UNIVERSIDADE FEDERAL DE ALFENAS

GIOVANI FESTA PALUDO

ESTATÍSTICA APLICADA À VANTAGEM DE CASA NO FUTEBOL

ALFENAS/MG

2022

GIOVANI FESTA PALUDO

ESTATÍSTICA APLICADA À VANTAGEM DE CASA NO FUTEBOL

Dissertação apresentada como parte dos requisitos para obtenção do título de Mestre em Estatística Aplicada e Biometria, pela Universidade Federal de Alfenas. Área de Concentração: Estatística Aplicada e Biometria.
Orientador: Prof. Dr. Eric Batista Ferreira

ALFENAS/MG

2022

Sistema de Bibliotecas da Universidade Federal de Alfenas
Biblioteca Central

Paludo, Giovani Festa.
Estatística Aplicada à Vantagem de Casa no Futebol / Giovani Festa Paludo. - Alfenas, MG, 2022.
86 f. : il. -

Orientador(a): Eric Batista Ferreira.
Dissertação (Mestrado em Estatística Aplicada e Biometria) - Universidade Federal de Alfenas, Alfenas, MG, 2022.
Bibliografia.

1. Estatística aplicada. 2. Vantagem de casa. 3. Inferência. 4. Futebol. 5. Distribuição de probabilidade. I. Ferreira, Eric Batista, orient. II. Título.

Ficha gerada automaticamente com dados fornecidos pelo autor.

GIOVANI FESTA PALUDO

Estatística Aplicada à Vantagem de Casa no Futebol

A Banca examinadora abaixo-assinada aprova a Dissertação apresentada como parte dos requisitos para a obtenção do título de Mestre em Estatística Aplicada e Biometria pela Universidade Federal de Alfenas. Área de concentração: Estatística Aplicada e Biometria.

Aprovada em: 25 de fevereiro de 2022.

Prof. Dr. Eric Batista Ferreira
Instituição: Universidade Federal de Alfenas - UNIFAL-MG

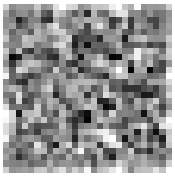
Profa. Dra. Josiane Magalhães Teixeira
Instituição: Universidade Federal do Vale do Jequitinhonha e Mucuri - UFVJM

Prof. Dr. Person Pereira Neves
Instituição: Universidade Federal de Alfenas - UNIFAL-MG

Documento assinado eletronicamente por **Eric Batista Ferreira**, **Presidente**, em 04/03/2022, às 15:19, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Josiane Magalhães Teixeira Ribeiro**, **Usuário Externo**, em 04/03/2022, às 17:13, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Person Pereira Neves**, **Professor do Magistério Superior**, em 04/03/2022, às 17:24, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.



A autenticidade deste documento pode ser conferida no site https://sei.unifal-mg.edu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **0683148** e o código CRC **ED085387**.

"*Lembre-se do seu fim e pare de odiar*."
(BÍBLIA, Eclesiástico 28:6)

AGRADECIMENTOS

Agradeço sobretudo à Deus pelo dom da vida.

Agradeço ao Professor Eric, pela orientação, confiança, ajuda, amizade, ensinamentos e por todo apoio prestado sempre! Nunca foi só uma profissão. Quem tem compromisso em ensinar e em ajudar as pessoas a crescerem sempre, são pessoas excelentes e que fazem toda a diferença para a sociedade! E aproveito e estendo esse meu agradecimento aos meus eternos orientadores: Professora Rose e Professores Mantovani e Maurício. São pessoas que não medem esforços e se dedicam continuamente na formação técnica e científica dos estudantes! E também incluo todos os professores do Programa de Estatística Aplicada da Unifal. Em especial o Professor Beijo, a quem sempre podemos contar, e que sempre ajudou muito frente à tantas coisas e mudanças que aconteceram nesses tempos. Meus agradecimentos também aos Professores Fabricio, Denismar, Patrícia, Natália, Juliana e Ricardo pelas excelentes aulas e os momentos de aprendizado que propiciaram. Agradeço a secretária da Pós Martha pela agilidade e presteza. E acredito que o suporte de todo o Programa em especial ao Professor Beijo e Professor Eric, foram fundamentais para a conclusão do meu curso. Eu agradeço a todo o Programa e Universidade, pois foi aqui que pude retomar na pós-graduação e na pesquisa, que sempre gostei muito. Sou muito grato por existir esse Programa sendo que Mestre em Estatística Aplicada e Biometria é uma enorme conquista para mim. Foi com muito esforço e com muita alegria que chego aqui.

Agradeço aos membros da banca por terem contribuído e ajudado bastante: Professora Josiane, Professor Person, Professor Beijo e Professor Denismar!

Agradeço ao meu amor Amandinha pela compreensão, carinho, companheirismo, empenho, paciência e tudo que tem feito por mim sempre! Sou muito grato à Deus por ter te encontrado nesse mundo!

Agradeço a minha família pelo apoio sempre: ao meu pai Zelindo, mãe Suzana, Elias, Manu, Carol, Junior e Arthur! Sou grato por ter vocês em minha vida. Peço perdão pelas minhas falhas e sempre quero e procuro o melhor para todos vocês!

Agradeço ao Nikolas pela amizade, parceria, ajuda e confiança! Que nossos estudos sobre o futebol rendam muitos frutos!

Agradeço à todos os Filhos do Barba. Acredito que esse encontro virtual foi muito importante e muito especial na pandemia. E obrigado à todos os colegas que sempre colaboraram com os estudos do futebol, e não vou citar nomes, mas agradeço a cada um! Obrigado mesmo!

Agradeço à todos aqueles que ajudaram. Em especial o Professor Cristiano Silva pela ajuda, disposição e entusiasmo!

Sou grato pelos meus colegas da minha turma da pós: Gabriel, Walef, Rafaela, Daiana e Marcos! Mesmo a distância, vocês foram fundamentais e considero-os muito!

Agradeço também ao Seu Gerson pelo apoio e acolhimento em mais de um momento: no começo e nessa etapa final do meu mestrado. Agradeço tmb ao André, Adriely, Amanda, Rômulo e Junior pelo acolhimento em São Paulo. Obrigado!

Agradeço aos meus amigos que sempre pude contar! Vai aqui uma saudação pros Pikapaus do Asfalto, Igor, Pame, Poli, Lipe, Max, Richard, Benini. Também agradeço aos colorados do “Se enrolado para o futebol"pelos numerosos exemplos de baixa qualidade técnica no futebol. E ao Matteus, Washington e Guilherme de Alfenas pela amizade e ajuda nessa caminhada! Agradeço Karin, João, João, Rony, André, Matteus, Rafael e todos pela harmoniosa convivência nesse período! Obrigado mesmo!

Essa foi uma grande conquista para mim, que teve bastante esforço e muitas pessoas ajudaram. Se eu não citei você aqui, peço desculpas e saiba que mesmo assim sou eternamente grato!

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

RESUMO

Os fatores que fazem com que o Campeonato Brasileiro tenha mais vantagem de casa do que, por exemplo, as ligas europeias, ainda não estão completamente elucidados. Para poder explicar melhor os fatores associados à vantagem de casa no Brasil, primeiro seria necessário um método para medir o efeito de casa que permitisse obter uma informação para cada participação de um time na competição e que fosse por pontos. Um dos principais estudos que obtém o efeito de casa por pontos no Campeonato Brasileiro, fez uma correção pela habilidade e não trouxe os valores por ano que um time participou por questões metodológicas. Assim, o nosso objetivo foi desenvolver uma forma de obter o efeito de casa por pontos, que possibilite trazer uma informação para cada time em cada participação e que não necessite uma correção pela habilidade. Para atingir esse objetivo foram desenvolvidos dois estudos, um no Capítulo 4 e outro no Capítulo 5. O Capítulo 4 teve como objetivo propor uma métrica para medir o efeito de casa no futebol baseada em pontos. O ponto inicial do Capítulo 4 foi partir de uma métrica bem conhecida para o efeito de casa e a partir de certas modificações propor uma nova métrica. Em seguida, utilizou-se o Campeonato Brasileiro de Futebol Série A de 2003 a 2020 como aplicação. Como resultado, foi obtida a nova métrica denominada de d, que atingiu os pontos que haviam sido listados. Ainda foi desenvolvido um teste sobre d que permitiu verificar quando há informação suficiente para afirmar se existe efeito de casa, tirando o efeito da oscilação quando o time ganha poucos pontos. Com base nisso, das 370 participações de 43 times em 18 edições, uma apresentou desvantagem de casa, 259 tiveram vantagem de casa e 110 não apresentaram efeito de casa. Além disso, d permitiu visualizar o comportamento individual de cada time ao longo das edições da competição. Já o Capítulo 5 tratou a métrica d como uma variável aleatória (v.a.) D procurando descrever sua distribuição e características através de uma distribuição de probabilidade. Foram trazidas algumas características importantes como média e variância, e, na impossibilidade de obter a distribuição exata, foram estudadas aproximações. Para avaliar e decidir qual distribuição melhor aproximava a distribuição de D, foi feito um estudo de simulação para verificar a melhor aproximação. Como resultado do estudo de simulação, a aproximação pela distribuição normal foi aquela que apresentou maior aderência e foi considerada para as aplicações. Para ilustrar a metodologia desenvolvida, foram feitas cinco aplicações utilizando a aproximação pela normal. Sendo que o conhecimento da distribuição possibilita a inferência estatística utilizando a variável aleatória D e permite um série de aplicações com testes bastante conhecidos. Assim, foi proposto uma variável aleatória para

medir o efeito de casa e estudou-se sua distribuição, que possibilita a realização de diversas inferências.

Palavras-chave: Vantagem de casa; efeito de casa; diferença de pontos relativa; desvantagem de casa; esportes coletivos; futebol; campeonato brasileiro; estudo de simulação; aproximação de uma distribuição.

ABSTRACT

The Brazilian *Série A* usually has higher home advantage than, for example, the European leagues and the factors that explain this pattern remain not completely cleared. To provide a better comprehension of those factors, we believe that the first issue is to obtain the home effect metric that permit observe the year to year variation of each club and that is based in points. One of the most important study in this subject obtained an average home advantage for the seasons that was corrected by ability and the study didn't bring the information of each season by methodological reasons. Then, our objective is to develop a metric to obtain the home effect based on points, that does not need an ability correction and that bring an information for a team in every season. To achieve this objective, we conducted two studies, one in Chapter 4 and other in the Chapter 5. Chapter 4 has the objective of propose a metric to obtain the home advantage based in points. This Chapter started transforming a well-known metric to a new metric. So, we have used data from *Série A* of *Campeonato Brasileiro* from 2003 to 2020 as a study application. As a result, the new metric was named d and it was obtained accomplishing the points we had expected to. As a solution, it was developed a test for d to verify if there is enough information to affirm if there is a home effect avoiding random effects as when a team win few points. Based on this test, we have obtained from 43 teams in 370 participations that: 259 had positive home effect, 1 negative and 110 had no effect. Moreover, d has permitted to visualize graphically the team's individual behavior though the years in the competition. Chapter 5 treated d as a random variable (v.a.) D and we described its distribution and some important characteristics as population mean and variance. Besides, as it was not possible to obtain the exact distribution of D, we obtained 2 approximations of the distribution of D: a binomial and a normal one. To evaluate and to decide which approximation was the best, we conducted a simulation study. As the main result, D was well approximatted by the normal distribution and so we used it in applications with real data. Knowing the proper distribution permit to conduct statistical inferences over the v.a. D and has permitted some applications using well known tests. So, as a general conclusion we have developed a v.a. D to measure the home effect and we have studied its distribution, which is approximately normal and permit to build inferences.

Keywords: home ground effect; home cocking; relative difference of points; home disadvantage; sports; round-robin tournament; distribution approximation.

SUMÁRIO

LISTA DE SÍMBOLOS

$\perp$	Indica independência entre eventos, exemplo: A $\perp$ B, isto é, A e B são eventos independentes.
a_c	Novo parâmetro para mandante;
a_f	Novo parâmetro para visitante;
B	Conjunto dos b_i;
b_i	Pontos que um time conquistou na i-ésima partida;
Bin	Distribuição binomial;
$_c$	Como mandante;
c_v	Constante que indica os pontos que um time conquista por vitória;
c_e	Constante que indica os pontos que um time conquista por empate;
c_d	Constante que indica os pontos que um time conquista por derrota;
d	Função para medir o efeito de casa proposta pelo presente estudo, nomeada de *diferença relativa de pontos*;
D	Variável aleatória chamada de diferença de pontos relativa para medir o efeito de casa (no estudo anterior consistia na d);
$\bar{D}$	Média amostral de valores de d;
da	Função alternativa para medir o efeito de casa proposta pelo presente estudo, nomeada de *diferença absoluta de pontos*;
$\bar{DA}$	Média amostral de valores de da;
$_f$	Como visitante;
gl	Graus de liberdade;
gl'	Graus de liberdade aproximados;
h	Função para medir o efeito de casa do estudo de Pollard, Silva e Medeiros (2008);
$\bar{H}$	Média amostral de valores de H;
h_0	Função para medir o efeito de casa do estudo de Pollard, Silva e Medeiros (2008) com escala modificada;
$\bar{H}_0$	Média amostral de valores de h_0;
$\mathscr{H}_0$	Hipótese nula;
$\mathscr{H}_0$	Hipótese alternativa;
IP_{LS}	Limite superior do intervalo de predição;
IP_{LI}	Limite inferior do intervalo de predição;
n_c	Número de partidas como mandante em uma competição;
m	Número de edições da competição que um time disputou;

m_+	Número de edições da competição que um time disputou e obteve efeito de casa significativo;
$Multi$	Distribuição multinomial;
N	Distribuição normal;
n	Número de partidas disputadas em um campeonato;
$n(W)$	Função $n(\cdot)$ que conta o número de vezes que um evento W aconteceu;
n_c	Número de partidas disputadas como mandante em um campeonato;
n_f	Número de partidas disputadas como visitante em um campeonato;
Ω	Espaço amostral;
$\mathbf{p}_c$	Vetor de probabilidades em casa, $\mathbf{p}_c = (p_{vc},p_{ec},p_{dc})^\top$;
$\mathbf{p}_d$	Vetor de probabilidades de derrotas em casa e fora, $\mathbf{p}_d = (p_{dc},p_{df})^\top$;
p_{dc}	Probabilidade de um time perder como mandante;
p_{df}	Probabilidade de um time perder como visitante;
$\mathbf{p}_e$	Vetor de probabilidades de empates em casa e fora, $\mathbf{p}_e = (p_{ec},p_{ef})^\top$;
p_{ec}	Probabilidade de um time empatar como mandante;
p_{ef}	Probabilidade de um time empatar como visitante;
$\mathbf{p}_f$	Vetor de probabilidades fora de casa, $\mathbf{p}_f = (p_{vf},p_{ef},p_{df})^\top$;
$\mathbf{p}_v$	Vetor de probabilidades de vitórias em casa e fora, $\mathbf{p}_v = (p_{vc},p_{vf})^\top$;
p_{vc}	Probabilidade de um time vencer como mandante;
p_{vf}	Probabilidade de um time vencer como visitante;
R	Conjunto com os mesmos elementos de B, porém com ordem aleatorizada;
s	Tamanho amostral;
S	Desvio padrão amostral;
S^2	Variância amostral;
$S^2_{\bar{D}}$	Variância amostral da média;
$_v$	Como visitante;
v	Pontos por vitória;
$v.a.$	Variável aleatória ou variáveis aleatórias;
$\mathbf{X}$	Vetor composto pelo número de vitórias, empates e derrotas que um time obtém como mandante e como visitante ao final do campeonato;
$\mathbf{X}^\top$	Transposto de $\mathbf{X}$.
$\mathbf{X}_c$	Vetor número de vitórias, empates e derrotas em partidas como mandante;
$\mathbf{X}_f$	Vetor número de vitórias, empates e derrotas em partidas como visitante;
X_{vc}	v.a. número de vitórias como mandante ao fim do campeonato;
X_{ec}	v.a. número de empates como mandante ao fim do campeonato;

X_{dc}	v.a. número de derrotas como mandante ao fim do campeonato;
X_{vf}	v.a. número de vitórias como visitante ao fim do campeonato;
X_{ef}	v.a. número de empates como visitante ao fim do campeonato;
X_{df}	v.a. número de derrotas como visitante ao fim do campeonato;
Y_c	Pontos conquistados como mandante ao final de uma edição da competição;
Y_v	Pontos conquistados como visitante ao final de uma edição da competição.

1 INTRODUÇÃO GERAL

O futebol é um esporte com expressiva importância social e cultural. Além de ser uma atividade de entretenimento amplamente difundida, o futebol gera um grande número de empregos diretos e indiretos, sendo que é bastante comum encontrarmos altos valores de patrocínios, premiações, salários e outros. Por exemplo, para a temporada de 2019/2020 o Manchester United recebeu cerca de 90 milhões de dólares para a marca que patrocina a camiseta do clube. Nessa mesma linha, temos a premiação da União das Federações Europeias de Futebol (UEFA), que distribuiu 1,98 bilhões de euros em premiações aos times que participaram da Champions League na temporada de 2018/2019. E quando observa-se o Brasil, o futebol no Brasil movimentou em 2018 um valor próximo a 0,7% do Produto Interno Bruto brasileiro.

A questão chave por trás de expressivos volumes financeiros é a audiência: o futebol é um dos esportes com maior audiência em nível global. Por exemplo, a Copa do Mundo de futebol da Rússia teve 2,65 bilhões de pessoas que assistiram mais de 20 minutos, enquanto que os jogos Jogos Olímpicos de Verão do Rio de Janeiro tiveram audiência 2,6 bilhões de pessoas que assistiram mais de 15 minutos. Toda essa audiência faz com que o futebol seja uma excelente vitrine para divulgação de marcas, mercadorias e produtos.

Embora não seja completamente estabelecido o porquê de tanta audiência no futebol, existem alguns fatores que podem estar associados a essa expressiva audiência no futebol. Um dos fatores é a baixa necessidade de recursos para a prática desse esporte. Outra possível razão apontada são os baixos placares quando comparado à outros esportes, sendo que esse fator pode provocar surpresa e choque em quem assiste, já que as vezes um time mais fraco acaba ganhando de um time forte, ou as vezes a partida é decidida com um gol no último minuto. E somado à isso, estudos mostram que o futebol é um dos esportes mais competitivos. E assim, a competitividade pode ser um dos importantes fatores levando a expressiva audiência no futebol. Se um esporte é muito previsível, os fãs poderão perder o interesse, pois se no início do campeonato, já se sabe quem vai ser o campeão, então o campeonato vai ficando menos interessante. É por isso que, ao longo das décadas, são realizadas mudanças em regras e regulamentos com intuito de melhorá-lo e torná-lo mais competitivo. Se é mais competitivo, os fãs ficarão mais interessados e isso vai gerar maiores retornos financeiros. Tem-se que o futebol é muito imprevisível de uma maneira geral. Um dos fatores que é sabido que acaba afetando um pouco a imprevisibilidade, é a vantagem de casa.

A vantagem de casa (do inglês *home advantage* e abreviada aqui como HA) diz respeito

ao quanto um time tem um desempenho melhor em casa do que fora de casa. Sendo que HA é um fenômeno e pode ser definido como um resultado consistente em que o time da casa vence mais de 50% dos jogos em um calendário que há a mesma quantidade de jogos em casa e fora de casa para cada time. Sendo que no presente texto, vantagem de casa será sinônimo de *efeito de casa positivo*, e será dado preferência ao termo *efeito de casa*, uma vez que o efeito de casa é um termo mais geral que inclui tanto o efeito positivo de casa (vantagem de casa) quanto o efeito negativo de casa (desvantagem de casa). Quanto à obtenção de um valor de efeito de casa para um time, há duas variáveis que são comumente utilizadas separadamente para criar métricas. Uma delas é a diferença de gols da partida (ou saldo de gols como é comumente chamado) e a outra variável são os pontos conquistados. Enquanto que o saldo de gols figura como importante para estudar a vantagem de casa em competições que são mata-mata em duas partidas (na casa do primeiro e na casa do segundo time), os pontos parecem fazer mais sentido para estudar a vantagem de casa em competições no formato de pontos corridos em que todos os times se enfrentam uma ou duas vezes (uma na casa do primeiro e outro na casa do segundo).

Quando se pretende utilizar pontos para medir o efeito de casa, há uma métrica muito bem conhecida, que consiste na razão entre os pontos conquistados em casa e o total de pontos conquistados por um time em uma competição (que chamaremos aqui de h). Ao observar a métrica h é possível notar que há um divisão pelo total de pontos, e se um time ganha poucos pontos, os pontos de casa serão divididos por um número menor e o valor de h tende a ficar inflacionado. Este é um exemplo de uma fragilidade desta métrica e há outras fragilidades que serão apresentadas na Capítulo 4. A ideia central da primeira parte da dissertação é o desenvolvimento de uma nova métrica baseada em h e que mitigue o efeito das fragilidades que serão discutidas na sequência. Por fim, será conduzido um estudo para o conhecimento da distribuição dessa métrica, o que vai facilitar a inferência sobre parâmetros de interesse.

2 OBJETIVO

2.1 OBJETIVO GERAL

Estudar o efeito de casa no futebol, desde a construção racional de uma métrica até a inferência estatística sobre os parâmetros de interesse.

2.2 OBJETIVOS ESPECÍFICOS

Os objetivos específicos do presente estudo podem ser escritos como:

a) Propor uma métrica otimizada para o efeito de casa em partidas de futebol;

b) Estimar por ponto e intervalo o efeito de casa populacional média para cada time que participou do Campeonato Brasileiro de Futebol de 2003 a 2020;

c) Modelar a métrica como variável aleatória e compreender a a estrutura da população e parâmetros de interesse;

d) Obter a distribuição de probabilidades dessa variável aleatória;

e) Inferir sobre os parâmetros de interesse, a partir da distribuição obtida;

f) Ilustrar a teoria com dados reais do Campeonato Brasileiro de Futebol, e ligas europeias.

Os itens de números *a)* e *b)* serão abordados no Capítulo 4 e os itens de números *c)*, *d)*, *e)* e *f)* serão abordados no Capítulo 5.

