as Figuras 1, 2 e 3 foram baseadas em valores hipotéticos para vetores de parâmetros $\mathbf{p}_c$ e $\mathbf{p}_f$.

A Figura 2 se baseou em parâmetros hipotéticos com vetor $\mathbf{p}_c$ sempre igual ao vetor $\mathbf{p}_f$. Como resultado, em apenas uma combinação os valores de aderência da binomial foram maiores que a normal, que foi na Figura 2g e que também foi a única situação que $\mathbf{p}_e$ e foi maior que a probabilidade de $\mathbf{p}_v$. É possível observar uma aderência maior da normal em relação à binomial em 6 combinações sendo que em duas situações a aderência da normal caiu expressivamente na medida que o s foi aumentando: a primeira foi quando $\mathbf{p}_v$ e $\mathbf{p}_e$ foram $0{,}5$ e a segunda foi quando $\mathbf{p}_e$ foi maior que $\mathbf{p}_v$. Quando analisa-se apenas a aproximação pela binomial, existiu maior aderência quando ambos $\mathbf{p}_c$ e $\mathbf{p}_f$ foram pequenos ou quando $\mathbf{p}_v$ ou $\mathbf{p}_e$ foi alta. Pode-se pontuar aqui que a variável aleatória D é constituída da combinação linear de duas v.a. dependentes, subtraída de combinação linear (de outras duas v.a. binomiais dependentes) independentes entre si. Uma possibilidade é que, quando o valor de $p_{vc} = p_{vf}$ é próximo de 1 e o valor de $p_{ec} = p_{ef}$ é próximo de 0, D tenderá a uma distribuição binomial (Figura 2f), pois nesse caso D será constituído principalmente da diferença de duas v.a. binomiais independentes, que também é binomial.

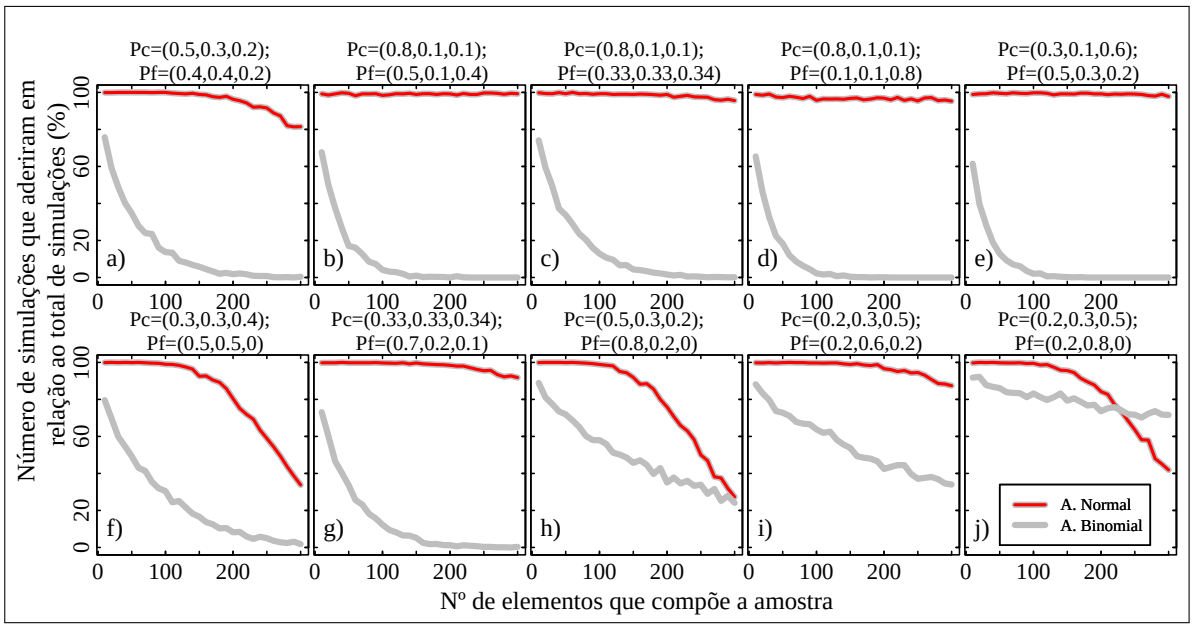
Figura 2 – Grau de aderência de dados gerados por vetores hipotéticos tais que $\mathbf{p}_c = \mathbf{p}_f$ em relação às duas aproximações (normal e binomial). Foram utilizadas 1000 iterações de geração aleatória de D em amostras de tamanho $s = 10, 20, 30, ..., 300$. Sendo que 0% representa nenhuma iteração com aderência e 100% representa que todas as iterações aderiram à distribuição. Ainda, os vetores $\mathbf{p}_c$ e $\mathbf{p}_f$ utilizados estão explicitados logo acima à area gráfica