3 REFERENCIAL TEÓRICO

Esta seção traz uma revisão de literatura sobre conceitos fundamentais da estatística que são utilizados ao longo do trabalho para desenvolvimento do estudo. Inicialmente são apresentadas definições sobre conceitos de probabilidade, em seguida são apresentadas brevemente as distribuições de probabilidade mencionadas no Capítulo 5. Também são apresentados alguns fundamentos sobre o efeito de casa, bem como algumas diferenças entre alguns métodos conhecidos e uma breve passagem de alguns estudos que trabalharam o efeito de casa no futebol brasileiro. Para finalizar, há uma descrição de abordagens possíveis para a modelagem do problema que está sendo estudado. Dessa forma, o leitor poderá ficar confortável e compreender melhor os estudos apresentados nos Capítulos 4 e 5.

3.1 FUNDAMENTOS DE ESTATÍSTICA

Algumas das principais definições necessárias para o desenvolvimento da métrica proposta para medir o efeito de casa, sob a perspectiva da estatística, foram incluídas na sequência. Nesta seção podem ser encontradas definições como a do espaço amostral, variável aleatória, distribuição de probabilidade, e fechando em uma construção sobre maneiras de representar os resultados de partidas de futebol de um campeonato do formato da Série A do Campeonato Brasileiro de Futebol.

3.1.1 Principais definições

A primeira definição que será mostrada é a de espaço amostral. Segundo Mood, Graybill e Boes (1974),

Definição *(espaço amostral):* *o espaço amostral é denotado por Ω e é a coleção ou totalidade de todas as saídas possíveis de um experimento conceitual*.

A título de exemplificação, considere um time de futebol que joga uma partida. O time pode obter uma vitória (V), empate (E) ou derrota (D). O espaço amostral para o experimento aleatório *resultado em uma partida* pode ser estabelecido como

$$\Omega = \{V, E, D\}.$$

A segunda definição necessária é a definição da σ-algebra. Embora que Mood, Graybill e Boes (1974), em seu livro texto, definem a função de probabilidade em uma álgebra, a definição da função de probabilidade em uma σ-algebra é mais comum, sendo que há uma indicação no próprio livro do Mood, Graybill e Boes (1974) para que isso seja feito. Assim, segundo Resnick (2005), temos que

Definição ***(σ-álgebra):*** *$\mathscr{B}$ é uma classe de subconjuntos de Ω não vazia e fechada sobre uniões contáveis, intersecções contáveis e complementos. Sendo que há alguns postulados para que $\mathscr{B}$ seja σ-algebra*:

1. $\Omega \in \mathscr{B}$.
2. $B \in \mathscr{B}$ implica que $B^C \in \mathscr{B}$.
3. $B_i \in \mathscr{B}, i \geq 1$ implica $\cup_{i=1}^{\infty} B_i \in \mathscr{B}$,

onde: $\mathscr{B}$ é a σ-álgebra; B é um evento qualquer pertencente à σ-álgebra; B^C é o evento complementar do evento B e; i é o índice que enumera os eventos pertencentes a $\mathscr{B}$, isto é B_1, B_2, ..., B_∞.

Para o exemplo de Ω dado acima, podemos obter a seguinte σ-álgebra $\mathscr{A}$, sendo que sabe-se que cada elemento da σ-álgebra é um evento:

$$\mathscr{A} = \sigma(\Omega) = \sigma(\{V, E, D\}) = \{\emptyset, \{V\}, \{E\}, \{D\}, \{VE\}, \{VD\}, \{ED\}, \Omega\}.$$

Com o Ω e a σ-álgebra definidas, podemos definir o espaço de probabilidade. Segundo Resnick (2005), um espaço de probabilidade é um triplo $(\Omega, \mathscr{A}, P)$. Onde Ω é um espaço amostral correspondendo às saídas de algum experimento (sendo que pode ser hipotético). Já $\mathscr{A}$ é a σ-álgebra de subconjuntos de Ω e estes subconjuntos recebem o nome de eventos. E ainda, temos a medida de probabilidade P, que é uma função com domínio $\mathscr{A}$ e imagem $[0,1]$, tal que

1. $P(A) \geq 0$ para todo $A \in \mathscr{A}$.
2. P é σ-aditivo: Se $A_n, n \geq 1$ são eventos em $\mathscr{A}$ que são disjuntos, então

$$P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n).$$

3. $P(\Omega) = 1$.

Uma vez que entende-se o espaço de probabilidade, pode-se definir variável aleatória (MOOD; GRAYBILL; BOES, 1974):

Definição ***(variável aleatória (v.a.):*** *para um dado espaço de probabilidade ($\Omega, \mathscr{A}, P[\cdot]$), uma v.a., denotada por X ou $X(\cdot)$, é uma função com domínio Ω e contradomínio a reta real. A função $X(\cdot)$ deve ser tal que o conjunto A_r definido por $A_r = \{\omega : X(\omega) \leq r\}$, pertence a $\mathscr{A}$ para todo número real r.*

Com isto, é possível estabelecer que a variável aleatória (v.a.) é uma função que tem como saída valores da reta real. Quando o Ω já tem como elementos os valores da reta real, são estes os números que poderão virar realizações da variável aleatória. Quando os elementos do Ω não são números, a v.a. associa números aos seus elementos. Como exemplo, se tomamos a v.a. número de vitórias em uma partida do espaço amostral $\Omega = \{V, E, D\}$, associaremos o valor 1 para o evento V e 0 para os eventos E e D. Isto é, a v.a. L, poderá assumir os seguintes valores:

$$L = \{0,1\}$$

Sobre v.a., é possível encontrar em alguns livros, a distinção entre v.a. discreta e contínua:

Definição (variável aleatória discreta): *uma variável aleatória X será definida como discreta se a série de valores de X é contável. Se uma v.a. é discreta, então sua correspondente função de distribuição acumulada $F_X(\cdot)$ será definda como discreta (MOOD; GRAYBILL; BOES, 1974).*

Assim, como exemplo de v.a. discreta temos a v.a. L, definida anteriormente, e que é o número de vitórias em uma partida.

Definição (variável aleatória contínua): *uma variável aleatória X é chamada de contínua se existe uma função $f_X(\cdot)$ tal que $F_X(x) = \int_{-\infty}^{x} f_X(u)du$ para cada número real x. A função de distribuição acumulada $F_X(\cdot)$ de uma v.a. contínua X é chamada absolutamente contínua (MOOD; GRAYBILL; BOES, 1974).*

O exemplo de v.a. contínua seria o "tempo necessário até que 1 gol seja marcado por um time". E, ainda, tem-se outros dois conceitos intimamente ligados ao conceito da v.a., é o conceito de amostra aleatória e vetor aleatório.

Definição (amostra aleatória): *seja as v.a. $X_1, X_2, ..., X_n$ com uma densidade conjunta*

$f_{X_1,X_2,...,X_n}$ *que fatora como segue:*

$$f_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) = f(x_1)f(x_2) \cdot \cdot \cdot \cdot \cdot f(x_n),$$

onde $f(\cdot)$ *é a densidade (comum) de cada* X_i*. Então* $X_1, X_2, ..., X_n$ *é definida para ser uma a.a. de tamanho* n *de uma população com densidade* $f(\cdot)$ (MOOD; GRAYBILL; BOES, 1974).

Ao que pode-se observar à respeito da definição acima, a amostra aleatória é composta por v.a. cuja distribuição conjunta é o produtório das distribuições marginais. Tal características, satisfaz também a condição de independência entre as v.a.. Sendo que é possível mencionar ainda que os parâmetros das distribuições marginais são os mesmos para todas as variáveis aleatórias, por isso diz-se que na amostra aleatória, as v.a. são identicamente distribuídas.

Já o *vetor aleatório* difere de amostra aleatória uma vez que no vetor aleatório não tem essa restrição sobre as marginais, pois pode ser definido como:

Definição (vetor aleatório discreto): *Para* $\boldsymbol{x}$ *ser um vetor aleatório discreto, a função de probabilidade conjunta é definida da seguinte forma:*

$$p(\mathbf{x}) = p(x_1, x_2, ..., x_m) = P(X_1 = x_1, X_2 = x_2, ..., X_m = x_m),$$

onde: **x** é um vetor aleatória discreto; $x_1, x_2, ..., x_m$ são cada uma das v.a. desse vetor aleatório; $p(\mathbf{x})$ é uma função de probabilidade do vetor **x** e; $P(X_1 = x_1, X_2 = x_2, ..., X_m = x_m)$ é a probabilidade de **X**.

A função de probabilidade marginal de $X_k, k = 1, 2, ..., m$ *é dada por:*

$$p_{X_k}(x_k) = P(X_k = x_k) = \sum_{x_i \forall i \neq k} p(\mathbf{x}) = \sum_{x_i \forall i \neq k} P(X_1 = x_1, ..., X_m = x_m);$$

com a soma para os valores possíveis em todas as coordenadas, exceto k *(MAGALHãES, 2011).*

Outra definição importante para a construção da presente dissertação é a definição de estatística, que, como a própria definição explicita, uma estatística é também uma v.a. e para ela estamos interessados em associar uma distribuição de probabilidades (que recebe o nome especial de distribuição de amostragem) e, a partir daí, fazer inferência sobre seus parâmetros.

Definição (estatística): *uma estatística é uma função observável de v.a., que é ela própria uma v.a. observável, que não contém nenhum parâmetro desconhecido (MOOD; GRAYBILL;*

BOES, 1974).

Ainda, pode ser considerado que nós temos a intenção de fazer inferência sobre a densidade da v.a., pois, caso a v.a. não for observável, ela não terá uso em se fazer inferências (MOOD; GRAYBILL; BOES, 1974). Como exemplificado pelos próprios autores, podemos dizer que se Y é uma v.a. e δ um parâmetro, $Y + \delta$ não é uma estatística, porém, $Y + 1$ seria uma estatística. Sendo que encontrar uma estatística adequada para representar o parâmetro populacional, é um dos problemas centrais da estatística (MOOD; GRAYBILL; BOES, 1974).

Definição (estimador): *Qualquer estatística (função conhecida de v.a. observáveis que é ela própria uma v.a.) cujos valores são usados para estimar o $\tau(\theta)$, onde $\tau(\cdot)$ é alguma função do parâmetro θ, é definido para ser um estimador de $\tau(\theta)$* (MOOD; GRAYBILL; BOES, 1974).

Cabe ressaltar que o estimador sempre é uma estatística e também é uma função e uma v.a. (MOOD; GRAYBILL; BOES, 1974). No presento estudo, a função $\bar{D}$ que será definida e apresentada no Capítulo 4, é um exemplo de um estimador utilizado.

3.1.2 Testes de aderência

Os testes de aderência, são métodos para examinar como uma amostra de dados aceita uma dada distribuição como sua população (D'AGOSTINO; STEPHENS, 1986). De uma maneira geral, a hipótese nula $\mathscr{H}_0$ é que uma dada v.a. X segue uma função de probabilidade acumulada $F_X(x)$, como exemplo, a distribuição normal. O teste de aderência mede a conformidade dos valores amostrados para a distribuição de interesse ou a discrepância em relação è ela (D'AGOSTINO; STEPHENS, 1986).

Uma classe de testes de aderência são os chamados *testes do tipo qui-quadrado*, que de uma maneira geral tinha a ideia baseando em um teste uma comparação de contagens observadas com os valores esperados de acordo com a hipótese a ser testada (MOORE, 1986). Como há uma redução na informação, então estes testes de qui-quadrado tendem a ser menos poderosos que outras classes de testes de aderência, porém, os testes podem ser aplicados a dados discretos ou contínuos, univariados ou multivariados, em suma, eles são os mais comumente aplicados testes de aderência (MORE, 1986).

3.2 ALGUMAS DISTRIBUIÇÕES DE PROBABILIDADE

Na presente dissertação foram utilizadas três distribuições: a distribuição binomial (Bin), multinomial (Multi) e normal (N).

3.2.1 Distribuição Multinomial

Inicialmente iremos considerar a distribuição multinomial. Conforme a descrição de Johnson, Kotz e Balakrishnan (1997), tem-se uma série de n ensaios independentes e em cada ensaio nós observamos um número k de eventos que são mutuamente excludentes, isto é, que não podem ocorrer simultaneamente. Ainda, temos que a probabilidade de ocorrência do evento E em qualquer ensaio é igual a **p** (com $p_1 + p_2 + ... + p_k = 1$). Ainda, defina a sequência de v.a. $Y_1, Y_2, ..., Y_k$ como o número de ocorrência dos eventos $E_1, E_2, ..., E_k$, respectivamente, nesses n ensaios, sendo o $\sum_{i=1}^{k} Y_i = n$. Assim, a distribuição conjunta de $Y_1, Y_2, ..., Y_k$ é dada por (JOHNSON; KOTZ; BALAKRISHNAN, 1997):

$$Pr[\bigcap_{i=1} k(Y_i = n_i)] = n! \prod_{i=1}^{k} (\frac{p_i^{n_i}}{n_i!}) = P(n_1,n_2, ..., n_k), \text{ com } n_i \geq 0, \sum n_i = n. \tag{3.1}$$

Na multinomial, temos que a esperança, variância e covariância são dadas, respectivamente por,

$$E[Y_i] = np_i, \tag{3.2}$$

$$var[Y_i] = np_i(1 - p_i), \tag{3.3}$$

$$cov[Y_i] = -np_ip_j. \tag{3.4}$$

Outro resultado importante é que o estimador de máxima verossimilhança de cada elemento do vetor de parâmetros $p_1, p_2, ..., p_k$ é a frequência relativa , ou seja,

$$\hat{p} = N_i/n, i = 1, 2, ..., k.$$

Sendo que $\hat{p}$ é o estimador de máxima verossimilhança; N_i é o número de vezes que o evento N aparece e n é a soma dos N_i de $i = 1, 2, ..., k$.

3.2.2 Distribuição Binomial

Sobre a distribuição binomial, temos que se X é uma v.a. que segue uma distribuição binomial, isto é, $X \sim Bin(n, p)$, então

$$Pr[X = x] = \binom{n}{x} p^x q^{n-x}, x = 0,1,2, ..., n,$$

onde $q + p = 1, p > 0, q > 0$ e n é um inteiro positivo, n é o número de ensaios e p é a probabilidade de que um evento irá ocorrer (JOHNSON; KEMP; KOTZ, 2005). Cabe ressaltar que quando $n = 1$, a distribuição é conhecida como distribuição de Bernoulli. A média na distribuição binomial é dada por $\mu = np$ e variância é dada por $\sigma^2 = npq$ (JOHNSON; KEMP; KOTZ, 2005).

A aproximação normal para distribuição binomial pode ser escrita como:

$$P[\alpha < (X - np)(npq)^{-1/2} < \beta] \approx \frac{1}{\sqrt{2\pi}} \int_{\alpha}^{\beta} e^{-u^2/2} du = \phi(\beta) - \phi(\alpha)$$

sendo que essa é uma aproximação cru, mas é útil quando o n é grande (JOHNSON; KEMP; KOTZ, 2005). Ainda segundo o mesmo autor, uma melhoria na aproximação é obtida com uma correção de continuidade.

Quando utiliza-se a aproximação normal para a distribuição binomial, há duas regras de ouro: (i) utilizar quando $np(1 - p) > 9$ e; (ii) utilizar quando $np > 9$ para $0 < p \leq 0.5q$. Em estudo de Schader e Schmid (1989), foi mostrado que o erro da aproximação normal aumenta enquanto o menor o valor do p, sendo que o erro é menor quando o $p = 1/2$. Sendo que esse fato pode ser esperado, pois, segundo Johnson, Kemp e Kotz (2005), a distribuição binomial é simétrica quando se tem $p = 1/2$.

Quando temos X_i, $i = 1,2,...$, como v.a. binomiais independentes com parâmetros (n_i,p), então $\sum_i X_i$ também tem uma distribuição binomial com parâmetros ($\sum_i n_i, p$), sendo que esta é a propriedade reprodutiva da distribuição binomial (JOHNSON; KEMP; KOTZ, 2005). Já quando se tem duas v.a. binomiais independentes, diga-se X_1 e X_2, então se definirmos X como $X = X_1 + X_2$, pode-se obter uma distribuição exata com forma fechada na referência supracitada.

Convém trazer o seguinte resultado sobre a situação em que temos uma constante multiplicada por uma v.a. binomial, questão pertinente para o Capítulo 5.

Uma v.a. binomial multiplicada por uma constante

A seguir é apresentado um resultado importante para o desenvolvimento do Capítulo V, que é a multiplicação de uma v.a. binomial por uma constate.

Seja $X \sim \text{Bin}(n,p)$. Considere:

$$X \sim \text{Bin}(n,p) \Rightarrow X \sim \text{N}(\mu = np, \sigma^2 = npq)$$

Seja agora $Y = \frac{1}{k}X$, em que k é constante e $k \in \mathbb{N}$.

Então $Y = \frac{X}{k}$ não é mais binomial. Afinal:

$$E[Y] = E\left[\frac{X}{k}\right] = \frac{1}{k}E[X] = \frac{np}{k}$$

$$Var[Y] = Var\left[\frac{X}{k}\right] = \frac{1}{k^2}Var[X] = \frac{1}{k^2}npq = \frac{npq}{k^2}$$

Na binomial, $Var[X] = npq = E[X](1-p)$. Se, $p^* = \frac{p}{k}$, então $E[Y] = np^* \Rightarrow q^* = 1 - p^* = 1 - \frac{p}{k}$.

Mas,

$$\begin{aligned} Var[Y] &= \frac{npq}{k^2} = np^*(1-p^*) = np^*\frac{(1-p)}{k} \\ &= np^*\left(\frac{1}{k} - \frac{p}{k}\right) = np^*\left(\frac{1}{k} - p^*\right) \neq np^*(1-p^*) \end{aligned}$$

3.2.3 Distribuição normal

Diz-se que uma v.a. X segue uma distribuição normal, isto é, $X \sim N(\xi, \sigma^2)$ se a sua função de densidade de probabilidade é dada por (JOHNSON; KOTZ; BALAKRISHNAN, 1994):

$$\frac{1}{\sqrt{2\pi}}exp\left[-\frac{1}{2}\left(\frac{x-\xi}{\sigma}\right)^2\right], \sigma > 0.$$

Antes dos trabalhos de Laplace e Gauss espalharem a importância teórica da normal, a normal foi referenciada como uma aproximação conveniente da distribuição binomial (JOHNSON; KOTZ; BALAKRISHNAN, 1994). Segundo os mesmos autores, na teoria da probabilidade ela tem um posição única, já que pode ser utilizada como aproximações de outras distribuições. Em termos práticos, se pode aplicar a distribuição normal com um pequeno risco

de sérios erros, quando distribuições não normais correspondem mais aos valores observados. Sendo que os argumentos teóricos para a normal ser utilizada são baseados nos teoremas centrais do limite. E o teorema diz que a soma de v.a. padronizadas, sendo ou não normais, tende a uma distribuição normal padronizada na medida que o número de v.a. na soma aumenta (JOHNSON; KOTZ; BALAKRISHNAN, 1994). Ou seja, em condições para garantir a distribuição normal padronizada assintótica.

Ainda, pode-se mencionar o estimador da variância populacional quando a média populacional (ξ) não é conhecida (JOHNSON; KOTZ; BALAKRISHNAN, 1994):

$$S_n = \left[n^{-1} \sum_{j=1}^{n} (X_j - \bar{X}^2 \right]^{1/2},$$

sendo que S_n é o estimador do desvio padrão populacional.

3.3 O EFEITO DE CASA EM ESPORTES E NO FUTEBOL

A primeira questão importante seria definir o que é a vantagem de casa (HA). Basicamente, podemos utilizar a definição trazida por Courneya e Carron (1992), onde HA pode ser definida como "o resultado consistente no qual o time de casa em competições esportivas vence mais que 50% dos jogos em um calendário de jogos balanceado entre casa e fora de casa". Quando utiliza-se uma v.a. para trazer informações sobre a vantagem de casa, não necessariamente o calendário precisa ser balanceado (por balanceado entende-se o campeonato com o mesmo número de jogos em casa e fora de casa). Então, com base nisso poderiámos reescrever a definição acima como "HA é o resultado no qual o time da casa vence mais que 50% dos jogos em casa do que fora". A definição não fala sobre empates, sendo que há esportes com e há esportes sem o empate. Então, uma maneira pela qual se lida com os empates é com base na criação de uma nova variável, que é chamada de "pontos", que condensa a informação de vitórias, empates e derrotas. Ainda, muitos esportes utilizam diferentes formas de definir quem vence a partida. No vôlei é quem ganha mais sets, no futebol pode-se dizer que é o saldo de gols (com um saldo positivo um time vence, com saldo negativo perde e com saldo igual a zero, empata), no basquete é quem obtém maior soma de pontuação em cestas, etc. Então, pode-se pensar em três principais variáveis para se obter a vantagem de casa:

a) o *número de vitórias e derrotas* (exemplo em: Stefani (2007); ou utiliza-se para esportes que não há empates, ou desconsideram-se os empates);

b) a *soma dos número de pontos atribuído à vitória, empate e derrota* (exemplo em Pollard, Silva e Medeiros (2008));

c) a mesma variável utilizada para definir o vencedor de uma partida (exemplo em (MAREK; VÁVRA, 2020)).

Quando o esporte que se estuda é o futebol, desconsiderar os empates até pode ser uma alternativa interessante, porém não foi considerada na presente dissertação. Já o saldo de gols na partida, que é a variável utilizada para definir o vencedor, pode não ser tão interessante. Considera-se aqui que as competições mais tradicionais dentro de um país, são as ligas nacionais de futebol, e geralmente nessas ligas todos os times se enfrentam uma ou duas vezes e o vencedor é aquele que conquista mais pontos ao final. Neste caso, o saldo de gols passa a ter um papel secundário nesse tipo de competição, sendo que é um dos critérios de desempate. Um time pode ganhar uma partida de 5 a 0 e perder 4 partidas de 1 a 0, que o saldo de gols fica positivo. Um time pode ganhar 4 partidas de 1 a 0 e perde uma partida de 5 a 0, que o saldo de gols fica negativo. Assim, a utilização de pontos como variável para medir a vantagem de casa figurou como a alternativa mais atrativa para o futebol. Dos três itens enumerados acima, os mais frequentemente utilizados são os pontos (POLLARD; SILVA; MEDEIROS, 2008; FAJARDO et al., 2019; LEITE, 2017; OLIVEIRA et al., 2020) e o saldo de gols (CLARKE; NORMAN, 1995; MAREK; VÁVRA, 2020). Pode-se mencionar também um estudo que obtém a HA pelo saldo de gols, porém utiliza os pontos para algum tipo de correção (GOUMAS, 2017). Os estudos citados nas duas frases acima são apenas exemplos, sendo que são numerosos os estudos sobre vantagem de casa.