Fonte: autores (2023)

Nas 10 situações apresentadas na Figura 3, foi possível observar que: (i) em 5 situações a aderência da aproximação normal começou a cair na medida que o tamanho de amostra é aumentado; (ii) excetuando-se o caso em que $p_{ec} < p_{ef}$ (Figura 3j), em todas as outras a normal apresentou maior aderência que a aproximação binomial e; (iii) a aproximação binomial teve um comportamento semelhante entre si e com baixa aderência nas Figuras 3a,b,c,d,e,f,g, porém, teve maior aderência quando a p_{ef} foi alta (Figura 3j com $p_{ef} = 0{,}8$ e Figura 3i, com $p_{ef} = 0{,}6$) ou quando a $p_{vf} = 0{,}8$ (Figura 3h).

Figura 3 – Grau de aderência de dados gerados por vetores hipotéticos com $\mathbf{p}_c \neq \mathbf{p}_f$ em relação à duas aproximações (normal e bionomial). Sendo que 0% representa nenhuma iteração com aderência e 100% representa que todas as iterações aderiram à distribuição) de 1000 iterações de geração aleatória de D em amostras de tamanho $s = 10, 20, 30, ..., 300$ em relação a duas distribuições aproximadas (binomial e normal) e com 3 pares de vetores $\mathbf{p}_c$ e $\mathbf{p}_f$ fixados, que estão explicitados logo acima da área gráfica de cada uma das figuras