No Campeonato Brasileiro de Futebol, Pollard, Silva e Medeiros (2008) utilizaram pontos para obtenção da HA, realizaram uma correção de acordo com habilidade e trouxeram valores de vantagem de casa médio para os times com mais participações no período de 2003 a 2007. Almeida, Oliveira e Silva (2011), também utilizando pontos, porém sem correção e trouxeram a vantagem de casa média para as séries A e B da competição, não apresentando valores para cada time. Há também estudo que utilizou as edições de 2018 a 2021, que trouxe valores médios por ano (RIBEIRO et al., 2022). Assim, o comportamento da vantagem de casa para cada time em cada ano é algo que não foi pesquisado de forma detalhada durante toda a sequência de anos desde que o Campeonato Brasileiro passou a ser disputado no formato de pontos corridos. Ressalta-se também que ainda não existem metodologias aplicadas ao Campeonato Brasileiro que permitiram observar o comportamento da vantagem de casa por pontos e com

correção para cada time em cada participação, por exemplo a metodologia de Pollard, Silva e Medeiros (2008) faz a correção pelo habilidade, porém nesse caso, não é possível observar cada valor de cada ano para um time. Ainda, também não há metodologias no Campeonato Brasileiro que verifiquem se uma certa configuração que aconteceu com um time, isto é, se o número de vitórias, empates e derrotas em casa e fora, consiste em informação suficiente para afirmar que existiu vantagem de casa para um time considerado. Por isso, na presente dissertação optou-se preferencialmente pelo termo efeito de casa, uma vez que esse termo inclui tanto a vantagem de casa (efeito de casa positivo), quanto a desvantagem de casa (efeito de casa negativo) ou a ausência de efeito de casa (sem vantagem e sem desvantagem de casa).

Estas foram algumas considerações necessárias ao entendimento do estudo desenvolvido. Sendo que na próxima seção foram trazidas algumas abordagens possíveis para representação do problema da obtenção da vantagem de casa.

3.4 PARTIDAS DE FUTEBOL EM CASA E FORA DE CASA

Ao considerar uma partida de futebol como um ensaio, nós podemos ter 3 resultados possíveis: vitória, empate ou derrota. Esse ensaio não pode ser considerado um ensaio de Bernoulli, contudo pode ser considerado um ensaio multinomial com $n = 1$.

$$\text{Resultado(partida)} = [\text{vitória}, \text{empate}, \text{derrota}]$$

Quando se deseja estabelecer todos os resultados possíveis considerando um time em um campeonato equilibrado (entende-se por equilibrado o campeonato com o mesmo número de partidas em casa e fora de casa) de 38 partidas, nós temos o seguinte número de arranjos ou resultados possíveis:

$$3^{38} = 1.350852 \times 10^{18} \text{ (resultados diferentes possíveis).}$$

Uma definição importante para ser trazida à discussão é sobre a dependência de eventos. Segundo Resnick (2005) a seguinte definição diz respeito a independência de dois eventos:

Definição (eventos independentes): *Suponha (Ω, $\mathscr{B}$, $\mathscr{P}$) é um espaço de probabilidade fixado. Eventos $A, B \in \mathscr{B}$ são independentes se:*

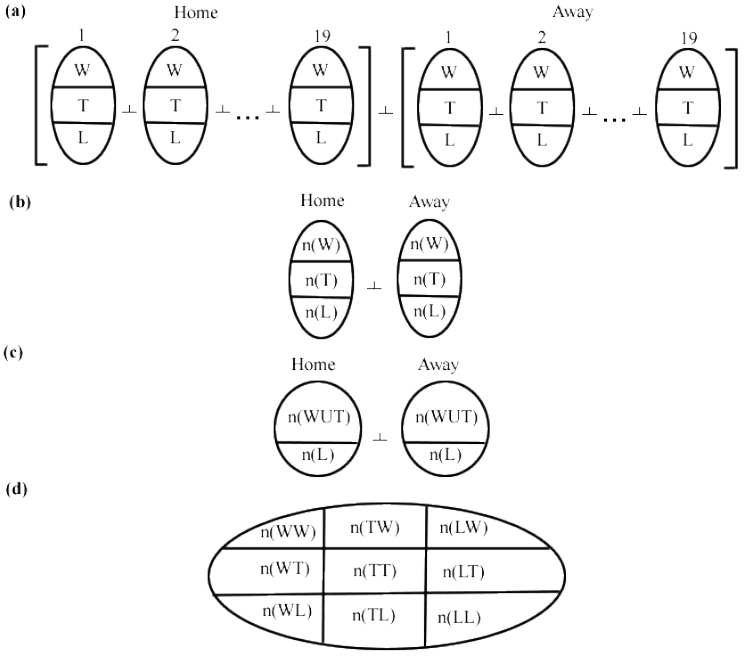
$$P(AB) = P(A)P(B).$$

Considere então como sendo A uma vitória em uma partida e B uma vitória na segunda partida. Assume-se aqui que a probabilidade de um time obter as duas vitórias é a probabilidade de obter a primeira vitória vezes a probabilidade de obter a vitória na segunda partida. Tal afirmação foi assumida, isto é, a partida em casa é independente da partida fora de casa e representou-se a independência com o símbolo "$\perp$ ". Se um partida de futebol é um ensaio e, embora possa haver várias covariáveis associadas aos resultados das partidas, o resultado de uma segunda partida é se dá pela observação do que acontece no tempo da segunda partida, não havendo uma dependência clara do resultado que aconteceu na partida anterior.

O presente estudo está focado em, dado um time, verificar qual a diferença dos resultados entre as partidas realizadas como mandante ou como visitante deste time. De maneira mais acertada, podemos dizer que queremos verificar a diferença nos pontos obtidos como mandante ou como visitante. Sendo que uma observação importante é que há o mesmo número de partidas como mandante e como visitante no modelo de competição que está sendo considerado no presente estudo. Desta maneira, foram estabelecidas quatro abordagens com as quais podemos encarar o presente problema de pesquisa, que estão resumidas na sequência.

A *primeira* forma que podemos escrever tal problema é estabelecer que cada ensaio é independente, ou seja, que o resultado da primeira partida é independente do resultado da segunda partida e assim por diante. Neste caso nós teríamos 38 distribuições multinomiais independentes. Como queremos verificar a diferença entre casa e fora, então teríamos dois vetores aleatórios, um para as 19 partidas de casa e um para as 19 partidas como visitante. Sendo que cada vetor aleatório é composto por 19 v.a. (Figura 1a), em que cada v.a. é o resultado de um partida (vitória, empate ou derrota) que segue uma distribuição multinomial.

Figura 1 – Representação de conjuntos com diagramas de quatro abordagens discutidas: *(a)* primeira abordagem; *(b)* segunda abordagem; *(c)* terceira abordagem e; (d) quarta abordagem. Onde: W é vitória, T é empate e L é derrota. *Home* se refere a partida como mandante e *away* refere-se a partidas como visitante. Ainda, $n(W)$ é uma função que indica o número de vezes que o evento W acontece.



Fonte: Próprios autores.

Uma *segunda* maneira de escrever o problema é com base na definição também de dois vetores aleatórios, um para as partidas jogadas em casa e outro para as partidas como visitante, que assumimos que são independentes. Porém, aqui cada vetor aleatório é composto apenas por 2 v.a. (Figura 1b): o número de vitórias e número de empates nas 19 partidas (uma vez que o número de derrotas pode ser determinado pelo número de empates e vitórias número de vitórias, então não é v.a.) . A *terceira* maneira de escrever o problema seria considerar o número de vitórias e o número de derrotas, sendo que a cada 3 empates, seria incluída uma vitória a mais (Figura 1c).

Há uma *quarta* maneira de escrever o problema, que utiliza um único vetor aleatório. Como é assumida a independência entre as partidas em casa e como visitante, e como cada time se enfrenta duas vezes, um na casa do primeiro e outro na casa do segundo, então observa-se esse par de partidas. Quando observa-se um par de partidas, nós temos 9 situações possíveis (Figura 1d). Sendo ainda, que da mesma forma como foi assumida a independência entre partidas como mandante e visitante, uma outra abordagem poderia ser considerado assumir dependência entre elas. Porém, foi interesse do presente estudo assumir que partidas como mandante e como visitante são independentes, ou seja, o resultado de uma partida como

mandante, não afeta diretamente o resultado de uma partida como visitante.

No Capitúlo 4 serão apresentadas mais características da métrica utiliza no estudo de Pollard, Silva e Medeiros (2008). E a partir dele vamos sugerir mudanças, e propor uma métrica com algumas características diferentes. Esta nova métrica será chamada de d e será exemplificada utilizando-se o Campeonato Brasileiro como aplicação. Já no Capítulo 5 serão apresentados um estudo sobre a distribuição dessa métrica, em que aqui receberá a denominação de variável aleatória e será representada por D. O Capítulo 6, por sua vez, traz o fechamento da dissertação.

4 PROPOSTA DE UMA MÉTRICA PARA O EFEITO DE CASA

PROPOSE OF A METRIC TO THE HOME EFFECT BASED ON AWARDED POINTS

Giovani Festa Paludo*

Nikolas Neves Figueiredo†

Eric Batista Ferreira‡

Resumo

Em competições esportivas de pontos corridos, o total de pontos ganhos em casa sobre o total de pontos consiste em uma métrica para medir o efeito de casa que é fácil de utilizar, de interpretar e que é muito conhecida. Porém, quando um time faz poucos pontos essa métrica fica inflacionada e oscila expressivamente, entre outras fragilidades. Assim, nosso objetivo foi construir uma nova métrica para o efeito de casa mantendo as características consideradas como positivas dessa métrica e ao mesmo tempo superando as fragilidades listadas. Com isso, foram propostos uma métrica e um teste para serem utilizados em campeonatos esportivos e utilizaram o Brasileirão Série-A (2003-2020) como modelo de estudo. Como principal resultado, a nova métrica não foi negativamente correlacionada com os pontos conquistados. O teste sobre a métrica permitiu verificar quando há informação suficiente para afirmar que existe vantagem de casa, tirando o efeito da oscilação quando o time ganha poucos pontos. Com base nisso, das 370 participações de 43 times em 18 edições, uma apresentou efeito de casa negativo, 259 tiveram efeito de casa positivo e 110 não apresentaram efeito de casa. Ainda, a métrica permitiu visualizar o comportamento individual de cada time ao longo das edições da competição.

Palavras-chave: vitória do mandante, estatísticas do esporte, futebol, Campeonato Brasileiro, liga.

Abstract

In double round-robin sports tournaments, the total of points awarded divided by the total of the points earned at home consists in metric to access the home advantage that is easy to use, to interpret and is widely known. However, when a team award a small number of points this metric is underestimated, vary expressively, among others fragilities. Thus, our objective was to build a new metric to measure home advantage keeping the positive aspects of the previous one and improving all the listed fragilities. It was proposed the new metric and a test to be used in sportive championships. As the study application, it was used data from Brazilian Championship Series-A (2003-2020). As our main result, the new metric was not negatively correlated to the awarded points. The test of the metric let to verify on which occasion there is enough information to affirm that exists home advantage, avoiding the effects of a small number of awarded points. Based on this, of the 370 participations of 43 teams in 18 editions, one participation presented home disadvantage, 259 presented home advantage and 110 did not present home effect. Besides, the metric permitted to visualize the individual behavior of each team throughout the league.

Keywords: home win, home ground effect, home cocking, soccer, sports statistics.

*http://lattes.cnpq.br/8897773821703545. Universidade Federal de Alfenas, gfpaludo@gmail.com.

†http://lattes.cnpq.br/3128218938439663. Universidade Federal de Alfenas, nikolasfig@gmail.com.

‡http://lattes.cnpq.br/9965398009651936. Universidade Federal de Alfenas, eric.ferreira@unifal-mg.edu.br

4.1 INTRODUÇÃO

Quando o time do coração vai jogar em casa, o torcedor já cria expectativas de que o time tem maiores chances de ganhar. E o torcedor não está errado, pois é bem conhecido no mundo dos esportes, o fato de que a probabilidade de um time obter uma vitória em uma partida é maior quando o time joga em casa (POLLARD; POLLARD, 2005). A esse fenômeno, de melhores resultados jogando em casa, é comumente atribuído o termo "vantagem de casa" (HA) para um efeito de casa positivo. Sendo que HA pode ser definida como "o resultado consistente no qual o time de casa em competições esportivas vence mais que 50% dos jogos em um calendário de jogos balanceado entre casa e fora de casa" (COURNEYA; CARRON, 1992). Ainda, a HA pode ser encontrada em diferentes esportes (POLLARD; POLLARD, 2005; DAWSON; MASSEY; DOWNWARD, 2020), é persistente ao longo do tempo (POLLARD; POLLARD, 2005; JACKLIN, 2005) e existem diferentes métricas para obtê-la (POLLARD; SILVA; MEDEIROS, 2008; GOUMAS, 2017; MAREK; VÁVRA, 2020).

Em competições esportivas no formato de pontos corridos, medir a vantagem de casa por pontos é bastante útil, pois conquistar pontos é o principal objetivo de cada time nesse tipo de competição. Por isso, entre as métricas utilizadas para obtenção de HA, as baseadas em pontos são frequentemente utilizadas (POLLARD; SILVA; MEDEIROS, 2008; LEITE, 2017; OLIVEIRA et al., 2020) e destaca-se a métrica obtida a partir da divisão entre total de pontos conquistados em casa e o total de pontos conquistados, que neste estudo será chamada de métrica h e que já foi utilizada em vários estudos com ou sem alguma correção (FAJARDO et al., 2019; TILP; THALLER, 2020; POLLARD; SILVA; MEDEIROS, 2008) ($h = (Y_c * 100)/(Y_f + Y_c)$; onde h é a função, Y_c são os pontos conquistados em casa e Y_f são os pontos conquistados fora de casa; sendo que h assume valores entre 0 e 100%, em que h = 50% é sem vantagem de casa, h = 100% é interpretado como o máximo de HA e h = 0% seria o máximo de desvantagem de casa). A métrica h é baseada em pontos, é fácil de ser calculada e interpretada, porém, é possível apontar algumas fragilidades.

A primeira fragilidade para se obter adequadamente um valor da função h para um time acontece quando o time conquista poucos ou muitos pontos do total de disputados. Para um exemplo, considere uma competição com as seguintes características: formato de pontos corridos, onde todos os times se enfrentam entre si duas vezes, uma vez na casa do primeiro e outra na casa do segundo; com 20 times; com 38 rodadas e; onde o vencedor de uma partida ganha 3 pontos e no caso de empate fica 1 ponto para cada time. Então, dentro dessa liga,

considere o time A e o time B (TABELA 1). Como fez poucos pontos, o valor da função h para o time A pode assumir valores entre 0% e 100%, mas o time B que fez muitos pontos só pode ter a métrica de h de 46% a 54%, assim a métrica h só permite valores de 100% quando o time pontua de 1 a até a metade do número total de partidas. Com base nisso, é possível observar que o valor de h depende dos pontos conquistados e provavelmente este é o motivo pelo qual pesquisas utilizaram metodologias para corrigir o valor de h em relação aos pontos conquistados (frequentemente chamada de correção pela habilidade) (CLARKE; NORMAN, 1995; POLLARD; SILVA; MEDEIROS, 2008; GOUMAS, 2017; OURS, 2019). Porém, essa correção por habilidade pode não ser suficiente de acordo com o próximo parágrafo.

Tabela 1 – Exemplo com 6 times hipotéticos (A, B, C, D, E e F) para mostrar as fragilidades da métrica h. Sendo que V significa vitória, E singifica empate e D significa derrota. Total se refere ao total de jogos em casa e como visitante.

	Casa			Total				
Time	V	E	D	V	E	D	Pontos	h
A	-	-	-	3	0	35	9	De 0% a 100%
B	-	-	-	35	0	3	105	De 46% a 54%
C	-	-	-	11	0	27	33	De 42% a 58%
D	-	-	-	0	33	5	33	De 0% a 100%
E	2	0	17	3	0	35	9	67%
F	1	0	18	3	0	35	9	33%

Fonte: Próprios autores.

A segunda fragilidade a ser destacada é o fato de que o número de pontos conquistados é importante, mas o número vitórias e o número de empates é mais importante que o número de pontos em si. Observe os dois times C e D (TABELA 1). Ambos tem a mesma pontuação, só que um ganhou os pontos em vitórias e outro ganhou os pontos em empates. O time C, matematicamente falando, poderá ter o valor de h entre 42% e 58%, enquanto que matematicamente falando, o time D poderá ter h de 0% a até 100%. Esse exemplo mostra que para realizar a correção no valor da métrica h deve-se considerar os pontos obtidos por empate e por vitória. Sendo que, embora hajam correções para a métrica h, os autores do presente estudo desconhecem algum estudo que tenha utilizada uma correção considerando o número de empates ou vitórias.

A terceira fragilidade está relacionada à situações em que um time ganha poucos pontos. Seguindo o mesmo formato de liga acima descrita e considerando como exemplo outros dois times E e F (TABELA 1). Ambos tiveram apenas 3 vitórias em um campeonato, porém o time E obteve uma vitória a mais em casa que o time F. Com isso, a h para o time E seria de 67%

e para o time F seria 33%. A diferença de uma vitória de um total de 3 vitórias é suficiente para afirmar que o E tenha 67% e o time F 33%? Os autores do presente estudo acreditam que uma vitória é pouco para justificar que o valor da métrica h varie de 33% para 67%. Ainda, fica evidenciado a oscilação que um número pequeno de vitórias provoca no valor da métrica. E por último, o time E e F podem ter tido essa diferença de 1 vitória por um efeito aleatório ou não relacionado, pois uma partida de diferença pode não ser suficiente para determinar que a h do time E foi 67% e que a h do F foi 33%.

Associado à essa terceira fragilidade, podemos observar que os estudos nem sempre trazem informações de efeito de casa para todos os times participantes. Eventualmente, é possível observar uma exclusão de certos times que participaram poucas vezes, por conta de uma limitação da metodologia empregada. Por exemplo, Goumas (2017) analisou uma competição com jogos do tipo mata-mata e mostrou os valores apenas para os times que tiveram mais de 50 partidas. Pollard, Silva e Medeiros (2008) calcularam valores de vantagem de casa apenas para times que participaram no mínimo em 3 edições do campeonato em formato de pontos corridos. Acredita-se no presente estudo que se um time participou uma única vez, isso é suficiente para que ele tenha alguma informação sobre o efeito da casa.

Considerando competições de esportes coletivos nas quais são jogadas o mesmo número de partidas em casa e fora de casa, o objetivo do presente estudo foi de desenvolver uma métrica para medir o efeito de casa que mantém as características positivas da métrica h, tais como: ser baseada em pontos conquistados, ser de fácil obtenção e interpretação. E também superar as três fragilidades e a observação apontadas acima, isto é: (i) espera-se que a nova métrica não é inflacionada quando o time obtém poucos pontos. Se essa afirmação é verdadeira, será confirmado que: a vantagem de casa obtida pela métrica h deverá ser negativamente correlacionada com os pontos conquistados, enquanto que a vantagem de casa obtida pela nova métrica não será negativamente correlacionada com os pontos conquistados. (ii) que avalie quando há evidências suficientes para concluir que exista efeito de casa e não seja apenas um efeito devido ao acaso. Ou seja, para superar a segunda e a terceira fragilidade também foi objetivo apresentar um teste para avaliar quando se tem evidências suficientes para concluir ou não que é um efeito de casa, e não apenas um efeito devido a questões ao acaso como pouca informação. Sendo que ainda o teste deve considerar a configuração de vitórias e empates que aconteceu ao invés de considerar apenas os pontos.

4.2 MATERIAL E MÉTODOS

4.2.1 Descrição dos dados

Como exemplo de aplicação foi utilizado o Campeonato Brasileiro de Futebol Série A, utilizando-se dados de todas as edições da história dos pontos corridos, ou seja, de 2003 a 2020. A partir de 2003, todos os times passaram a se enfrentar duas vezes e o campeão passou a ser o time que acumulou mais pontos somando-se todos os enfrentamentos. Sendo que em 2003 e 2004, o campeonato tinha 24 times participantes, 46 rodadas e 552 partidas. Já em 2005, foram 22 times, 42 rodadas e 462 partidas. No de 2006 até os dias atuais o campeonato tem 20 times participantes, 380 partidas em 38 rodadas. Os dados utilizados no presente estudo foram obtidos no <www.soccerway.com>, website que também foi utilizado nos estudos de Pollard, Silva e Medeiros (2008), Silva et al. (2018). As variáveis coletadas foram: nome do time mandante, nome do time visitante, gols do time da casa e gols do time visitante.

Cabe ainda ressaltar que, por exemplo, em uma partida entre Grêmio *versus* Vasco. Entendeu-se o primeiro time (Grêmio) como mandante e o segundo time como visitante (Vasco), não importando o estádio em que o jogo aconteceu. Assim, todas as vezes que o termos "partida em casa" e "partida fora de casa" são utilizados, eles se referem, respectivamente à "partida como mandante" e "partida como visitante".