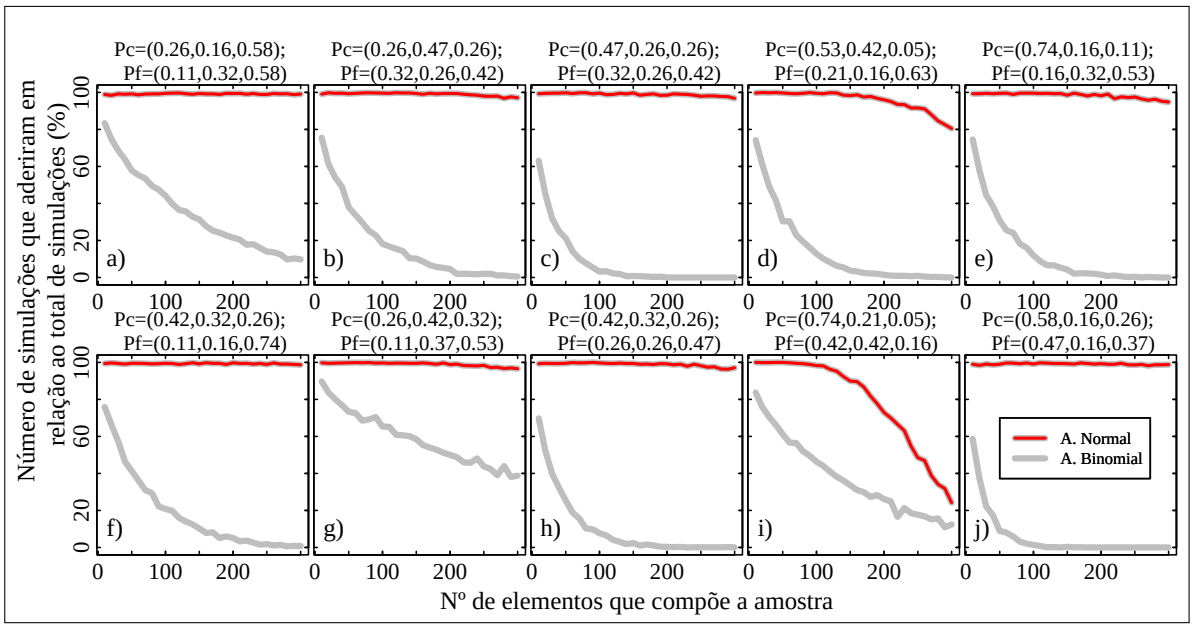


Fonte: autores (2023)

Já nas situações com parâmetros $\mathbf{p}_c$ e $\mathbf{p}_f$ obtidos no Campeonato Brasileiro (Figura 4), são evidenciados alguns pontos: (i) maior aderência da aproximação normal em relação a binomial em todos os tamanhos amostrais e em todas as combinações de parâmetros $\mathbf{p}_c$ e $\mathbf{p}_f$ considerados; (ii) comportamento variado da aderência das duas aproximações dependendo da combinação de parâmetros. Em dois gráficos (Figura 4g,i), a binomial teve valores de aderência mais próximos aos observados para a normal: um time que ganhou muito pontos e um time que conquistou poucos

pontos. Embora que a aproximação binomial teve melhores aderências nessas situações, a aderência da normal sempre apresentou valores superiores nas 10 combinações consideradas.

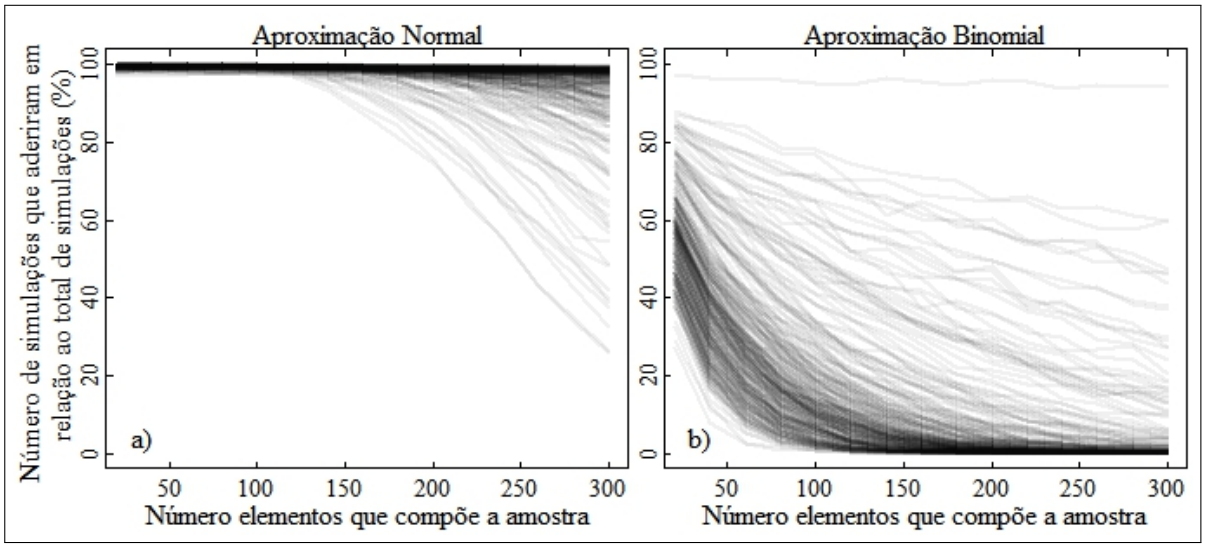
Figura 4 – Grau de aderência de dados gerados a partir de vetores de probabilidade obtidos no Campeonato Brasileiro de Futebol em relação à duas aproximações (normal e binomial). Sendo que 0% representa nenhuma iteração com aderência e 100% representa que todas as iterações aderiram à distribuição. Foram 1000 iterações de geração aleatória de D em amostras de tamanho $s = 10, 20, 30, ..., 300$ em relação a duas distribuições aproximadas (binomial e normal)



Fonte: autores (2023)

Quando todos os vetores de probabilidade obtidos em cada uma das 280 participações de um time no Campeonato Brasileiro foram colocados em um único gráfico, gerou 278 curvas que estão apresentadas com transparência na Figura 5 (duas curvas não foram possíveis de serem simuladas). Como a linha utilizada para cada uma das 278 curvas tem a mesma cor cinza e com transparência, então os regiões da área gráfica com a tonalidade mais escura representa os locais com maior sobreposição de curvas. Com base nesse resultado é possível observar uma aderência maior da normal em relação à binomial. Sendo que em parte das combinações a normal teve aderência reduzida a partir de tamanhos de amostra de 150. Como pode ser observado, valores maiores de aderência da binomial são exceções, já que a maioria das curvas da binomial iniciam com cerca de 50% de aderência e ficam com menos de 20% de aderência em tamanhos amostrais maiores que $s = 100$.

Figura 5 – As 278 curvas de aderência com dados gerados a partir de parâmetros reais obtidos em competição e comparadas em relação à aproximação normal e a aproximação binomial. Isto é, cada curva cinza foi construída com uma combinação de parâmetros que aconteceu para um time em uma participação no Campeonato Brasileiro de Futebol de 2006 a 2019. Foram realizadas 1000 iterações de geração aleatória de D em amostras de tamanho $s = 20, 40, 60, ..., 300$



Fonte: autores (2023)

Quando os valores da v.a. D observados foram comparados com um teste de aderência em relação às duas aproximações, a distribuição normal teve maiores valores de aderência em todas as combinações de $\mathbf{p}_c$ e $\mathbf{p}_f$ analisadas. Sendo que ao todo foram simuladas 18 situações hipotéticas e 10 situações com vetores de probabilidade observados no Campeonato Brasileiro em gráficos isolados e 278 situações do Campeonato Brasileiro em dois gráficos, um para a aproximação normal (Figura 5a) e outro para a binomial (Figura 5b), de maneira que fossem representadas as diferentes situações que aconteceram.

De uma maneira geral, têm-se os seguintes resultados: (i) a aproximação binomial teve melhor aderência quando o valores gerados de D foram obtidos por $\mathbf{p}_c$ e $\mathbf{p}_f$ igualmente pequenos, quando a $p_{vc} = p_{vf}$ era próximo de 1 ou quando a $\mathbf{p}_e > \mathbf{p}_v$; (ii) a aproximação binomial só teve aderência maior que a aproximação normal em algumas classes de tamanho de amostra quando $\mathbf{p}_e > \mathbf{p}_v$, isto é, a aproximação normal teve maior aderência na grande maioria de situações analisadas e; (iii) a aderência da aproximação normal sempre foi próxima a 100% em tamanhos amostrais de até s=100.

Assim, a partir dos resultados do estudo de simulação pôde-se concluir que a aproximação normal melhor se aderiu aos dados. Desse modo, assumiu-se que os dados seguem aproximadamente uma distribuição normal parametrizada com média e variância dadas por
$\mu = a_c - a_f$ e $\sigma^2 = \frac{1}{c_v n_c}[c_v p_{vc}(1-p_{vc}) + \frac{c_e^2}{c_v} p_{ec}(1-p_{ec}) + c_v p_{vf}(1-p_{vf}) + \frac{c_e^2}{c_v} p_{ef}(1-p_{ef})]$.
Em uma amostra, pode-se então utilizar a média amostral ($\bar{d}$) e a variância amostral (S^2) para inferências.

3.3 Resultados das aplicações

São apresentados os resultados de 3 aplicações possíveis a partir da contribuição do presente estudo, isto é, uma vez que D é aproximadamente normal, podem ser utilizados os testes bem estabelecidos, como por exemplo, o teste t para médias. A aplicação 1 consistiu em um exemplo da utilização do teste de hipóteses para um par de times que disputam a mesma competição, que no caso foi o Internacional e o Grêmio em todas as participações no Campeonato Brasileiro de Futebol Série A entre 2006 a 2020. Onde estabeleceu-se:

$\mathcal{H}_0 : \mu_{int} = \mu_{gre}$

$\mathcal{H}_1 : \mu_{int} \neq \mu_{gre}$

onde μ_{int} é a média populacional da v.a. D para o Internacional e μ_{gre} é a média populacional para o Grêmio.