Ainda, há algumas observações sobre dados não considerados no presente estudo: (i) como aqui foi utilizado a pontuação conquistada por partida, então todas as punições que implicaram na subtração ou adição de pontos na tabela de pontuação do campeonato não foram consideradas no presente estudo. Tais subtrações ou adições aconteceram em 2003, 2004, 2005, 2010 e 2013 (2003: -4 Ponte Preta, -8 Paysandu, +3 São Caetano, +3 Ponte Preta, +2 Corinthians, +2 Fluminense, +3 Juventude e +2 Internacional; 2004: -24 São Caetano; 2005: -1 Brasiliense e +2 Vasco; 2010: -3 Grê.Barueri; 2013: -4 Portuguesa e -4 Flamengo); (ii) há 23 partidas ocorridas nos anos de 2016, 2018 e 2019 nos quais o time que jogaria em casa vendeu o mando de campo provavelmente por questões financeiras. Com isso, a partida ocorreu em outro local, sendo que essas partidas não foram removidas do banco de dados; (iii) não foram consideradas as partidas anuladas no ano de 2005, sendo que foram consideradas apenas as novas partidas e; (iv) a pontuação da partida da Chapecoense na última rodada da edição de 2016 foi considerada de maneira igual à considerada pela Confederação, que foi de derrota para ambos os times.

4.2.2 Métrica proposta e análise de dados

Para quantificar o quanto o desempenho de um time é melhor em casa do que como visitante, foi necessário estabelecer algumas definições. A primeira delas foi a diferença de pontos absoluta (da) que ficou definida como a diferença entre os pontos conquistados como mandante (Y_c) e os pontos conquistados como visitante (Y_v) no final do campeonato, isto é:

$$da = Y_c - Y_v.$$

Em seguida, estabeleceu-se a diferença de pontos relativa (d), que ficou definida como a diferença de pontos absoluta multiplicada por um fator que faça com que o maior valor possível de diferença, seja 100%, isto é,

$$d = \frac{Y_c - Y_v}{n_c \times v} \times 100, \tag{4.1}$$

em que v são os pontos por vitória; n_c é o número de partidas ou rodadas que um time joga em casa em uma edição do campeonato.

Ainda, o efeito de casa médio amostral de um time ($\bar{D}$) obtido pela métrica d, será definido como a média dos valores de d, isto é:

$$\bar{D} = \frac{\sum_{i=1}^{m} d_i}{m},$$

em que i é cada participação do time em uma edição do campeonato e m é o somatório do i, ou seja, o total de vezes que um time participou. Sendo que $\bar{H}$ e $\bar{DA}$ são médias obtidas da mesma forma que $\bar{D}$, isto é, $\bar{DA} = \sum_{i=1}^{m} da_i/m$ e $\bar{H} = \sum_{i=1}^{m} h_i/m$. Onde da_i são os valores de da para cada participação i de um total de m participações e h_i são os valores de h de um time em uma participação i de um total de m participações. Ressalta-se que tanto o $\bar{H}$, $\bar{DA}$ e $\bar{D}$ são médias amostrais obtidas de um número finito de participações de um espaço amostral com infinitas participações possíveis. Considerando a métrica d, porém pode-se obter uma média amostral $\bar{D}$, dado a impossibilidade de obtenção da média populacional δ, e ainda pode-se obter o intervalo de confinaça para a média populacional ($IC[\delta; 95\%]$). O mesmo acontece para a métrica da, sendo que a média populacional ξ é desconhecida, sendo que o que pode ser obtido é uma média amostral $\bar{DA}$, e um intervalo de confiança para a média populacional ($IC[\xi; 95\%]$). Para a obtenção das estimativas intervalares para a média populacional foi utilizado a estatística t a 95% de probabilidade do intervalo conter o verdadeiro

parâmetro populacional.

Uma maneira bastante comum para se medir o efeito positivo da casa é a métrica h sem correção, que diz respeito à porcentagem dos pontos conquistados em casa em relação ao total de pontos conquistados. Podendo ser escrita na forma de

$$h = \frac{Y_c}{Y_c + Y_v} \times 100, \tag{4.2}$$

em que: Y_c é a variável pontos conquistados em casa; Y_v é a variável pontos conquistados como visitante e; o valor de $h = 100\%$ indica o máximo de vantagem de casa, $h = 50\%$ indica nenhuma vantagem de casa e $h = 0\%$ indicaria o máximo de desvantagem de casa.

Para permitir a comparação da estatística criada no presente estudo com a estatística h (4.2) já desenvolvida, foi estabelecida a vantagem de casa centrada em 0 (h_0), que ficou definida como:

$$h_0 = 2h - 100,$$

em que h_0 varia de -100% quando a desvantagem de casa é total, 0 quando não há vantagem e nem desvantagem e $h = 100\%$ quando a vantagem de casa é total.

4.2.3 Inferência para uma participação

Como exposto na terceira fragilidade, há diferentes valores possíveis de vantagem de casa dependendo se o ponto veio de um empate ou vitória. Então, inicialmente foi escrito o vetor denominado P que foi composto por todos os valores de pontos obtidos por partida de um certo time em uma edição do campeonato, indo de 1 a 38 no Campeonato Brasileiro.

$$B = \{b_1, b_2, b_3, ..., b_{38}\},$$

em que b_1 é a pontuação obtida na partida da rodada 1, b_2 na rodada 2, e assim por diante.

Com isso foi definido o vetor R, um novo vetor constituído com o mesmos valores do vetor B, porém com os valores em uma ordem diferente da original. Sendo que a ordem foi obtida com um processo de aleatorização, e assim foi escrito um novo vetor:

$$R = \{r_1,r_2,r_3,...,r_{38}\}.$$

Para gerar uma distribuição com os possíveis valores da métrica d, foi então definido que os pontos em casa seriam a primeira metade dos elementos e os pontos conquistados fora de casa a segunda metade do conjunto.

$$Y_c = \sum_{i=1}^{19} r_i \text{ e } Y_v = \sum_{i=20}^{38} r_i$$

O próximo passo constituiu-se na obtenção do conjunto com os valores de d. Ou seja, aplicou-se a expressão (4.1) para cada sorteio e obteve-se um valor de d que foi chamado de ds_i, onde i é o número da simulação, e ds é a diferença de pontos relativa de cada simulação. Ao final de 1000000 de simulações (procedimento detalhado na Apêndice A), obteve-se o conjunto:

$$DS = \{ds_1, ds_2, ..., ds_{1.000.000}\}$$

Em seguida ordenou-se esse conjunto e foram obtidos os percentis de interesse, que no caso foi deixando 20% dos valores nas caudas, isto é, os percentis de 10% e 90%. Assim definimos o intervalo de predição (IP) composto pelo limite inferior (IP_{LI}) e pelo limite superior (IP_{LS}) que podem ser escritos da forma:

$$IP_{LI} = Sort(DS)_{0.10*k} \text{ e } IP_{LS} = Sort(DS)_{0.90*k},$$

em que $Sort$ é uma função do Programa R que ordena os elementos do conjunto DS e k é o número de reamostras.

A verificação da existência do efeito de casa não nulo foi realizada pela comparação entre a métrica d e os IP_{LS} e IP_{LI}. Admite-se neste trabalho que, se $d < IP_{LI}$, então há efeito de casa negativo que pode ser chamado de desvantagem de casa; se $IP_{LI} \leq d \leq IP_{LS}$, então não há evidencias suficientes para afirmar que há efeito de casa e; se $IP_{LS} < d$, então há efeito de casa positivo, ou seja, há vantagem de casa.

Ainda, para verificar se existia correlação entre h e os pontos obtidos e entre d e os pontos obtidos, foi utilizado uma correlação linear simples e um teste t para avaliar a significância do coeficiente r de correlação de linear simples. E nessas duas correlações foram utilizados as edições com 20 times participantes, ou seja, todas as 15 edições que ocorreram de 2006 à 2020. Todas as análises e gráficos foram realizadas utilizando-se o software estatístico R (R Core Team, 2021) sendo que na Apêndice A do presente artigo foi disponibilizado um código em linguagem R que calcula d, da e o intervalo de predição de d.

4.3 RESULTADOS E DISCUSSÃO

4.3.1 Características e média global da métrica d

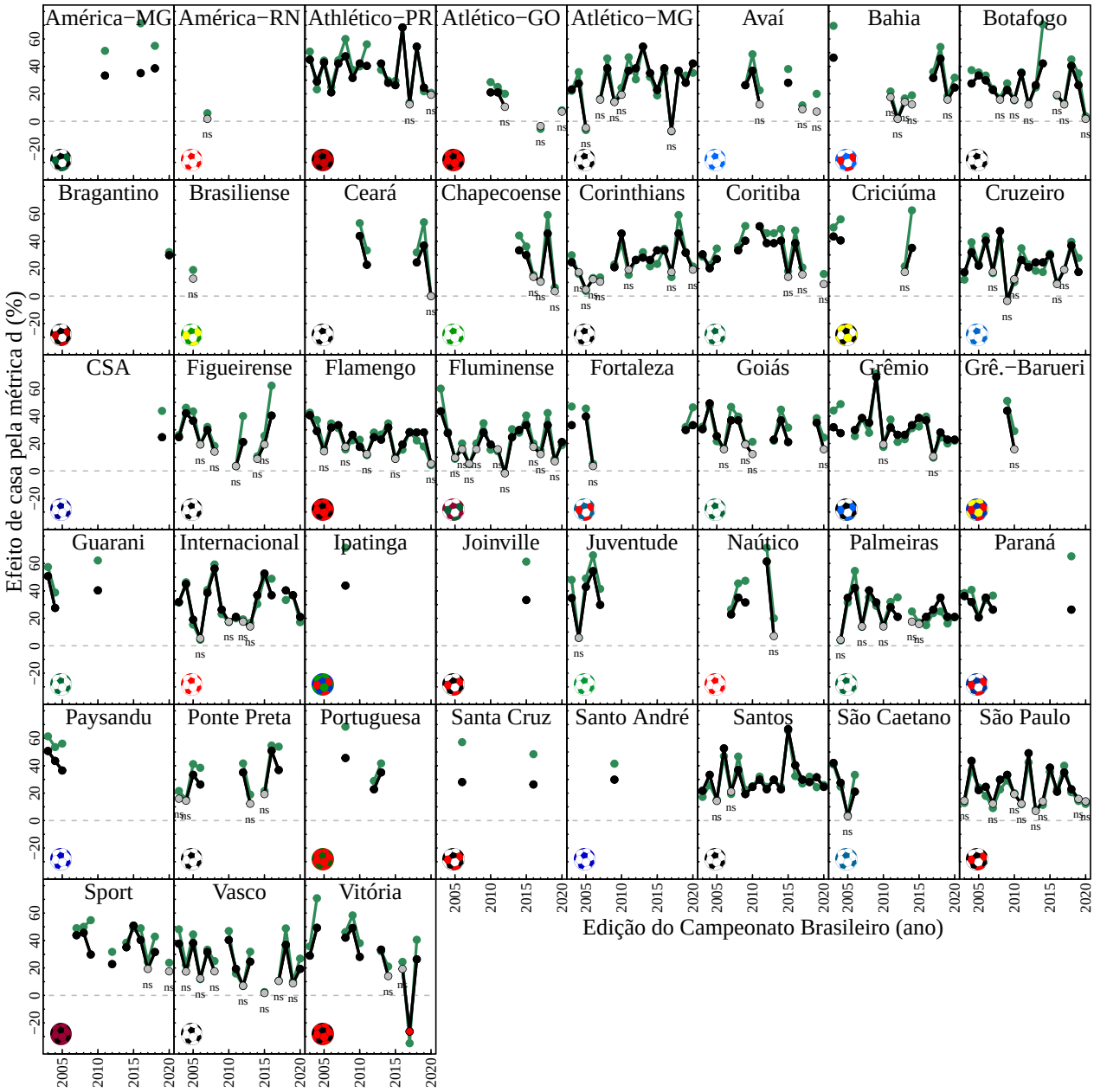
A Série A do Campeonato Brasileiro de Futebol de 2003 a 2020 teve 12879 pontos conquistados em casa e 7014 pontos conquistados fora de casa. Isso gerou uma diferença absoluta global de pontos, isto é, um $da = 5865$ pontos. A média global de todos os 370 valores obtidos da métrica d foi 26,80% e quando utilizou-se a métrica h obteve-se uma média global de 64,74%, que resultou em um h_0 global de 29,48%. Três aspectos podem ser observados em relação à métrica d nesses três resultados.

O primeiro aspecto é sobre sua interpretabilidade. Como a métrica d pode receber valores de no mínimo -100% quando o time conquistou todos os pontos possíveis fora de casa e nenhum em casa, pode receber o valor máximo de 100% quando o time conquistou todos os pontos possíveis em casa e nenhum fora de casa e d é 0 quando o mesmo número de pontos foi conquistado em casa ou fora de casa. Então, a média 26,80% de todas os valores de d significa que, do total de pontos disputados, em média 26,80% foram conquistados a mais em casa do que como visitante. Sendo que d é uma métrica que reflete diretamente a porcentagem dos pontos ganhos a mais em casa do que fora de casa. Essa porcentagem pode ser utilizada para comparação de competições de pontos corridos de diferentes esportes que utilizem sistema semelhante de pontuação.

Um segundo aspecto a ser destacado é que os valores das métricas d e h são valores próximos e podem ser comparados desde que h seja multiplicado por dois e subtraído em 100 unidades, isto é, que seja obtido o h_0. Esta proximidade pode ser observada na Figura 2 que mostra tanto os valores de d quanto os valores de h_0 por time e por ano. Esse valor médio da métrica d encontrado no presente trabalho está próximo ao encontrado na literatura. Pollard, Silva e Medeiros (2008) estudaram as edições de 2003 a 2007 e utilizaram a mesma métrica h, porém com uma correção, sendo que encontraram um h médio de 65% de vantagem de casa que gera um h_0 médio global de 30%. Já Fajardo et al. (2019) encontraram a média da métrica h de 65,6% de 2012 a 2016, que gera um h_0 global de $31{,}2\%$. Por outro lado, Oliveira et al. (2020) estudaram a edição de 2017 utilizando uma métrica diferente das utilizadas no presente artigo e observaram 54% de vantagem de casa. No presente estudo, a média de d para quem participou em 2017 foi de 15,0%, a menor média registrada na sequência de 2003 a 2020. Assim, um característica da métrica d é que a métrica corrige as fragilidades supracitadas e ao

mesmo tempo traz um valor não muito distante ao que foi obtido pela métrica h em estudos anteriores também sobre o Campeonato Brasileiro, mesmo que tenham ou não utilizado alguma correção (POLLARD; SILVA; MEDEIROS, 2008; FAJARDO et al., 2019).

Figura 2 – Comparação entre as métricas h_0 (pontos com contorno e preenchimento verdes) e d (pontos com contorno em preto) para medir o efeito de casa para cada participação de cada time ao longo das edições do Campeonato Brasileiro de Futebol (Série A) de 2003 à 2020. Dos pontos com contorno em preto (métrica d), os preenchidos em preto indicam vantagem de casa, em vermelho desvantagem de casa e em cinza indicam ausência de efeito significativo da casa.

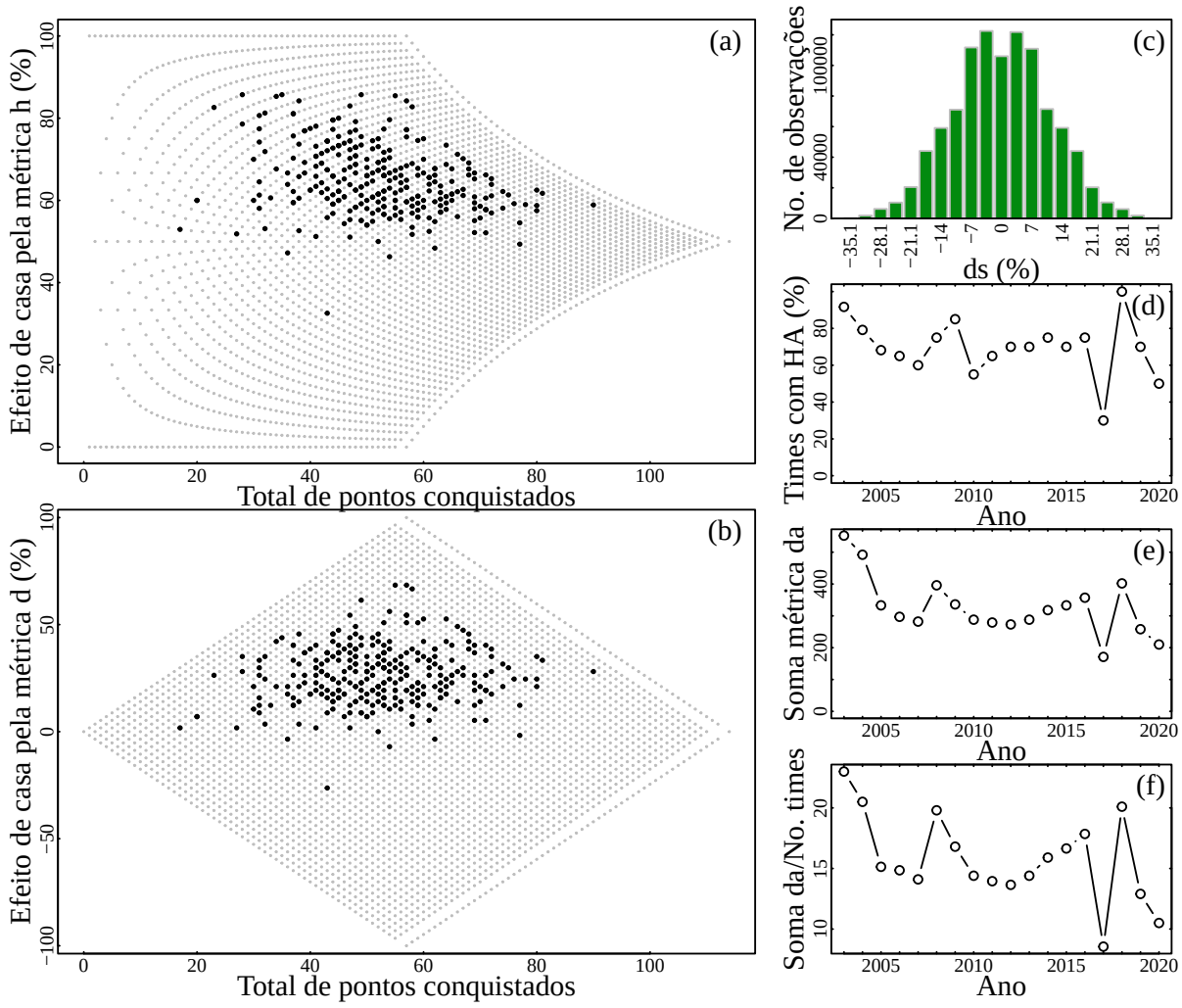


Fonte: Próprios autores.

Um terceiro aspecto importante a ser destacado é que a principal diferença entre os va-

lores de d e h_0 obtidos para o Campeonato Brasileiro, se deve principalmente à natureza da métrica h, que, ao dividir a diferença de pontos pelo total de pontos, acaba inflacionando a métrica, pois quanto menor o número de pontos, o valor da métrica tenderá a ficar maior. A métrica d não faz essa divisão e como é possível observar nos pontos em cinza na Figura 3a e 3b, os possíveis valores da métrica d são simétricos, enquanto que os possíveis valores de h são assimétricos quando observados ao longo dos pontos totais conquistados. Esse resultado é corroborado pela análise de correlação linear simples entre o total de pontos conquistados de cada time e o respectivo valor da métrica h (4.2) que mostrou uma correlação linear negativa (r=-0,296; n=300; valor-p$<$0,0001; FIGURA 3a). Ao contrário, a análise de correlação linear simples entre os pontos conquistados de cada time em cada ano e o valor da métrica d (4.1), não apresentou correlação significativa (r=0,101; n=300; valor-p=0,081; FIGURA 3b). Assim, confirmou-se o que era esperado que a vantagem de casa obtida pela nova métrica d não foi negativamente correlacionada com os pontos ganhos. Com isso, o resultado suporta que a nova métrica não é inflacionada quando se tem situações de poucos pontos, isto é, a métrica não é inflacionada pela habilidade do time. Vários estudos sobre o efeito positivo da casa fazem uma correção pelos pontos conquistados e tal correção não é recente (CLARKE; NORMAN, 1995; POLLARD; GóMEZ, 2007; POLLARD; SILVA; MEDEIROS, 2008). Assim, ao contrário do que acontece com a métrica h sem correção, a métrica d não foi inflacionada quando um time fez poucos pontos, não sendo mais necessário realizar uma correção pela habilidade. Característica importante para uma métrica para acessar a vantagem de casa baseada em pontos. Ainda, a eliminação da inflação pode contribuir também com estudos que fazem previsões de resultados de jogos de futebol (exemplo: RAMOS; FERNANDES; BATISTA, 2021).

Figura 3 – *(a)* O efeito de casa obtido pela métrica h para 300 participações em 15 edições (2006 à 2020) do Campeonato Brasileiro (pontos pretos) e todos os valores possíveis para a métrica h (pontos cinza); *(b)* O efeito de casa obtida pela nova métrica d para as 300 participações de 2006 à 2020 do Campeonato Brasileiro (pontos pretos) e todos os valores possíveis para a métrica d (pontos cinza). *(c)* Distribuição dos valores de diferença de pontos relativa simulada obtidas nas 1000000 reamostras para a configuração de vitórias, empates e derrotas do Naútico em 2013. Essa distribuição empírica indica quais são os valores mais prováveis para serem encontrados em um novo ensaio, dado o que já aconteceu. *(d)* Número de times que tiveram efeito de casa positivo em relação ao total de times que participou da edição do campeonato para as 18 edições de 2003 à 2020. *(e)* Soma da métrica da para todos os times que participaram de 2003 à 2020; *(f)* Soma da métrica da dividida pelo número de times que participaram naquela edição, considerando os anos de 2003 à 2020.



Fonte: Próprios autores.