O primeiro passo foi a obtenção dos d_i valores da variável aleatória D para as i participações do time em uma competição, isto é:

$d_{int} = \{0{,}053;\ 0{,}386;\ 0{,}561;\ 0{,}263;\ 0{,}175;\ 0{,}211;\ 0{,}175;\ 0{,}140;\ 0{,}368;\ 0{,}526;\ 0{,}368;\ 0{,}404;\ 0{,}368\}$
$d_{gre} = \{0{,}386;\ 0{,}351;\ 0{,}684;\ 0{,}193;\ 0{,}316;\ 0{,}263;\ 0{,}263;\ 0{,}333;\ 0{,}386;\ 0{,}368;\ 0{,}105;\ 0{,}281;\ 0{,}228;\ 0{,}228\}$

onde d_{gre} e d_{int} são os conjuntos com os valores observados de d. Estes dois conjuntos geraram as médias amostrais iguais a $\bar{d}_{int} = 0{,}3077;\ S_{int} = 0{,}15221$ e $\bar{d}_{gre} = 0{,}3133;\ S_{gre} = 0{,}1334$.

Inicialmente precisamos verificar se as variâncias são homogêneas através do teste F ($F = 0{,}1334^2/0{,}15221^2 = 0.7682$). Como o valor tabelado da estatística $F_{13,12,\alpha=0,05} = 3{,}15$ é maior do que o valor calculado, não rejeita-se a hipótese nula de que as variâncias são iguais. Assim, utilizou-se a estatística t para variâncias

homogêneas (Zar, 2010). Inicialmente, obteve-se a variância ponderada $s_p^2 = (SS_1 + SS_2)/(\nu_1 + \nu_2) = (0{,}27800 + 0{,}23137)/(12 + 13) = 0{,}020374$, em que SS_1 e SS_2 são as somas de quadrados com ν_1 e ν_2 graus de liberdade, do Internacional e do Grêmio, respectivamente. Assim, obteve-se o t calculado, $t = (\bar{D}_1 - \bar{D}_2)/\sqrt{s_p^2/n_1 + s_p^2/n_2} = (0{,}3077 - 0{,}3133)/(\sqrt{(}0{,}020375/13 + 0{,}020375/14) = -0{,}1016924$. Para um nível de significância $\alpha = 0{,}05$ e graus de liberdade $gl = n_1 + n_2 - 2 = 25$, encontram-se os valores de $t_{critico} = \pm 2{,}05954$. Portanto, não rejeita-se a hipótese nula de que não há diferença entre os times. Além disso, é importante ressaltar a interpretação que pode ser obtida. O valor da média amostral obtida para o Internacional, isto é $\bar{d}_{int} = 0{,}3077$ e do Grêmio $\bar{d}_{gre} = 0{,}3133$, representa que, no período analisado, o Internacional ganhou $30{,}77\%$ dos pontos a mais em casa do que fora da casa. Já o Grêmio ganhou 31,33% dos pontos a mais em casa do que fora. Se este valor é de 100% representa que o time ganha todos os pontos em casa e nenhum fora. Se o valor é de -100% significa que ganha todos os pontos fora de casa e nenhum em casa. A v.a. D tem a escala de variação da mesma forma que a função apresentada como "método de Pollard reescalado" (Matos et al., 2020).