4.3.2 Inferência para participação de um time em uma única edição da competição e representação longitudinal

A principal novidade do presente estudo é a possibilidade de obtenção de uma informação sobre o efeito de casa para uma única participação de um time em uma competição. Como um exemplo dessa aplicação da métrica e do respectivo teste, considerou-se o time do Naútico em 2013, que fez 12 pontos em casa e 8 pontos como visitante. A diferença de pontos absoluta, $da = P_c - P_v = 12 - 8 = 4$, e a diferença de pontos relativa, $d = (P_c - P_v) \times 100/(n_c \times v) = (12 - 8) \times 100/(19 \times 3) = 7{,}02\%$. E a partir da distribuição de todos os valores de d obtidos no processo de reamostragem quando os pontos ganhos foram sorteados entre casa e fora (FIGURA 3c), o intervalo de predição deixando 20% das observações nas caudas foi $IP_{LI} = -14{,}035\%$ e $IP_{LS} = 14{,}035\%$. Como o valor de d para o Naútico em 2013 foi de 7,02% e está situado dentro do intervalo de predição foi de -14,0% e 14,0%, então conclui-se que não há evidências suficientes para afirmar que existiu efeito de casa para o Naútico em 2013. Já utilizando a métrica h, Naútico obteve $h = (P_c \times 100)/(P_c + P_v) = (12)/(12+8) = 60{,}0\%$, e $h_0 = 2h - 100 = 120{,}0 - 100{,}0 = 20{,}0\%$. Embora que o valor de 20% da métrica h_0 indicaria a existência de vantagem de casa, não é possível afirmar que exista efeito positivo da casa a partir do teste sobre a métrica d. A diferença encontrada não foi suficiente para afirmar que existe vantagem de casa para o Naútico de acordo com o teste considerado no presente estudo. Sendo que o valor de $h_0 = 20\%$ do presente estudo, é um exemplo de um valor que pode ser considerado inflacionado, pois foi resultado de apenas uma vitória e um empate a mais em casa do que como visitante.

Quando a métrica e o teste da métrica do exemplo acima são aplicados à todas 370 participações geradas pelos 43 times nas 18 temporadas do Campeonato Brasileiro de Futebol, obtiveram-se 370 valores de efeito de casa. Desses, uma única participação apresentou efeito de casa negativo (desvantagem de casa; que foi o time do Vitória em 2017), 259 apresentaram efeito de casa positivo (vantagem de casa) e 110 não apresentaram efeito de casa a um nível de significância de 20% (FIGURA 2). O ano com mais times com efeito de casa positivo foi 2018 que todos os 20 times participantes apresentaram HA e o ano com menos times com HA foi 2017 com 6 times de um total de 20 participantes (FIGURA 3d). Já em 2020, que teve todas as partidas durante a pandemia do coronavírus, 10 times tiveram efeito positivo da casa. Seria esperado que na edição de 2020 haveriam os menores valores de vantagem de casa, porém 2017 foi o ano com menores valores. Ainda, poderia ser esperado uma redução na vantagem de casa

ao longo do tempo, como discutido em Leite (2017). Porém, os resultados do presente estudo, tanto sobre o número de times com efeito de casa positivo (FIGURA 3d), quanto a soma da métrica da (FIGURA 3e) e a soma da métrica da dividida pelo número de times (FIGURA 3f), mostraram expressiva variação da vantagem de casa ao longo do tempo, sem um padrão de queda quando analisado graficamente.

O efeito de casa se mostrou variável entre os clubes (FIGURA 2), informação coerente com o apontado pela literatura para o Campeoanto Brasileiro (POLLARD; SILVA; MEDEIROS, 2008) e outras competições de outros países (CLARKE; NORMAN, 1995; GOUMAS, 2017). Ainda, existiu uma expressiva variação para um mesmo clube ao longo do tempo, também verificado em outro estudo que utilizou saldo de gols para acessar a vantagem de casa ao invés de pontos no campeonato inglês (CLARKE; NORMAN, 1995). Sendo que esta visualização ao longo do tempo é importante e foi proporcionada pela métrica e pelo teste apresentado no presente estudo. Por exemplo, o Athlético-PR passou a utilizar gramado sintético na temporada de 2016, ano que apresentou o maior valor observado de efeito positivo de casa para o time. Sendo que em um estudo realizado em país europeu, o gramado sintético proporcionou maior vantagem de casa (OURS, 2019) e o gramado sintético pode ser uma das causas para esse valor observado para o Athlético-PR em 2016, fato que merece estudos adicionais. Como outro exemplo, pode ser citado o time do Internacional que em 2013 não utilizou o seu estádio e não apresentou vantagem de casa, mas apresentou vantagem de casa em 2014 e 2015 (FIGURA 2), após a reforma do estádio. Essa detecção de variação ao longo do tempo é um passo importante para auxiliar na explicação dos principais fatores que afetam a vantagem de casa, questão que ainda não está completamente elucidada na literatura. Seria importante que fosse observada essa variação ano a ano para buscar associações do que pode explicar o efeito positivo da casa.

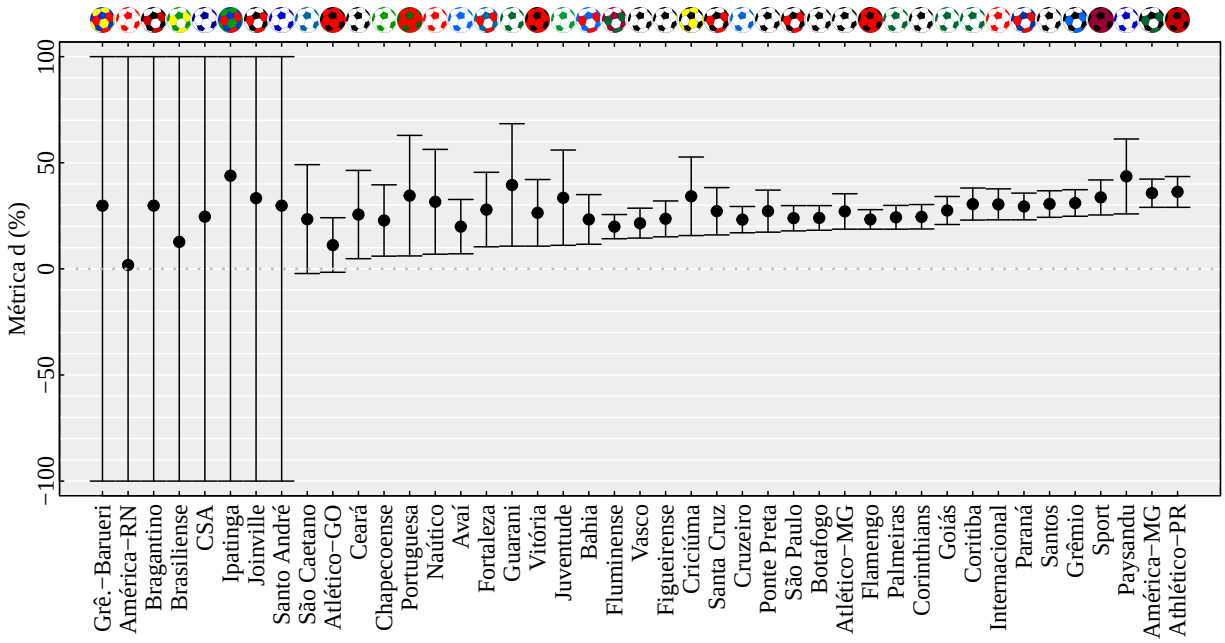
Outro aspecto que pode ser observado é sobre o rebaixamento de times da Série A para a Série B do campeonato. Dos 43 times que participaram no Campeonato Brasileiro Série A, 38 deles foram rebaixados em algum momento. Na Figura 2, é possível visualizar quando um time caiu da Série A para a Série B, que foi quando um ponto não foi sucedido por uma linha. Ainda, em algumas destas participações que terminaram com queda para outra divisão, há efeito de casa positivo significativo no ano anterior, mas no ano da queda não há efeito de casa significativo. Esse padrão de não haver significância no ano que caiu, aconteceu em 14 times, como exemplo o Atlético Goianense em 2007, o Atlético-MG em 2005, Avaí em 2011, entre outros. Eis duas prováveis explicações: (i) se o time está indo mal na tabela, a

cobrança da torcida pode passar a ter um efeito negativo sobre o desepenho do time ou; (ii) se um time cai para a outra divisão é porque fez poucos pontos, e com poucos pontos é mais fácil acontecer a situação da informação ser insuficiente para o teste acusar que exista um efeito de casa. Questão essa que merece estudos adicionais.

4.3.3 Intervalo de confiança para a média populacional de d

Quando não é mais realizada uma avaliação de cada participação de um time, mas obtem-se uma média amostral $\bar{D}$ da métrica d de um número n de participações (TABELA 2), pode-se realizar o ranqueamento dos times de acordo com a vantagem de casa (FIGURA 4). Assim, ranqueando-se pelo intervalo de confiança inferior, os times dos quais se pode esperar maiores valores de efeito positivo da casa são: Atlético-PR, América-MG e Paysandu (FIGURA 4; TABELA 2). Os três times que se espera um efeito menos expressivo são: São Caetano, Atlético-GO e Ceará. No estudo de Pollard, Silva e Medeiros (2008) os times com maior vantagem de casa foram o Paysandu, Athlético-PR e Juventude, semelhante ao encontrado no presente estudo. O Paysandu obteve 24,9% no estudo de Pollard, Silva e Medeiros (2008), e no presente estudo foi 21,8%. Ainda, os 8 primeiros times expostos na Figura 4 têm o intervalo de confiança em toda a faixa possível da métrica d, uma vez que 7 deles participaram uma única vez e isso faz com que não seja possível a obtenção de um intervalo de confiança.

Figura 4 – Intervalo de confiança para o efeito da casa medido pela métrica d para todos os times que participaram do Campeonato Brasileiro Série A de 2003 a 2020. Em que o ponto do centro é a média da diferença de pontos relativa e as barras verticais nas extremidades representam os limites do intervalo de confiança para a média populacional com 95% de probabilidade de conter um novo valor de vantagem de casa em uma nova participação de um time.



Fonte: Próprios autores.

Tabela 2 – Número de participações de cada time (m), média amostral da métrica da ou diferença de pontos absoluta ($\bar{DA}$), desvio padrão da métrica da ($S_{\bar{DA}}$), intervalo de confiança para a média populacional de da (IC[ξ;95%]), média amostral da métrica d ou diferença de pontos relativa ($\bar{D}$), desvio padrão da métrica d, ($S_{\bar{D}}$), intervalo de confiança para a média populacional de d (IC[δ;95%]), valores obtidos pela métrica h e número de vezes que um time apresentou efeito positivo da casa (m_+). São dados de vantagem de casa para todas as participações dos 43 clubes nas 18 edições do Campeonato Brasileiro de Futebol Série A.

Time	m	$\bar{DA}$	S_{da}	IC[ξ;95%]	$\bar{D}$	S_d	IC[δ;95%]	$\bar{H}$	m_+
América-MG	3	20,3	1,5	16,5; 24,1	35,7	2,7	29; 42,3	79,6	3
América-RN	1	1			1,8			52,9	0
Athlético-PR	17	21,4	8,4	17; 25,7	36,3	14,1	29; 43,5	69,3	15
Atlético-GO	5	6,4	5,9	-0,9; 13,7	11,2	10,3	-1,6; 24,1	57,6	2
Atlético-MG	17	15,8	9,3	11; 20,5	27,1	16,2	18,7; 35,4	63,8	12
Avaí	6	11,3	6,9	4; 18,6	19,9	12,2	7,1; 32,7	64	3
Bahia	9	13,9	9,9	6,3; 21,5	23,3	15,3	11,6; 35	65	4
Botafogo	16	14	6,5	10,5; 17,5	24	10,9	18,2; 29,8	64	10
Bragantino	1	17			29,8			66	1
Brasiliense	1	8			12,7			59,5	0
Ceará	5	14,6	9,6	2,7; 26,5	25,6	16,8	4,8; 46,4	67,2	4
Chapecoense	6	13	9,1	3,4; 22,6	22,8	16	6; 39,6	64,3	3
Corinthians	17	14,3	6,3	11; 17,6	24,5	11,2	18,8; 30,3	62,3	10
Coritiba	13	18	7	13,7; 22,3	30,5	12,6	23; 38,1	68	10
Criciúma	4	22	9,1	7,5; 36,5	34,2	11,6	15,7; 52,7	73,8	3
Cruzeiro	17	13,6	7,1	10; 17,3	23,2	12,1	17; 29,4	61,9	12
CSA	1	14			24,6			71,9	1
Figueirense	11	14,4	8,3	8,8; 19,9	23,6	12,6	15,1; 32	64,9	6
Flamengo	18	13,8	6	10,8; 16,8	23,3	9,3	18,7; 27,9	61,7	13
Fluminense	18	11,8	7,4	8,2; 15,5	19,9	11,5	14,2; 25,6	61,4	9
Fortaleza	5	17,2	9,1	5,9; 28,5	27,9	14,1	10,4; 45,5	67,6	4
Goiás	13	16,5	7,4	12; 21	27,5	11	20,9; 34,1	65,6	9
Grêmio	17	18,1	6,9	14,6; 21,7	31	12,1	24,8; 37,3	65,7	15
Grê.-Barueri	2	17	11,3	-84,6; 118,6	29,8	19,8	-148,5; 208,2	70	1
Guarani	3	25,7	8,3	5; 46,4	39,5	11,6	10,7; 68,4	76,4	3
Internacional	17	17,9	8,5	13,6; 22,3	30,4	14,2	23,1; 37,7	65	13
Ipatinga	1	25			43,9			85,7	1
Joinville	1	19			33,3			80,6	1
Juventude	5	20,6	10,6	7,4; 33,8	33,5	18,1	11,1; 56	71	4
Naútico	5	18	11,3	3,9; 32,1	31,6	19,9	6,9; 56,3	71,1	4
Palmeiras	16	14	6,1	10,8; 17,2	24,3	10,5	18,7; 29,9	62,3	11
Paraná	6	18,3	4,7	13,4; 23,3	29,4	6	23,1; 35,7	69,6	6
Paysandu	3	29,3	6	14,4; 44,3	43,6	7,1	25,9; 61,2	78,5	3
Ponte Preta	9	16,1	7,1	10,7; 21,6	27,2	12,9	17,3; 37,1	67,1	5
Portuguesa	3	19,7	6,5	3,5; 35,8	34,5	11,4	6,1; 62,9	73,2	3
Santa Cruz	2	15,5	0,7	9,1; 21,9	27,2	1,2	16; 38,3	76,4	2
Santo André	1	17			29,8			70,7	1
Santos	18	17,8	7,1	14,3; 21,4	30,6	12,6	24,3; 36,8	64,9	16
São Caetano	4	15,5	11,4	-2,6; 33,6	23,4	16,1	-2,2; 49,1	62,8	3
São Paulo	18	14,1	7,3	10,4; 17,7	23,9	11,9	17,9; 29,8	60,9	10
Sport	10	19,2	6,6	14,5; 23,9	33,7	11,5	25,4; 41,9	70,7	8
Vasco	15	12,9	7,9	8,5; 17,2	21,5	12,7	14,5; 28,6	62,9	8
Vitória	10	16	13,4	6,4; 25,6	26,4	21,9	10,7; 42,1	66,6	7

Fonte: Próprios autores.

4.4 CONSIDERAÇÕES FINAIS

Este trabalho trouxe a proposta de uma métrica, isto é, uma função de variáveis aleatórias para se obter o efeito de casa considerando alguns aspectos, sendo que o principal aspecto é o não inflacionamento da métrica quando um time ganha poucos pontos. A métrica exprime o resultado como uma percentagem em relação ao total de pontos conquistados possíveis de serem conquistados, o que permite comparar diferentes competições de pontos corridos ou até diferentes esportes que atribuem pontuação por vitória e por empates (quando existirem empates).

O presente estudo também utilizou um teste, que na forma como foi aplicado levou em consideração se os pontos foram obtidos em empates ou vitórias. Sendo que o teste permite verificar se é possível afirmar que há efeito de casa quando existe efeito de casa positivo (vantagem de casa), efeito de casa negativo da casa (desvantagem de casa) ou se não há evidências suficientes para afirmar que exista efeito de casa. Um resultado principal é que existiu o efeito positivo da casa (vantagem de casa) em 70% das participações dos times na competição, sendo que a desvantagem de casa aconteceu apenas uma única vez. Ainda, a métrica e o teste permitiram observar que há expressiva variação na vantagem de casa de um time ao longo do tempo.

Estudos futuros poderiam obter valores de vantagem de casa para cada time e com um maior nível de detalhamento, como por exemplo filtrar os ruídos promovidos pelas perdas judiciais de mando de campo, vendas de mando de campo, entre outros. Ou ainda, estabelecer metodologias para mitigar o efeito desses ruídos na obtenção do valor da vantagem de casa por time. Também seria interessante responder qual métrica consegue predizer melhor valores futuros de vantagem de casa. E ainda, poderia ser investigado quais fatores afetam a vantagem de casa utilizando a presente métrica d, que não é inflacionada como a métrica h, especialmente no sentido de investigar o que pode estar associado a variação observada para cada clube em cada participação na competição.

REFERÊNCIAS

CLARKE, S. R.; NORMAN, J. M. Home ground advantage of individual clubs in english soccer. **Journal of the Royal Statistical Society: Series D (The Statistician)**, [*s. l.*], v. 44, n. 4, p. 509–521, 1995.

COURNEYA, K. S.; CARRON, A. V. The home advantage in sport competitions: A literature review. **Journal of Sport and Exercise Psychology**, [*s. l.*], v. 14, n. 1, p. 13–27, 1992.

DAWSON, P.; MASSEY, P.; DOWNWARD, P. Television match officials, referees, and home advantage: Evidence from the european rugby cup. **Sport Management Review**, [*s. l.*], v. 23, n. 3, p. 443–454, 2020.

FAJARDO, L. et al. A vantagem de jogar em casa em relação às séries do campeonato brasileiro de futebol. **Revista Brasileira de Futebol**, Viçosa, v. 10, n. 2, p. 25–34, 2019.

GOUMAS, C. Modelling home advantage for individual teams in uefa champions league football. **Journal of Sport and Health Science**, Xangai, v. 6, n. 3, p. 321–326, 2017.

JACKLIN, P. B. Temporal changes in home advantage in english football since the second world war: What explains improved away performance? **Journal of sports Sciences**, [*s. l.*], v. 23, n. 7, p. 669–679, 2005.

LEITE, W. S. S. Home advantage: Comparison between the major european football leagues. **Athens Journal of Sports**, Atenas, v. 4, n. 1, p. 65–74, 2017.

MAREK, P.; VÁVRA, F. Comparison of home advantage in european football leagues. **Risks**, [*s. l.*], v. 8, n. 3, p. 87, 2020.

OLIVEIRA, P. V. S. R. et al. Vantagem de jogar em casa na Série A do Campeonato Brasileiro e na copa do brasil. **Revista Brasileira de Futsal e Futebol**, São Paulo, v. 12, n. 48, p. 180–186, 2020.

OURS, J. C. van. A note on artificial pitches and home advantage in dutch professional football. **De Economist**, Países Baixos, v. 167, n. 1, p. 89–103, 2019.

POLLARD, R.; GóMEZ, M. A. Home advantage analysis in different basketball leagues according to team ability. **Iberian Congress on Basketball Research**, [*s. l.*], v. 4, p. 61–64, 2007.

POLLARD, R.; POLLARD, G. Long-term trends in home advantage in professional team sports in north america and england (1876-2003). **Journal of Sports Sciences**, [*s. l.*], v. 23, p. 337–50, 2005.

POLLARD, R.; SILVA, C. D.; MEDEIROS, N. C. Home advantage in football in brazil: Differences between teams and the effects of distance traveled. **Revista Brasileira de Futebol**, Viçosa, v. 1, n. 1, p. 3–10, 01 2008.

R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria, 2021. Disponível em: <https://www.R-project.org/>.

RAMOS, L. F. P.; FERNANDES, H. C.; BATISTA, B. D. O. Modelagem matemática para previsão de jogos de futebol. **Revista Sergipana de Matemática e Educação Matemática**, Aracaju, v. 6, n. 1, p. 46–64, 2021.

SILVA, C. D. et al. Competitive balance in football: A comparative study between brazil and the main european leagues (2003-2016). **Journal of Physical Education**, Maringá, v. 29, 2018.

TILP, M.; THALLER, S. Covid-19 has turned home-advantage into home-disadvantage in the german soccer bundesliga. **Frontiers in Sports and Active Living**, [*s. l.*], v. 2, p. 165, 2020.

APÊNDICE A – Algoritmo para calcular D, DA Limites de predição superior e inferior

Código para utilização no Software R. O código retorna o limite superior e o limite interior do intervalo de predição para uma dada configuração que aconteceu para um time e também retorna a métrica diferença de pontos relativa (d) e a diferença de pontos absoluta (da) observada para o time em um dado ano. O número recomendado de reamostras é de 1.000.000.

A função `teste_reamostragem(·)` tem as seguintes entradas: `vc` é o número de vitórias em casa; `ec` é o número de empates em casa; `dc` é o número de derrotas em casa; `vv` é o número de vitórias como visitante; `ev` é o número de empates como visitante, `dv` é o número de derrotas como visitante, `n_simu` é o número de reamostras obtidas, `percentil_bilateral` é a proporção (em número decimal) de observações que ficaram nas duas caudas somadas e recomenda-se utilizar 0,20; `pto_v` são os pontos por vitória e; `pto_e` são os pontos por empate.

```
#Metrica e teste
teste_reamostragem<-function(vc, ec, dc, vv, ev, dv, n_simu,
percentil_bilateral, pto_v, pto_e)
 {
  p1<-vc+ec+dc ; p2<-p1+1; p3<-p1*2
  conjunto_c<-c(rep(3,vc),rep(1,ec),rep(0,dc))
  conjunto_v<-c(rep(3,vv),rep(1,ev),rep(0,dv))
  conjunto<-c(conjunto_c,conjunto_v)
  reamostras<-matrix(0,n_simu,1)
  for(i in 1:n_simu){
  cr<-sample(conjunto)
  reamostras[i]<-((sum(cr[1:p1])-sum(cr[p2:p3]))*100)/((p3/2)*pto_v)
     }
  ip_linf<-sort(reamostras)[(percentil_bilateral/2)*n_simu]
```

```
  ip_lsup<-sort(reamostras)[(1-percentil_bilateral/2)*n_simu]
  da<-vc*pto_v+ec*pto_e - vv*pto_v - ev*pto_e
  d<- (da*100)/((vc+ec+dc)*pto_v)
  return(list(limite_inferior_predicao = ip_linf,
  limite_superior_predicao = ip_lsup,
  diferenca_pontos_relativa=d, diferenca_pontos_absoluta=da))
  }

#Exemplo para o Nautico em 2013
teste_reamostragem(vc=3, ec=3, dc=13, vv=2, ev=2, dv=15, n_simu
   =1000000,
percentil_bilateral=0.2, pto_v=3, pto_e=1)


#Resultados
#$limite_inferior_predicao
#[1] -14.03509
#
#$limite_superior_predicao
#[1] 14.03509
#
#$diferenca_pontos_relativa
#[1] 7.017544
#
#$diferenca_pontos_absoluta
#[1] 4
```

5 MODELAGEM PROBABILÍSTICA E INFERÊNCIA DO EFEITO DE CASA EM PARTIDAS ESPORTIVAS

Giovani Festa Paludo[§]

Eric Batista Ferreira[¶]

Resumo

Uma vez que foi proposta uma nova métrica para medir o efeito de casa em esportes, faz-se necessário conhecer sua distribuição para realização de inferências. O objetivo do presente estudo foi estudar a distribuição da variável aleatória (v.a.) D, denominada como diferença relativa de pontos e detalhada no Capítulo anterior. A v.a. D é composta pela combinação linear de outras 4 v.a.. Sendo que foi assumido que o problema surge de duas multinomiais independentes, e dada as definições foram obtidas médias e variâncias. Com isso, foram desenvolvidas duas aproximações, uma aproximação pela distribuição normal e outra pela binomial. E em seguida foi feito um estudo de simulação para avaliação de qual aproximação poderia ser utilizada, em seguida foram apresentadas aplicações utilizando a distribuição de D. Os dados simulados de D, tanto partindo de vetores hipotéticos quanto observados no Campeonato Brasileiro, aderiram bem à aproximação binomial em apenas algumas situações específicas. Na grande maioria das situações analisadas, os dados 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 satisfatória que pode ser utilizada para inferências acerca de D.

Palavras-chave: Estudo de simulação, variáveis aleatórias, diferença relativa de pontos (D).

5.1 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 a bastante tempo, mas que não 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; MASSEY; DOWNWARD, 2020), psicologia e medicina do esporte (NEVILL; HOLDER, 1999; MCCARRICK et al., 2021), pesquisa operacional (GOLLER; KRUMER, 2020), estatística (BENZ; LOPEZ, 2021), economia (OURS, 2019; HEGARTY, 2021), riscos do mercado de apostas (MAREK; VÁVRA, 2020), entre outras. Ainda, a HA pode ser encontrada em vários esportes. No futebol, a vantagem de casa

[§]http://lattes.cnpq.br/8897773821703545. Universidade Federal de Alfenas, gfpaludo@gmail.com.
[¶]http://lattes.cnpq.br/9965398009651936. Universidade Federal de Alfenas, eric.ferreira@unifal-mg.edu.br

pode ser obtida de duas maneiras principais: utilizando-se pontos ou; saldo de gols. Como as principais competições de futebol tem por objetivo a conquista de pontos, isso faz com que estudar a vantagem de casa por pontos seja bastante importante.

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; SILVA; MEDEIROS, 2008). Porém, foi desenvolvido uma métrica para medir a vantagem de casa, que não é inflacionada ou não sofre efeitos diretamente relacionados à habilidade do time (PALUDO; FIGUEIREDO; FERREIRA, submetido). 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. Por isso, o objetivo do presente estudo foi de obter: (i) a distribuição da v.a. D; (ii) os estimadores dos parâmetros da distribuição de D; (iii) testes para os parâmetros da distribuição de D e; (iv) ilustrar com dados reais os resultados obtidos.

5.2 MATERIAL E MÉTODOS

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

5.2.1 Variável aleatória D e sua respectiva distribuição de probabilidade

Paludo, Figueiredo e Ferreira (submetido) definiram uma métrica d para obtenção da vantagem de casa, que a partir de então será denominada de variável aleatória D:

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,*

$$D = \frac{Y_c - Y_f}{c_v \times n_c}, \tag{5.1}$$

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.

Quando um time conquista todos os pontos em casa e nenhum ponto fora de casa, o valor que a diferença de pontos 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 $\mathbf{X}$. Seja $\mathbf{X}$ o vetor aleatório tri-dimensional composto pelas variáveis *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,

$$\mathbf{X}^\top = (X_v, X_e, X_d). \tag{5.2}$$

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

$$n = X_v + X_e + X_d. \tag{5.3}$$

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,

$$\begin{aligned} n_c &= n_f, \\ n &= n_c + n_f = 2n_c, \end{aligned}$$

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

$$\mathbf{X} = \begin{pmatrix} X_v = X_{vc} + X_{vf} \\ X_e = X_{ec} + X_{ef} \\ X_s = X_{dc} + X_{df} \end{pmatrix} = \begin{pmatrix} X_{vc} \\ X_{ec} \\ X_{dc} \end{pmatrix} + \begin{pmatrix} X_{vf} \\ X_{ef} \\ X_{df} \end{pmatrix} = \mathbf{X}_c + \mathbf{X}_f. \tag{5.4}$$

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 (5.3), 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:

$$\mathbf{X}_c \sim Multi(n_c, p_{vc}, p_{ec}, p_{dc}), \tag{5.5}$$
$$\mathbf{X}_f \sim Multi(n_f, p_{vf}, p_{ef}, p_{df}), \tag{5.6}$$

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 $\{i,j\} \neq \{i',j'\}$

$$E[X_{ij}] = n_i p_{ij}, \tag{5.7}$$
$$Var[X_{ij}] = n_i p_{ij}(1 - p_{ij}), \tag{5.8}$$
$$Cov[X_{ij},X_{i'j'}] = -n_i p_{ij} p_{i'j'}. \tag{5.9}$$

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[**]

$$X_{vc} \sim Bin(n_c,p_{vc}), \tag{5.10}$$
$$X_{ec} \sim Bin(n_c - x_{vc},p_{ec}). \tag{5.11}$$

Com base nisso, considere a variável aleatória *número de pontos ganhos* (Y), definida

[‖] Dados os números de vitórias e empates, o número de derrotas está automaticamente determinado.
[**] O mesmo para fora de casa.

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):

$$Y_c = c_v X_{vc} + c_e X_{ec}, \tag{5.12}$$

$$Y_f = c_v X_{vf} + c_e X_{ef}. \tag{5.13}$$

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

$$E[Y_c] = c_v n_c \left(p_{vc} + \frac{c_e}{c_v} p_{ec} \right), \tag{5.14}$$

$$E[Y_f] = c_v n_f \left(p_{vf} + \frac{c_e}{c_v} p_{ef} \right). \tag{5.15}$$

Considere a seguinte reparametrização:

$$a_c = p_{vc} + \frac{c_e}{c_v} p_{ec}, \tag{5.16}$$

$$a_f = p_{vf} + \frac{c_e}{c_v} p_{ef}. \tag{5.17}$$

A partir disso, temos que (5.14) e (5.15) podem ser reescritos como:

$$E[Y_c] = c_v n_c a_c, \tag{5.18}$$

$$E[Y_f] = c_v n_f a_f. \tag{5.19}$$

Além disso, utilizando (5.8), (5.9), (5.12) e (5.13), temos que as variâncias de Y_c e Y_f

são

$$
\begin{aligned}
Var[Y_c] &= Var[c_v X_{vc} + c_e X_{ec}] \\
&= c_v^2 Var[X_{vc}] + c_e^2 Var[X_{ec}] + 2c_v c_e Cov[X_{vc}, X_{ec}] \\
&= c_v^2 n_c p_{vc}(1 - p_{vc}) + c_e^2 n_c p_{ec}(1 - p_{ec}) - 2c_v c_e n_c p_{vc} p_{ec} \\
&= n_c c_v \left[c_v p_{vc}(1 - p_{vc}) + \frac{c_e^2}{c_v} p_{ec}(1 - p_{ec}) - 2c_e p_{vc} p_{ec} \right] \qquad (5.20) \\
Var[Y_f] &= n_f c_v \left[c_v p_{vf}(1 - p_{vf}) + \frac{c_e^2}{c_v} p_{ef}(1 - p_{ef}) - 2c_e p_{vf} p_{ef} \right] \qquad (5.21)
\end{aligned}
$$

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 e Punnen (2001) e Butler e 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.

5.2.2 Estudo de simulação

Para avaliar a adequabilidade dos dados às 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 3 e 4). Os vetores hipotéticos foram escolhidos de maneira que pudessem ser comparadas diferentes situações que estão descritas na Tabela 3, como por exemplo vetores $\mathbf{p}_c = \mathbf{p}_f$, $p_{vc} > p_{vf}$, entre outros. 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 resultados. O primeiro resultado foi com base na seleção de 10 combinações de vetores de 10 participações de times de maneira que diferentes situações

que aconteceram no Campeonato Brasileiro fossem representadas (considerando-se as edições de 2006 à 2021). Desses 10, 5 foram escolhidos de maneira que fossem incluídas diferentes combinações de em casa e os outros 5 para incluir diferentes situações fora de casa. E por sua vez, o segundo resultado consistiu na utilização de todos os parâmetros observados entre 2006 e 2019 no Campeonato Brasileiro. Isto é, foram simuladas 278[††] combinações de vetores e as linhas foram sobrepostas num gráfico.

Cabe ressaltar que o vetor de probabilidade de 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 ganhou 10 partidas em 19 que disputou em casa, ficaria com um $p = 0{,}526$.

Tabela 3 – Vetores p_c e p_f hipotéticos utilizados nas simulações avaliação das duas aproximações. Note que * foi o único caso no presente Capítulo que utilizou uma informação da edição de 2021 do Campeonato Brasileiro.

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: Próprios autores.

[††]Foram 280 participações em 14 anos, porém, 2 participações não puderam ser utilizadas.

Tabela 4 – Vetores p_c e p_f com dados observados no Campeonato Brasileiro de Futebol e utilizados nas simulações para avaliação das duas aproximações. Note que * foi o único caso no presente Capítulo que utilizou uma informação da edição de 2021 do Campeonato Brasileiro.

Número	Descrição	$\mathbf{p}_c^\top$	$\mathbf{p}_f^\top$
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: Próprios autores.

A partir de cada combinação dos dois vetores de probabilidade foram calculadas a média, 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, gerou-se uma amostra de dados de tamanho variável 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()`. As funções utilizadas e um *script* em linguagem R utilizado foi incluído como apêndice ao final desse Capítulo.

Para verificar se a cada iteração de um tamanho amostral s ou de uma situação (nas 28 situações descritas na Tabela 3 e 4), os dados se ajustavam a cada uma das 2 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:

$\mathscr{H}_0$: se os dados podiam ser considerados provenientes da distribuição em questão e;

$\mathscr{H}_1$: se os dados não podiam ser considerados como provenientes da distribuição em questão.

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

[‡‡]Exceto no gráfico com as 278 curvas, que foi construído com 300 simulações e $s = 20, 40, ..., 300$.

que $\mathscr{H}_0$ 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.

5.2.3 Aplicações

Utilizando dados de campeonatos de futebol, 5 aplicações serão apresentadas. A saber: a comparação entre dois times que participaram de um mesmo campeonato; comparação de múltiplos times; intervalo de confiança para a média populacional de D; aplicação de D para comparação de ligas e; exemplo para obtenção da média amostral de D para campeonatos não balanceados.

Dados utilizados

As aplicações tiveram como ideia central representar diferentes possibilidades de uso da distribuição de D, e portanto os dados utilizados variaram entre as aplicações. As aplicações 1 e 2 se basearam em dados do Campeonato Brasileiro que foram obtidos através do website <www.soccerway.com> que já foi utilizado nos estudos de Pollard, Silva e Medeiros (2008) e Silva et al. (2018). Ressalta-se que no presente Capítulo foram utilizados dados de 2006 à 2019, diferente do Capítulo anterior.

Já nas aplicações 3 e 4, além dos dados da Série A do Brasil, também foram utilizados dados da Série A do Brasil, *La Liga* da Espanha, *Premier League* da Inglaterra e da *Serie A* da Itália. Sendo que os dados das ligas da Espanha, Itália e Inglaterra foram obtidos através do pacote do Software R `engsoccerdata` (CURLEY, 2020). E por último, na aplicação 5 foi utilizado os resultados de um time ao final do liga nacional da Argentina de 2018/2019. 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.

Etapas para realização do teste de hipóteses das aplicações 1, 2 e 3

Nas aplicações 1, 2 e 3 foi utilizado o seguinte procedimento 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:

$$\mathscr{H}_0 : \mu_1 = \mu_2$$
$$\mathscr{H}_1 : \mu_1 \neq \mu_2.$$

O procedimento para a realização do teste de hipóteses acima envolveu as seguintes etapas:

1. Obtenção das médias amostrais ($\bar{D}$) e o desvio padrão amostral (S^2);

2. Aplicação do teste F para verificar se as variâncias são homogêneas, que segundo Bolfarine e Sandoval (2001), a estatística do teste F para comparação de variâncias pode ser escrita como:
$$F = \frac{(n_2 - 1)S_2^2/(n_2 - 1)}{(n_1 - 1)S_1^2/(n_1 - 1)} \sim F_{n_2-1,n_1-1}; \tag{5.22}$$

3. Se as variâncias são homogêneas

   (a) aplica-se o teste t para variâncias homogêneas, dado por (ZAR, 2010),
   $$t = \frac{\bar{D}_1 - \bar{D}_2}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}} \sim t_{n_1+n_2-2}$$
   onde,
   $$S_p^2 = \frac{SQ_1^2 + SQ_2^2}{(n_1 - 1) + (n_2 - 1)}$$

   (b) se as variâncias são heterogêneas, utilizar a estatística (ZAR, 2010):
   $$t = \frac{\bar{D}_1 - \bar{D}_2}{\sqrt{S_{\bar{D}_1}^2 + S_{\bar{D}_2}^2}},$$
   onde, $S_{\bar{D}_1}^2$ é a variância amostral da média $\bar{D}_1$ e $S_{\bar{D}_2}^2$ é a variância amostral da média $\bar{D}_2$. A variância amostral da média é dada por $S_{\bar{D}_1}^2 = S^2/n_1$. Ainda, para utilização da distribuição t, os graus de liberdade podem ser aproximados a partir de (ZAR, 2010):
   $$gl' = \frac{(S_{\bar{D}_1}^2 + S_{\bar{D}_2}^2)^2}{\frac{(S_{\bar{D}_1}^2)^2}{n_1-1} + \frac{(S_{\bar{D}_1}^2)^2}{n_2-1}}, \tag{5.23}$$

onde gl' são os graus de liberdade aproximados.

4. Realizar a decisão e conclusão sobre o resultado do teste de hipóteses.

Aplicações 4 e 5: Intervalo de confiança e estimação da média para dados de campeonatos não balanceados

Para a construção de estimativas intervalares para os times, foi utilizada a distribuição t e um nível de confiança de 95%.

Já para a obtenção da média de 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, $\hat{p} = N/n$. De posse das estimativas dos vetores $\mathbf{p}_c$ e $\mathbf{p}_f$, calcula-se:

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

e

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

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 obtém-se $\bar{D}$ para um time quando o campeonato é desbalanceado.

5.3 RESULTADOS E DISCUSSÃO

5.3.1 Resultados metodológicos

Aproximação Binomial para Y

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 (5.20) e (5.21) 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 (5.18) e (5.19) — e variâncias aproximadas, para garantir a exigência

$$Var[\cdot] = E[\cdot](1-p) \tag{5.24}$$

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

$$Var[Y_c] = n_c c_v a_c(1-a_c) \tag{5.25}$$

$$Var[Y_f] = n_f c_v a_f(1-a_f). \tag{5.26}$$

Agora, precisamos estabelecer o número máximo de sucessos para as variáveis Y. 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,

$$Y_c \dot{\sim} Bin\left(c_v n_c,\ a_c\right) \tag{5.27}$$

$$Y_f \dot{\sim} Bin\left(c_v n_f,\ a_f\right). \tag{5.28}$$

Aproximação Normal para Y

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 - (5.18) e (5.19) - e variâncias aproximadas - (5.25) e (5.26). Sendo assim, temos que:

$$Y_c \dot{\sim} N\left(n_c c_v a_c,\ n_c c_v a_c(1-a_c)\right) \tag{5.29}$$

$$Y_f \dot{\sim} N\left(n_f c_v a_f,\ n_f c_v a_f(1-a_f)\right). \tag{5.30}$$

E no segundo, utilizamos esperanças exatas - (5.18) e (5.19) - e variâncias exatas - (5.20) e (5.21). Sendo assim, temos que:

$$Y_c \dot{\sim} N\left(n_c c_v a_c,\ n_c c_v\left[c_v p_{vc}(1-p_{vc}) + \frac{c_e^2}{c_v} p_{ec}(1-p_{ec}) - 2c_e p_{vc} p_{ec}\right]\right) \tag{5.31}$$

$$Y_f \dot{\sim} N\left(n_f c_v a_f,\ n_f c_v\left[c_v p_{vf}(1-p_{vf}) + \frac{c_e^2}{c_v} p_{ef}(1-p_{ef}) - 2c_e p_{vf} p_{ef}\right]\right). \tag{5.32}$$

Aproximação Binomial para 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

$$D = \frac{Y_c - Y_f}{c_v n_c} \tag{5.33}$$

Assim, a variável aleatória D tem esperança

$$\begin{aligned} E[D] &= E\left[\frac{Y_c - Y_f}{c_v n_c}\right] \\ &= \frac{1}{c_v n_c}\left(E[Y_c] - E[Y_f]\right) \\ &= \frac{1}{c_v n_c}\left(c_v E[X_{vc}] + c_e E[X_{ec}] - c_v E[X_{vf}] - c_e E[X_{ef}]\right) \\ &= \frac{1}{c_v n_c}\left(c_v n_c p_{vc} + c_e n_c p_{ec} - c_v n_f p_{vf} - c_e n_f p_{ef}\right) \\ &\overset{n_c = n_f}{=} \frac{1}{c_v n_c}\left(c_v n_c p_{vc} + c_e n_c p_{ec} - c_v n_c p_{vf} - c_e n_c p_{ef}\right) \\ &= p_{vc} + \frac{c_e}{c_v} p_{ec} - p_{vf} - \frac{c_e}{c_v} p_{ef} \\ &= a_c - a_f. \end{aligned} \tag{5.34}$$

(5.35)

E a variância de D fica

$$\begin{aligned}
\sigma_D^2 &= Var[D] \\
&= Var\left[\frac{Y_c - Y_f}{c_v n_c}\right] \\
&= \frac{Var\left[Y_c\right] + Var\left[Y_f\right] - 2\cancelto{0}{Cov[Y_c,Y_f]}}{(c_v n_c)^2} \\
&= \frac{Var\left[c_v X_{vc} + c_e X_{ec}\right] + Var\left[c_v X_{vf} + c_e X_{ef}\right]}{(c_v n_c)^2} \\
&= \frac{c_v^2 Var[X_{vc}] + c_e^2 Var[X_{ec}] + 2Cov[X_{vc},X_{ec}] + c_v^2 Var[X_{vf}] + c_e^2 Var[X_{ef}] + 2Cov[X_{vf},X_{ef}]}{(c_v n_c)^2} \\
&= \frac{c_v^2 n_c p_{vc}(1-p_{vc}) + c_e^2 n_c p_{ec}(1-p_{ec}) + c_v^2 n_c p_{vf}(1-p_{vf}) + c_e^2 n_c p_{ef}(1-p_{ef})}{(c_v n_c)^2} \\
&= \frac{1}{c_v n_c}\left[c_v p_{vc}(1-p_{vc}) + \frac{c_e^2}{c_v} p_{ec}(1-p_{ec}) + c_v p_{vf}(1-p_{vf}) + \frac{c_e^2}{c_v} p_{ef}(1-p_{ef})\right]. \qquad (5.36)
\end{aligned}$$

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 a abordagem "b", descrita no final do Capítulo 3.

Assim, 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 (5.35) e aproximamos a variância (5.36), de tal forma que a exigência (5.24) da binomial seja satisfeita.

Assim encontramos a probabilidade de sucesso

$$p_1^* = \frac{E[D]}{c_v n} = \frac{a_c - a_f}{c_v n} \qquad (5.37)$$

e

$$\begin{aligned}
Var[D]_1^* &= c_v n \left(\frac{a_c - a_f}{c_v n}\right)\left(1 - \frac{a_c - a_f}{c_v n}\right) \\
&= (a_c - a_f)\left(1 - \frac{a_c - a_f}{c_v n}\right). \qquad (5.38)
\end{aligned}$$

No entanto, note que (5.37) precisa variar entre 0 e 1, o que não acontece nessa forma.

Sendo assim, sugere-se a reparametrização a seguir para garantir essa condição:

$$\begin{aligned} p_2^* &= 0{,}5 + c_v n_c p \\ &= 0{,}5 + \frac{a_c - a_f}{2} \\ &= \frac{1 + a_c - a_f}{2}. \end{aligned} \tag{5.39}$$

Isso implica a seguinte mudança na variância (5.38):

$$\begin{aligned} Var[D]_2^* &= c_v n p^*(1 - p^*) \\ &= c_v n \frac{1 + a_c - a_f}{2}\left(1 - \frac{1 + a_c - a_f}{2}\right) \\ &= c_v n \left(\frac{1 + a_c - a_f}{2}\right)\left(\frac{1 - a_c + a_f}{2}\right) \\ &= \frac{c_v n}{4}(1 + a_c - a_f)(1 - a_c + a_f). \end{aligned} \tag{5.40}$$

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

$$D \dot{\sim} Bin\left(c_v n,\ \frac{1 + a_c - a_f}{2}\right), \tag{5.41}$$

sendo que está será denominada como aproximação pela binomial.

Aproximação Normal para D

Apresenta-se uma maneira de aproximar a distribuição de D pela Normal. Sendo que nessa aproximação utiliza-se a esperança exata (5.35) e a variância exata (5.36). Dessa forma,

$$D \dot{\sim} N\left(a_c - a_f,\ \sigma_D^2\right) \tag{5.42}$$

Assim, está será denominada de aproximação pela normal.