A aplicação 2 baseou-se na comparação entre quatro ligas principais de quatro países. Comparou-se se havia diferenças entre as ligas: Série A brasileira, *Premier League* inglesa, *Serie A* italiana e *La Liga* espanhola. A média amostral da liga brasileira foi de $\bar{d} = 0{,}2680$, da inglesa $\bar{d} = 0{,}1746$, da italiana $\bar{d} = 0{,}1763$ e espanhola foi $\bar{d} = 0{,}1957$. A média da liga brasileira foi diferente das demais (Tabela 3). Graficamente também é possível observar que a Série A brasileira, com exceção do ano de 2017, sempre apresentou valores de vantagem de casa superiores às outras ligas (Figura 6). Novamente, ressalta-se que o teste de comparação de médias utilizado baseou-se em um nível de significância de 5% para cada par de ligas, sendo que o nível de significância global ficou superior a 5%, questão desconsiderada no presente estudo. No presente estudo, a primeira divisão do Campeonato Brasileiro apresentou uma média de efeito de casa superior às outras três ligas que foram comparadas. Um padrão semelhante foi observado em Silva & Moreira (2008), que estudaram ligas nacionais e encontraram que a liga brasileira e a francesa tiveram valores significativamente iguais, mas a liga brasileira apresentou valores superiores às ligas da Itália, Espanha, Inglatera, Portugal, Alemanha e Argentina. É possível observar

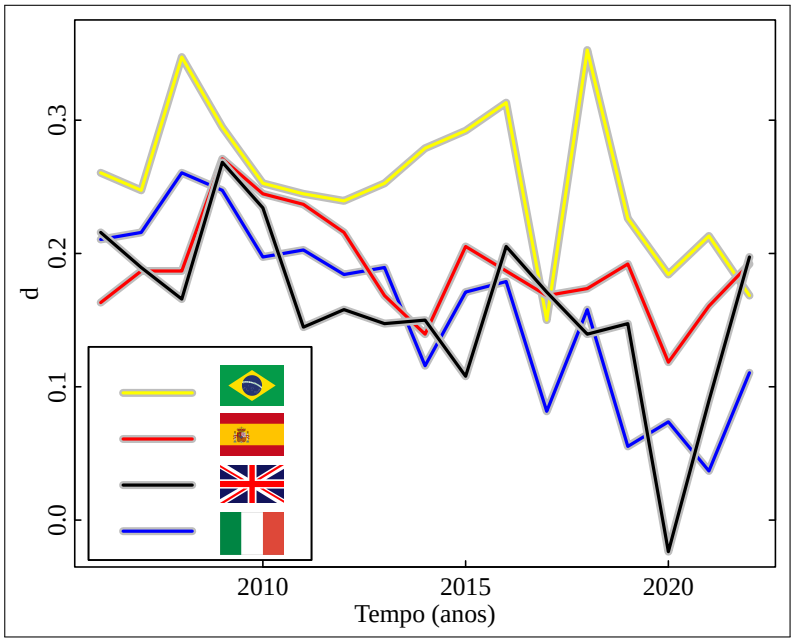
também uma redução expressiva no efeito de casa no ano da pandemia do Coronavírus, especialmente para a *Premier League*, seguido de uma retomada no ano de 2023.

Tabela 3 – Comparação entre as primeiras divisões de quatro ligas de quatro diferentes países para todas as participações dos times de 2006 a 2022 para o Campeonato Brasileiro e de 2006/2007 à 2022/2023 para as 3 ligas Européias (*La Liga* (Espanha), *Premier League* (Inglaterra) e *Serie A* (Itália))

	Média de D	
Série A (Brasil)	0,254	a
La Liga (Espanha)	0,189	b
Premier League (Inglaterra)	0,159	c
Serie A (Itália)	0,158	c

Fontes: autores (2023)

Figura 6 – Média amostral ($\bar{d}$) por ano para cada uma das ligas: Série A (Brasil; linha amarela); *La Liga* (Espanha; linha vermelha); *Premier League* (Inglaterra; linha preta) e; *Serie A* (Itália; linha azul). Foram utilizados dados das edições do Campeonato Brasileiro de 2006 à 2022 e das três ligas europeias de 2006/2007 à 2022/2023



Fonte: autores (2023)

E por último, a aplicação 3, mostra que pode-se obter valores de vantagem de casa para competições desbalanceadas, isto é, as competições em que um time não joga o mesmo número de partidas em casa e fora de casa. Por exemplo, o River Plate no Campeonato Argentino de futebol de 2018/2019 jogou 25 partidas, sendo que destas,

13 foram e casa (7 vitórias, 2 empates e 4 derrotas) e 12 fora (6 vitórias, 4 empates e 2 derrotas). A partir destes valores pode-se obter os dois vetores de probabilidade $\mathbf{p}_c = (7/13; 2/13; 4/13)^T$ e $\mathbf{p}_f = (6/12; 4/12; 2/12)^T$. Em seguida, calcula-se, $a_c = p_{vc} + \frac{c_e}{c_v} p_{ec} = 7/13 + \frac{1}{3}\frac{2}{13} = 0{,}589744$ e $a_f = p_{vf} + \frac{c_e}{c_v} p_{ef} = 6/12 + \frac{1}{3}\frac{4}{12} = 0{,}611111$. Desta maneira, pode-se obter uma média amostral de d, isto é, $\bar{d} = a_c - a_f = 0{,}589744 - 0{,}611111 = -0{,}021367$. Então, neste caso o River Plate conquistou mais pontos fora de casa, uma vez que teve um valor de d negativo, de 2,1%. Ressalta-se que para este caso, os adversários em casa e fora de casa são diferentes, o que pode exigir cuidados adicionais e que não foram o foco do presente estudo. Ressalta-se que este resultado é uma novidade do presente estudo, uma vez que é possível estimar o valor de d a partir das estimativas de p_{vc}, p_{ec}, p_{vf} e p_{ef}. No estudo anterior (Paludo et al., 2023), não era possível obter valores de d para campeonatos desbalanceados.

Com base nos resultados do presente estudo é possível construir outras aplicações utilizando-se resultados prontos para a distribuição normal, como exemplos os intervalos de confiança utilizando a distribuição t para as participações de um time em uma competição como exemplificado em Paludo et al. (2023). No mesmo estudo, também foi apresentado uma alternativa de teste bootstrap para verificar se os resultados obtidos por um time eram suficientes para afirmar sobre a existência de efeito de casa em cada edição do campeonato que o time participou. A partir dos resultados do presente artigo, será possível a utilização de testes já estabelecidos para a normal.

4 CONSIDERAÇÕES FINAIS

Este estudo serve de modelo para trabalhos que desenvolvem variáveis aleatórias e procuram descobrir a sua distribuição. Sendo que o conhecimento da sua distribuição permite a realização de inferências de maneira facilitada. Caso a distribuição da variável aleatória seja conhecida, pode-se utilizar testes já estabelecidos e conhecidos para tal varivável aleatória.

Com base no presente estudo foi possível aproximar uma distribuição de probabilidade para D, o que possibilita a construção de inferências utilizando-se resultados que são bem estabelecidos para a distribuição normal, como exemplo os teste de hipóteses e a construção de intervalos de confiança utilizando a distribuição t.

Além disso, como é possível estimar os vetores de probabilidade de vitória em casa e fora com base nos resultados das partidas de uma participação na competição, fica possível calcular d também para campeonatos desbalanceados, aspecto que não era possível de ser calculado antes do presente estudo.

REFERÊNCIAS

Alzubaidi, M., Hasan, K. N., & Meegahapola, L. (2022). Impact of Probabilistic Modelling of Wind Speed on Power System Voltage Profile and Voltage Stability Analysis. *Electric Power Systems Research*, 206:107807.

Benz, L. S. & Lopez, M. J. (2021). Estimating the change in soccer's home advantage during the Covid-19 pandemic using bivariate Poisson regression. *AStA Advances in Statistical Analysis*, pages 1–28.

Bolfarine, H. & Sandoval, M. C. (2001). *Introdução à Inferência Estatística*, volume 2. SBM.

Butler, K. & Stephens, M. A. (2017). The distribution of a sum of independent binomial random variables. *Methodology and Computing in Applied Probability*, 19(2):557–571.

Curley, J. (2020). *engsoccerdata: English and European Soccer Results 1871-2020*. R package version 0.1.7.

Dawson, P., Massey, P., & Downward, P. (2020). Television match officials, referees, and home advantage: Evidence from the European Rugby cup. *Sport Management Review*, 23(3):443–454.