5.3.2 Resultados do estudo de simulação

O estudo de simulação foi realizado para verificar qual das aproximações seria utilizada como distribuição de D. Sendo que, para o estudo de simulação foram gerados dados, isto é, valores hipotéticos de D, gerados por meio de computação e com parâmetros previamente definidos e foram testados se esses dados aderiam ou não à cada uma dessas aproximações

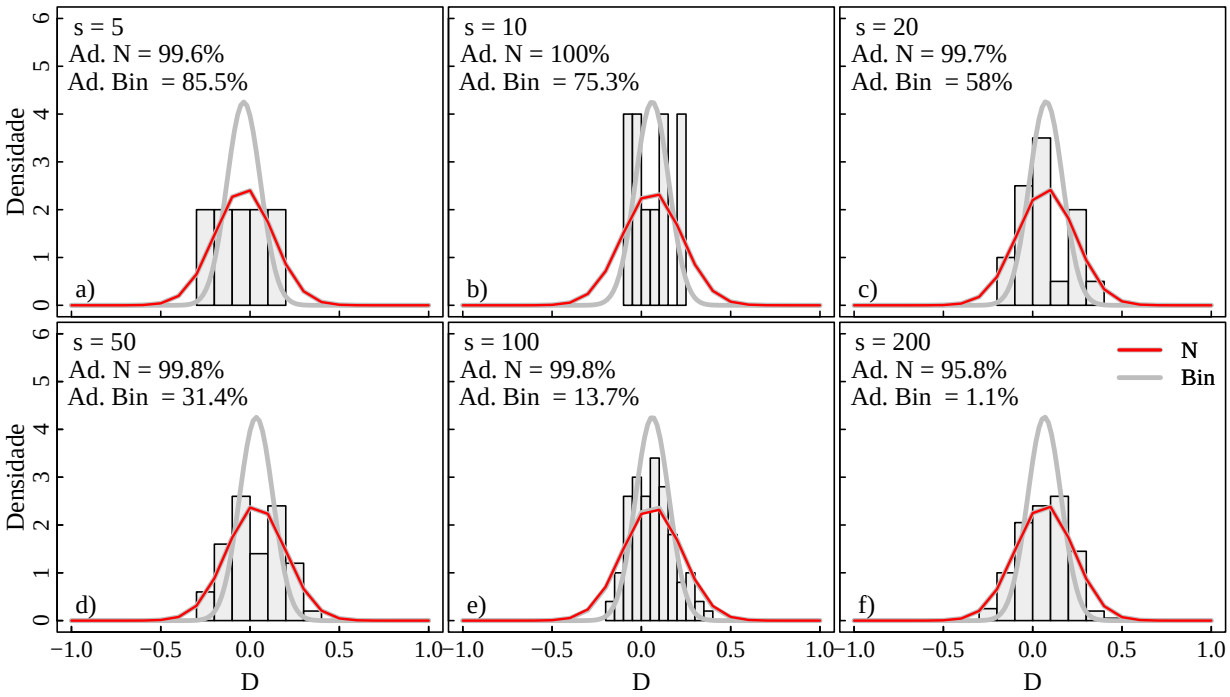
através do teste de aderência de qui-quadrado. Sendo que foram apresentadas as 5 figuras que trazem os resultados do estudo de simulação, das quais 3 delas foram baseadas em situações hipotéticas e 2 com situações observadas no Campeonato Brasileiro.

O primeiro resultado, 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 5). Sendo que o primeiro gráfico (FIGURA 5a) corresponde ao tamanho amostral de $s = 5$, indo até $s = 200$ no sexto gráfico (FIGURA 5f), 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 3).

Ainda, para a obtenção dos valores de aderências em cada gráfico, foram utilizadas 1000 simulações, sendo que uma aderência de $86{,}2\%$ (FIGURA 5a) indicou que a $\mathscr{H}_0$ não foi rejeitada em $86{,}2\%$ das simulações, isto é, que em $86{,}2\%$ 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 $1{,}9\%$ no tamanho amostral de $s = 200$. 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 5).

Destaca-se que em todos os gráficos da Figura 5, 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 5 – Histograma dos valores simulados de D e as curvas das distribuições (linhas cinzas e vermelhas). No qual 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 $\mathbf{p}_f = (0{,}4; 0{,}4; 0{,}2)^\top$.



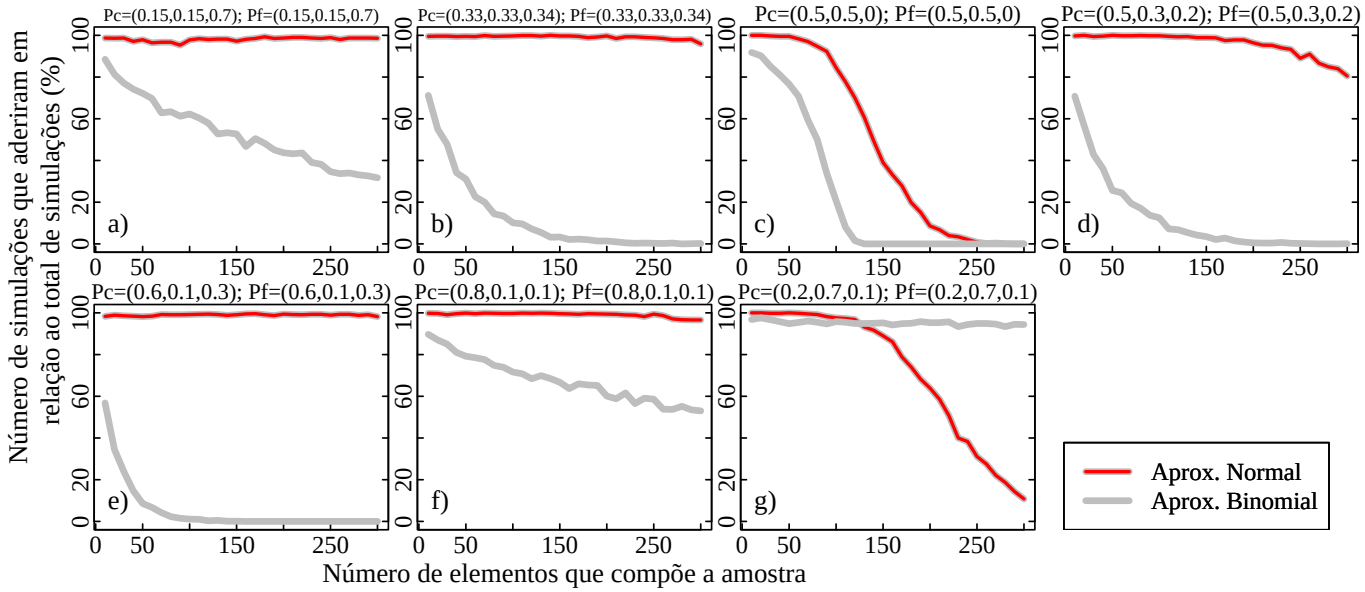
Fonte: Próprios autores.

Enquanto que a Figura 5 trouxe resultados para uma única combinação de vetores $\mathbf{p}_c$ e $\mathbf{p}_f$, nas Figuras 6,7 e 8, cada gráfico corresponde a uma combinação diferente dos $\mathbf{p}_c$ e $\mathbf{p}_f$. Ainda, as Figuras 6,7 e 8 não foram feitas para um único tamanho amostral, mas sim para uma sequência de tamanhos amostrais s, ou seja, $s = 10, 20, ..., 300$. E, enquanto que a Figura 8 e 9 foi baseada em vetores $\mathbf{p}_c$ e $\mathbf{p}_f$ observados no Campeonato Brasileiro, as Figuras 5, 6 e 7 foram baseadas em valores hipotéticos para vetores de parâmetros $\mathbf{p}_c$ e $\mathbf{p}_f$.

A Figura 6 se baseou em parâmetros hipotéticos sendo nesse resultado, o vetor $\mathbf{p}_c$ foi sempre igual $\mathbf{p}_f$. Como resultado, em apenas uma combinação os valores de aderência da binomial foram maiores que a normal, que foi na Figura 6g 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. Então, 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 6f), pois nesse caso D será constituído principalmente da diferença de duas v.a. binomiais independentes, que também é binomial.

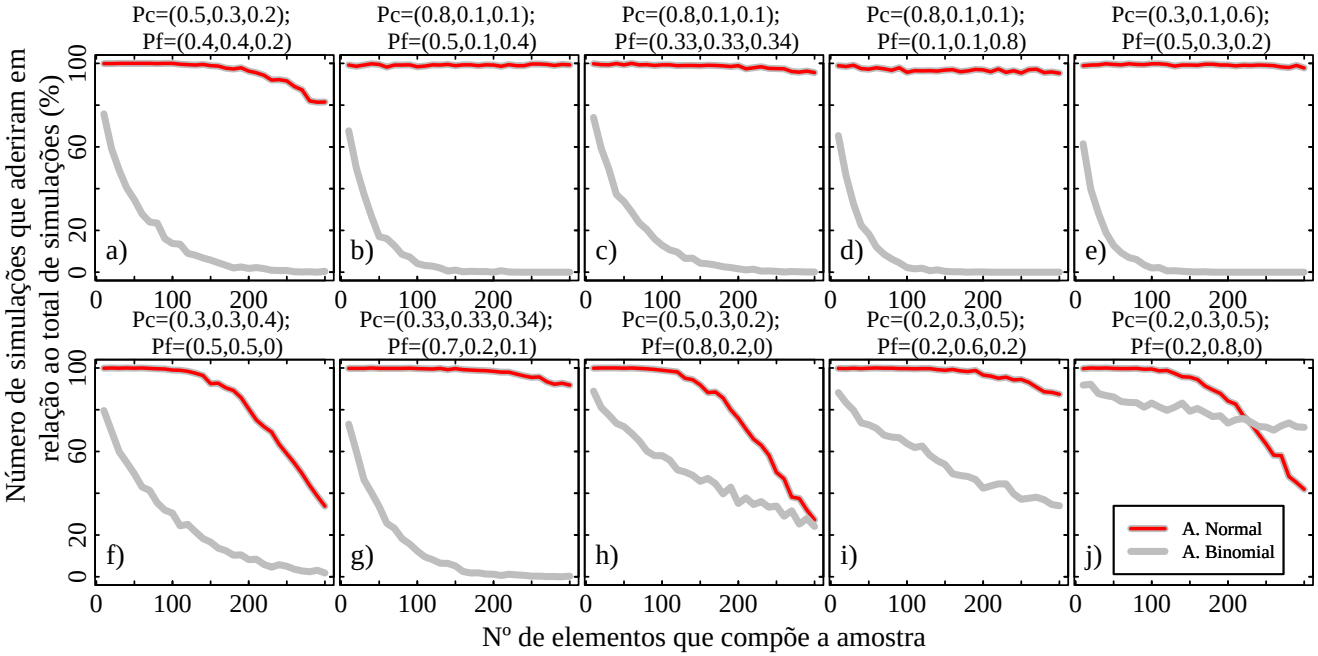
Figura 6 – Grau de aderência de dados gerados por vetores hipotéticos tais que $\mathbf{p}_c = \mathbf{p}_f$ em relação à duas aproximações (Normal e Binomial). Foram 1000 iterações de geração aleatória de D em amostras de tamanho $s = 10, 20, 30, ..., 300$. Sendo que 0% representa nenhuma iteração com aderência e 100% representa que todas as iterações aderiram à distribuição. Ainda, os vetores $\mathbf{p}_c$ e $\mathbf{p}_f$ estão explicitados logo acima à area gráfica.



Fonte: Próprios autores.

Nas 10 situações ou gráficos apresentados na Figura 7, 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 $p_{ec} < p_{ef}$, em todas as outras a normal apresentou maior aderência que a aproximação binomial e; (iii) a aproximação binomial teve um comportamento parecido e com pouca aderência nas Figuras 7a,b,c,d,e,f,g. Porém teve maior aderência quando a p_{ef} foi alta (FIGURA 7j com $p_{ef} = 0{,}8$ e FIGURA 7i, com $p_{ef} = 0{,}6$) ou quando a p_{vf}=0,8 (FIGURA 7h).

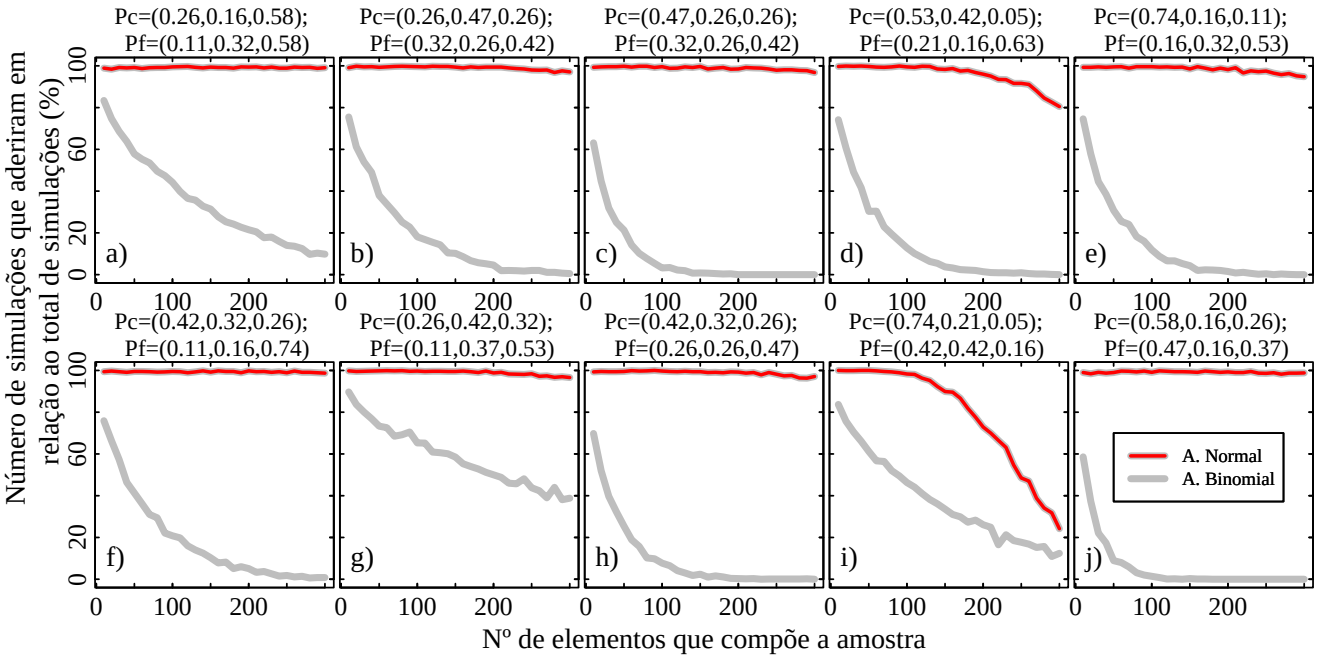
Figura 7 – Grau de aderência de dados gerados por vetores hipotéticos com $\mathbf{p}_c \neq \mathbf{p}_f$ em relação à duas aproximações (normal e bionomial). Sendo que 0% representa nenhuma iteração com aderência e 100% representa que todas as iterações aderiram à distribuição) de 1000 iterações de 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: Próprios autores.

Já nas situações com parâmetros $\mathbf{p}_c$ e $\mathbf{p}_f$ obtidos no Campeonato Brasileiro (FIGURA 8), 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 8g,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 teve sempre valores superiores nas 10 combinações consideradas.

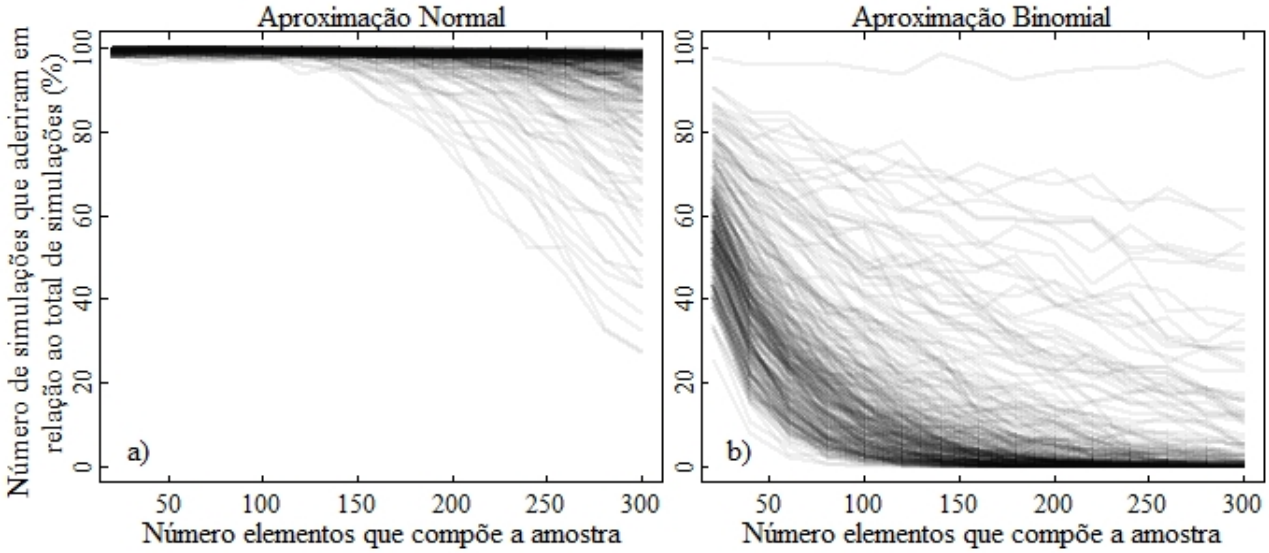
Figura 8 – Grau de aderência de dados gerados a partir de vetores de probabilidade obtidos no Campeonato Brasileiro de Futebol em relação à duas aproximações (normal e binomial). Sendo que 0% representa nenhuma iteração com aderência e 100% representa que todas as iterações aderiram à distribuição. Foram 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: Próprios autores.

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 9 (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 9 – 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 300 iterações de geração aleatória de D em amostras de tamanho $s = 20, 40, 60, ..., 300$.



Fonte: Próprios autores.

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 9a) e outro para a binomial (FIGURA 9b), 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 $\mathbf{p}_c$ igualmente pequenos, quando a $p_{vc} = p_{vf}$ era próximo de 1 ou quando a $\mathbf{p}_e > p_v$; (ii) a aproximação binomial só teve aderência maior que a aproximação normal em algumas classes de tamanho de amostra quando $\mathbf{p}_e > p_v$, 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 o $\bar{X}$ e o S^2 para inferências.

5.3.3 Resultados das aplicações

A aplicação 1 consistiu em um exemplo da utilização do teste de hipóteses para um par de times que disputa as mesma competição, que no caso foi considerado o Internacional e o Flamengo em todas as participações no Campeonato Brasileiro de Futebol Série A entre 2006 a 2020.

$$\mathscr{H}_0 : \mu_{int} = \mu_{fla}$$
$$\mathscr{H}_1 : \mu_{int} \neq \mu_{fla}$$

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

$$d_{int} = \{0{,}053; 0{,}386; 0{,}561; 0{,}263; 0{,}175; 0{,}211; 0{,}175; 0{,}140; 0{,}368; 0{,}526; 0{,}368; 0{,}404; 0{,}368\}$$

e

$$d_{fla} = \{0{,}32; 0{,}33; 0{,}18; 0{,}26; 0{,}18; 0{,}12; 0{,}23; 0{,}23; 0{,}32; 0{,}09; 0{,}19; 0{,}28; 0{,}28; 0{,}28\}$$

Estes dois conjuntos geraram as médias amostrais iguais a $\bar{d}_{int} = 0{,}3077; S_{int} = 0{,}15221$ e $\bar{d}_{fla} = 0{,}2356; S_{fla} = 0{,}07498$.

Inicialmente precisamos verificar se as variâncias são homogêneas através do teste de F

$$F = \frac{0{,}07498^2}{0{,}15221^2} = 4{,}1212 \tag{5.43}$$

Como o valor tabelado da estatística $F_{13,12,\alpha=0,05} = 3{,}15$ é menor que o valor calculado, rejeita-se a hipótese nula de que as variâncias são iguais.

Assim, utilizou-se a estatística t para variâncias heterogêneas. Inicialmente, obteve-se a variância amostral da média ($S^2_{\bar{D}_2}$),

$$S^2_{\bar{D}_2} = \frac{S_1^2}{n_1} = \frac{0{,}023167}{13} = 0{,}0017821$$

$$S^2_{\bar{D}_2} = \frac{S_2^2}{n_1} = \frac{0{,}0056213}{14} = 0{,}00040152$$

onde, $S^2_{\bar{D}_1}$ é a variância amostral da média $\bar{D}_1$ e $S^2_{\bar{D}_2}$ é a variância amostral da média $\bar{D}_2$. A variância amostral da média foi dada por $S^2_{\bar{D}_1} = S^2/n_1$. Em seguida, obtém-se o t calculado,

$$t = \frac{\bar{D}_1 - \bar{D}_2}{\sqrt{S^2_{\bar{D}_1} + S^2_{\bar{D}_2}}} = \frac{0{,}3077 - 0{,}2356}{\sqrt{0{,}00178208 + 0{,}00040152}} = 1{,}543013.$$

Ainda, para utilização da distribuição t, os graus de liberdade podem ser aproximados a partir de (ZAR, 2010):

$$gl' = \frac{(S^2_{\bar{D}_1} + S^2_{\bar{D}_2})^2}{\frac{(S^2_{\bar{D}_1})^2}{n_1 - 1} + \frac{(S^2_{\bar{D}_1})^2}{n_2 - 1}} = \frac{0{,}023167 + 0{,}0056213}{\frac{0{,}0017821^2}{13-1} + \frac{0{,}00040152^2}{14-1}} = 17{,}21023, \tag{5.44}$$

onde gl' são os graus de liberdade aproximados.

Para um nível de significância $\alpha = 0{,}05$ e $gl = 17{,}21023$, encontram-se os valores de $t_{critico} = \pm 2{,}104571$. Portanto, não rejeita-se a hipótese nula de que não há diferença entre os times.