Goller, D. & Krumer, A. (2020). Let's meet as usual: Do games played on non-frequent days differ? Evidence from top European soccer leagues. *European Journal of Operational Research*, 286(2):740–754.

Hegarty, T. (2021). Information and price efficiency in the absence of home crowd advantage. *Applied Economics Letters*, pages 1–6.

Marek, P. & Vávra, F. (2020). Comparison of Home Advantage in European Football Leagues. *Risks*, 8(3):87.

Matos, R. M., Amaro, N., & Pollard, R. (2020). How best to quantify home advantage in team sports: an investigation involving male senior handball leagues in Portugal and Spain. *RICYDE. Revista Internacional de Ciencias del Deporte.*, 16(59):12–23.

McCarrick, D., Bilalic, M., Neave, N., & Wolfson, S. (2021). Home advantage during the COVID-19 pandemic: Analyses of european football leagues. *Psychology of Sport and Exercise*, 56:102013.

Mood, A. M., Graybill, F. A., & Boes, D. C. (1974). *Introduction to the Theory of Statistics 1974*. McGraw-Hill Kogakusha.

Nevill, A. M. & Holder, R. L. (1999). Home Advantage in Sport: An Overview of Studies on the Advantage of Playing at Home. *Sports Medicine*, 28(4):221–236.

Paludo, G. F., Figueiredo, N. N., & Ferreira, E. B. (2023). Proposta de uma métrica para a vantagem de casa baseada em pontos ganhos. *Ciência e Natura*, 45(e3):01–29.

Pollard, R., Prieto, J., & Gómez, M.-Á. (2017). Global differences in home advantage by country, sport and sex. *International Journal of Performance Analysis in Sport*, 17(4):586–599.

Pollard, R., Silva, C. D., & Medeiros, N. C. (2008). Home advantage in football in Brazil: differences between teams and the effects of distance traveled. *Brazilian Journal of Soccer Science*, 1(1):3–10.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Silva, C. D., Abad, C. C. C., Macedo, P. A. P., Fortes, G. O. I., & Nascimento, W. W. G. (2018). Competitive balance in football: A comparative study between Brazil and the main European leagues (2003-2016). *Journal of Physical Education*, 29.

Silva, C. D. & Moreira, D. G. (2008). A vantagem em casa no futebol: comparação entre o campeonato Brasileiro e as principais ligas nacionais do Mundo. *Revista Brasileira de Cineantropometria Desempenho Humano*, 10(2):184–188.

Usman, M., Zubair, M., Shiblee, M., Rodrigues, P., & Jaffar, S. (2018). Probabilistic Modeling of Speech in Spectral Domain using Maximum Likelihood Estimation. *Symmetry*, 10(12):750.

Van-Ours, J. C. (2019). A Note on Artificial Pitches and Home Advantage in Dutch Professional Football. *De Economist*, 167(1):89–103.

Vellaisamy, P. & Punnen, A. P. (2001). On the Nature of the Binomial Distribution. *Journal of applied probability*, 38(1):36–44.

Zar, J. H. (2010). *Biostatistical Analysis.* Pearson.

Contribuições dos autores

1 – Giovani Festa Paludo
Mestrando em Estatística Aplicada
https://orcid.org/0000-0002-8046-8409 • gfpaludo@gmail.com
Contribuição: Conceptualization; Data curation; Formal Analysis; Investigation; Methodology; Visualization; Writing - original draft

2 – Eric Batista Ferreira
Doutorado em Estatística e Experimentação Agropecuária
https://orcid.org/0000-0003-3361-0908 • eric.ferreira@unifal-mg.edu.br
Contribuição: Conceptualization; Project Administration; Supervision; Validation; Writing - review & Editing

Como citar este artigo

Paludo, G. F., & Ferreira, E. B. (2025). Modelagem probabilística e inferência do efeito de casa em partidas esportivas. *Ciência e Natura*, Santa Maria, v. 47, e85273. DOI: https://doi.org/10.5902/2179460X85273.