Já a aplicação 2 consistiu na repetição do procedimento da aplicação 1 para todos os times do Campeonato Brasileiro com mais de 3 participações entre 2006 e 2020. Assim, foi obtida a Tabela 5 com os valores-p do teste para as comparações par a par. Ressalta-se que foi utilizado o teste t para comparações de médias 2 a 2. Quando realizam-se comparações múltiplas, comete-se mais erro do tipo I, fato que não foi considerado no presente estudo. Isto é, o nível de significância foi de 5% nas comparações par a par. E sabe-se que o nível de significância global é superior a 5% e não foi levado em consideração.

Tabela 5 – Valores-p do teste t para cada par de times que disputaram o Campeonato Brasileiro de 2006 a 2020 com mais de 3 participações.

	CAP	ACG	CAM	AVA	BAH	BOT	CEA	CHA	COR	CFC	CRU	FIG	FLA	FLU	GOI	GRE	INT	NAU	PAL	PON	SAN	SAO	SPT	VAS	VIT
CAP	1,00	0,01	0,18	0,03	0,02	0,02	0,55	0,08	0,07	0,69	0,02	0,01	0,01	0,00	0,08	0,35	0,31	0,54	0,03	0,36	0,43	0,02	0,80	0,00	0,11
ACG	0,01	1,00	0,07	0,35	0,40	0,05	0,04	0,29	0,04	0,01	0,16	0,33	0,03	0,34	0,05	0,02	0,04	0,13	0,04	0,07	0,01	0,10	0,00	0,36	0,40
CAM	0,18	0,07	1,00	0,24	0,23	0,45	0,69	0,46	0,76	0,35	0,32	0,17	0,30	0,05	0,70	0,56	0,72	0,74	0,51	0,84	0,48	0,39	0,26	0,12	0,53
AVA	0,03	0,35	0,24	1,00	0,99	0,37	0,14	0,73	0,22	0,04	0,62	0,95	0,42	0,77	0,28	0,07	0,14	0,26	0,31	0,20	0,06	0,50	0,02	0,91	0,75
BAH	0,02	0,40	0,23	0,99	1,00	0,38	0,17	0,73	0,23	0,04	0,61	0,97	0,54	0,79	0,30	0,07	0,14	0,26	0,32	0,22	0,06	0,49	0,02	0,92	0,74
BOT	0,02	0,05	0,45	0,37	0,38	1,00	0,22	0,76	0,58	0,05	0,72	0,29	0,74	0,12	0,71	0,13	0,24	0,34	0,89	0,34	0,09	0,86	0,02	0,24	0,87
CEA	0,55	0,04	0,69	0,14	0,17	0,22	1,00	0,34	0,43	0,73	0,23	0,10	0,08	0,04	0,36	0,98	0,88	0,97	0,24	0,82	0,94	0,24	0,59	0,09	0,49
CHA	0,08	0,29	0,46	0,73	0,73	0,76	0,34	1,00	0,51	0,13	0,97	0,66	0,92	0,46	0,61	0,20	0,31	0,44	0,68	0,42	0,17	0,86	0,09	0,61	0,97
COR	0,07	0,04	0,76	0,22	0,23	0,58	0,43	0,51	1,00	0,15	0,41	0,16	0,35	0,05	0,88	0,31	0,48	0,54	0,67	0,60	0,24	0,50	0,09	0,12	0,68
CFC	0,69	0,01	0,35	0,04	0,04	0,05	0,73	0,13	0,15	1,00	0,05	0,02	0,01	0,00	0,14	0,63	0,55	0,73	0,06	0,52	0,73	0,05	0,86	0,01	0,22
CRU	0,02	0,16	0,32	0,62	0,61	0,72	0,23	0,97	0,41	0,05	1,00	0,53	0,90	0,30	0,53	0,09	0,17	0,29	0,63	0,29	0,07	0,85	0,03	0,46	0,98
FIG	0,01	0,33	0,17	0,95	0,97	0,29	0,10	0,66	0,16	0,02	0,53	1,00	0,33	0,80	0,21	0,04	0,09	0,19	0,24	0,14	0,03	0,41	0,01	0,95	0,69
FLA	0,01	0,03	0,30	0,42	0,54	0,74	0,08	0,92	0,35	0,01	0,90	0,33	1,00	0,14	0,46	0,05	0,14	0,42	0,62	0,18	0,04	0,92	0,00	0,29	0,97
FLU	0,00	0,34	0,05	0,77	0,79	0,12	0,04	0,46	0,05	0,00	0,30	0,80	0,14	1,00	0,09	0,01	0,02	0,07	0,09	0,05	0,00	0,20	0,00	0,86	0,58
GOI	0,08	0,05	0,70	0,28	0,30	0,71	0,36	0,61	0,88	0,14	0,53	0,21	0,46	0,09	1,00	0,30	0,45	0,52	0,80	0,55	0,24	0,62	0,08	0,19	0,73
GRE	0,35	0,02	0,56	0,07	0,07	0,13	0,98	0,20	0,31	0,63	0,09	0,04	0,05	0,01	0,30	1,00	0,85	0,97	0,15	0,80	0,88	0,11	0,49	0,02	0,28
INT	0,31	0,04	0,72	0,14	0,14	0,24	0,88	0,31	0,48	0,55	0,17	0,09	0,14	0,02	0,45	0,85	1,00	0,93	0,28	0,93	0,74	0,20	0,43	0,06	0,38
NAU	0,54	0,13	0,74	0,26	0,26	0,34	0,97	0,44	0,54	0,73	0,29	0,19	0,42	0,07	0,52	0,97	0,93	1,00	0,53	0,89	0,90	0,32	0,64	0,15	0,52
PAL	0,03	0,04	0,51	0,31	0,32	0,89	0,24	0,68	0,67	0,06	0,63	0,24	0,62	0,09	0,80	0,15	0,28	0,53	1,00	0,38	0,11	0,76	0,03	0,20	0,82
PON	0,36	0,07	0,84	0,20	0,22	0,34	0,82	0,42	0,60	0,52	0,29	0,14	0,18	0,05	0,55	0,80	0,93	0,89	0,38	1,00	0,71	0,32	0,41	0,12	0,53
SAN	0,43	0,01	0,48	0,06	0,06	0,09	0,94	0,17	0,24	0,73	0,07	0,03	0,04	0,00	0,24	0,88	0,74	0,90	0,11	0,71	1,00	0,08	0,59	0,02	0,24
SAO	0,02	0,10	0,39	0,50	0,49	0,86	0,24	0,86	0,50	0,05	0,85	0,41	0,92	0,20	0,62	0,11	0,20	0,32	0,76	0,32	0,08	1,00	0,03	0,34	0,94
SPT	0,80	0,00	0,26	0,02	0,02	0,02	0,59	0,09	0,09	0,86	0,03	0,01	0,00	0,00	0,08	0,49	0,43	0,64	0,03	0,41	0,59	0,03	1,00	0,01	0,20
VAS	0,00	0,36	0,12	0,91	0,92	0,24	0,09	0,61	0,12	0,01	0,46	0,95	0,29	0,86	0,19	0,02	0,06	0,15	0,20	0,12	0,02	0,34	0,01	1,00	0,62
VIT	0,11	0,40	0,53	0,75	0,74	0,87	0,49	0,97	0,68	0,22	0,98	0,69	0,97	0,58	0,73	0,28	0,38	0,52	0,82	0,53	0,24	0,94	0,20	0,62	1,00

Fonte: Próprios autores.

A aplicação 3 baseou-se na comparação entre quatro ligas principais de quatro países. Comparou-se se haviam diferenças entre as ligas: Série A brasileira, *Premier League* inglesa, *Serie A* italiana e *La Liga* espanhola. A média 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 6). 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 10). 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 fica superior a 5%. No presente estudo, o Brasil apresentou diferenças significativas nos valores da métrica quando comparado com outros países. Um padrão semelhante foi observado em Silva e Moreira (2008), que estudaram ligas nacionais e encontraram que a liga brasileira e a francesa tiveram valores significativamente iguais, mas

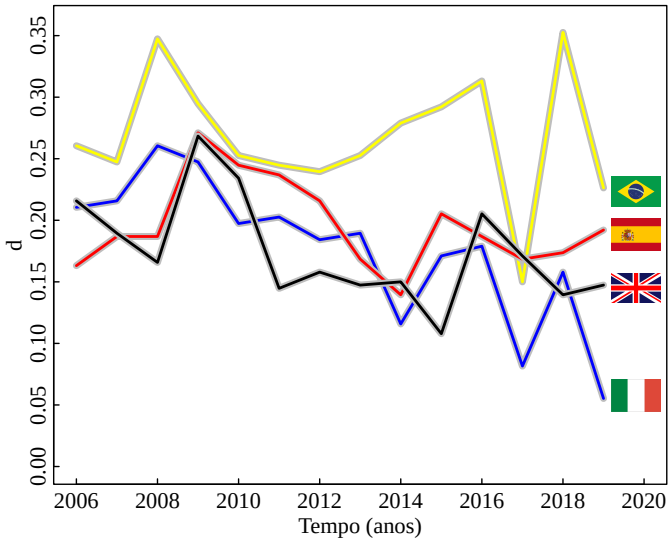
a liga brasileira apresentou valores superiores às ligas da Itália, Espanha, Inglatera, Portugal, Alemanha e Argentina.

Tabela 6 – Comparação entre as quatro ligas de quatro diferentes países para todas as participações de times de 2006 a 2019 para o Campeonato Brasileiro e de 2006/2007 à 2019/2020 para as 3 ligas Européias (*La Liga* (Espanha), *Premier League* (Inglaterra) e *Serie A* (Itália)).

	Média de D	
Série A (Brasil)	0,268	a
La Liga (Espanha)	0,196	b
Serie A (Itália)	0,176	b, c
Premier League (Inglaterra)	0,175	c

Fonte: Próprios autores.

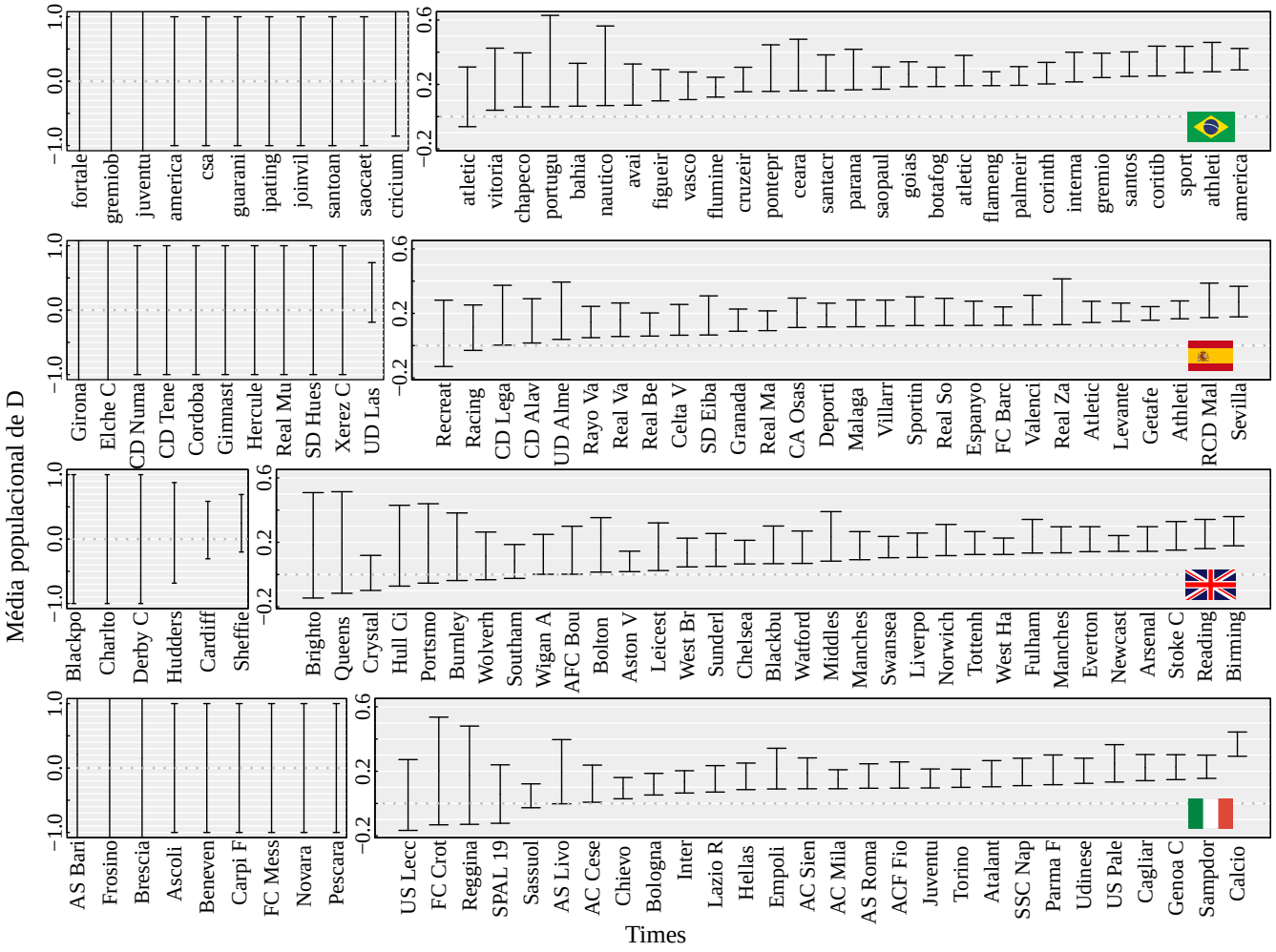
Figura 10 – $\bar{D}$ por ano para cada uma das ligas: Série A (Brasil); *La Liga* (Espanha); *Premier League* (Inglaterra) e; *Serie A* (Itália). Foram utilizados dados das edições do Campeonato Brasileiro de 2006 à 2019 e das três ligas europeias de 2006/2007 à 2019/2020.



Fonte: Próprios autores.

Na aplicação 4, foram construídos intervalos de confiança para a média de D, para cada time que participou das 4 ligas (FIGURA 11). Este é um exemplo de uma inferência que pode ser realizada para as participações dos times a partir da distribuição desenvolvida no presente Capítulo da dissertação.

Figura 11 – Intervalo de confiança para a v.a. D aplicada aos times de 4 ligas de futebol: Série A (Brasil); *La Liga* (Espanha); *Premier League* (Inglaterra) e; *Serie A* (Itália). Foram utilizados dados das edições do Campeonato Brasileiro de 2006 à 2019 e das três ligas europeias de 2006/2007 à 2019/2020.



Fonte: Próprios autores.

E por último, a aplicação 5, 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.

5.4 CONSIDERAÇÕES FINAIS

Quando estamos interessados em uma variável aleatória é importante conhecermos ou encontramos a distribuição dessa variável, para que assim possamos fazer inferências sobre ela. Em alguns casos não é possível ou não é simples de se obter a distribuição de forma fechada para a v.a., nesses casos é possível utilizarmos uma distribuição aproximada. Devido ao fato de que a v.a. D é composta da diferença de duas v.a. independentes, mas estas por sua vez são dadas por uma combinação linear de duas binomiais dependentes, a obtenção da distribuição exata de D não é simples. Como o objetivo do presente estudo foi de conseguir descrever a variável D a partir de uma distribuição de probabilidade, optamos por obter uma distribuição aproximada da mesma. Utilizamos duas aproximações para a distribuição de D, a primeira sendo uma distribuição binomial e a segunda uma distribuição normal. E, a partir de um estudo de simulação, no qual foram comparadas a adequabilidade de ambas as distribuições aos dados observados de D, verificou-se a aderência dessas aproximações para diferentes tamanhos amostrais com base no resultado do teste de aderência. Os resultados indicaram que a distribuição que melhor aproximou a distribuição dos dados foi a normal com uma aderência expressivamente superior à binomial. Assim, poderão ser utilizados os resultados inferenciais para essa distribuição, que já são amplamente conhecidos na literatura e algumas possibilidades foram mostradas nas aplicações.

Desse modo conseguimos obter estimativas pontuais e intervalares para os parâmetros populacionais que descrevem a distribuição de D, como por exemplo, tornou-se possível a realização de comparações entre as médias de D para dois times distintos ou duas ligas. Ressaltamos que até o momento não haviam sido apresentadas estimativas intervalares para média de D (isto é, da diferença relativa de pontos) para os times do Campeonato Brasileiro.

Outro aspecto positivo a ser destacado é que, com o presente artigo, é possível obter a diferença de pontos relativa média de um time, também para campeonatos desbalanceados, isto é, que o número de partidas em casa e fora de casa é diferente.

REFERÊNCIAS

BENZ, L. S.; LOPES, M. J. Estimating the change in soccer's home advantage during the covid-19 pandemic using bivariate poisson regression. **AStA Advances in Statistical Analysis**, Alemanha, v. (sem volume), n. (sem número), p. 1-28, 2021.

BOLFARINE, H.; SANDOVAL, M. **Introdução à inferência estatística**. Rio de Janeiro: SBM, 2001.

BUTLER, K.; STEPHENS, M. A. The distribution of a sum of independent binomial random variables. **Methodology and Computing in Applied Probability**, [*s. l.*], v. 19, n. 2, p. 557-571, 2017.

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

DAWSON, P.; MASSEY, P.; DOWNWARD, P. Television match officials, referees, and home advantage: Evidence from the european rugby cup. **Sport Management Review**, [*s. l.*], v. 23, n. 3, p. 443-454, 2020.

GOLLER, D.; KRUMER, A. Let's meet as usual: Do games played on non-frequent days differ? Evidence from top european soccer leagues. **European Journal of Operational Research**, [*s. l.*], v. 286, n. 2, p. 740-754, 2020.

HEGARTY, T. Information and price efficiency in the absence of home crowd advantage. **Applied Economics Letters**, [*s. l.*], v. 28, n. 21, p. 1902-1907, 2021.

MAREK, P.; VÁVRA, F. Comparison of home advantage in european football leagues. **Risks**, [*s. l.*], v. 8, n. 3, p. 87, 2020.

McCARRICK, D. *et al.* Home advantage during the Covid-19 pandemic: Analysis of european football leagues. **Psychology of Sport and Exercise**, [*s. l.*], v. 56, n. 1, p. 102013, 2021.

NEVILL, A. M.; HOLDER, R. L. Home advantage in sport. **Sports Medicine**, [*s. l.*], v. 28, n. 4, p. 221-236, 1999.

OURS, J. C. van. A note on artificial pitches and home advantage in dutch professional football. **De Economist**, Países Baixos, v. 167, n. 1, p. 89-103, 2019.

PALUDO, G. F.; FIGUEIREDO, N. N.; FERREIRA, E. B. Proposta de uma métrica para a vantagem de casa. Submetido.

POLLARD, R.; SILVA, C. D.; MEDEIROS, N. C. Home advantage in football in Brazil: Differences between teams and the effects of distance traveled. **Revista Brasileira de Futebol**, Viçosa, v. 1, n. 1, p. 3-10, 2008.

SILVA, C. D. *et al*. Competitive balance in football: A comparative study between brazil and the main european leagues (2003-2016). **Journal of Physical Education**, Maringá, v. 29,

2018.

SILVA, C. D.; MOREIRA, D. G. A vantagem em casa na futebol: comparação entre o campeonato brasileiro e as principais ligas nacionais do mundo. **Revista Brasileira de Cineantropometria e Desempenho Humano**, Florianópolis, v. 10, n. 2, p. 184-188, 2008.

VELLAISAMY, P.; PUNNEN, A. P. On the nature of the binomial distribution. **Journal of Applied Probability**, [*s. l.*], v. 38, n. 1, p. 36-44, 2001.

ZAR, J. H. **Biostatistical Analysis**. 5 ed. Upper Saddle River: Pearson, 2010.

APÊNDICE B – Um código utilizado no estudo de simulação

Código em linguagem R com o estudo de simulação utilizado para a avaliação das duas aproximações da distribuição da v.a. D.

```
###Analisando a aderencia dos dados as aproximacoes binomial e
   normal utilizando o teste do qui-quadrado
rm(list=ls(all=TRUE))
suporte<-seq(-1,1,0.1)
fe_normal1<-numeric(0)
mean_sigma_2_obs<-0
mean_aproximacao_sigma_2_obs<-0
n_simu<-1000
sequencia<-seq(10,300,10)

#Funcao que realiza o calculo de aderencia
uma_aderencia<-function(prob_c_1,prob_f_1)
{
  ad_qui_1<-matrix(0,length(sequencia),2)
  ac_qui_1<-matrix(0,n_simu,2)
  for(j in 1:length(sequencia))
  {
    for(i in 1:n_simu)
    {
      cont = 0
      while(cont == 0)
      {
        cont2 = 0
        while(cont2 == 0)
        {
          x_c<-rmultinom(sequencia[j],size=19,prob=prob_c_1)#Geracao
              da amostra com n's tamanhos.
          x_f<-rmultinom(sequencia[j],size=19,prob=prob_f_1)
          c<-c(3,1,0)
          y_c<-t(x_c)%*%c
          y_f<-t(x_f)%*%c
          a_c_obs<-x_c[1,]/19+x_c[2,]/(19*3)
          a_f_obs<-x_f[1,]/19+x_f[2,]/(19*3)
```

```
            d_obs<-a_c_obs-a_f_obs
            d_obs_binomial<-d_obs*19*3+57
            prob_c_obs<-x_c/19
            prob_f_obs<-x_f/19

            mean_sigma_2_obs<-mean(1/(3*19)*(3*prob_c_obs[1,]*(1-prob_
               c_obs[1,])+1/3*prob_c_obs[2,]*(1-prob_c_obs[2,])+3*prob
               _f_obs[1,]*(1-prob_f_obs[1,])+1/3*prob_f_obs[2,]*(1-
               prob_f_obs[2,])))
            mean_aproximacao_sigma_2_obs<-mean((a_c_obs-a_f_obs)*(1-(a
               _c_obs-a_f_obs)/(3*38)))

            p_binomial<-(1+a_c_obs-a_f_obs)/2
            fe_binom<-dbinom(1:114, 114, mean(p_binomial))
            d_obs_bin_fator<-factor(d_obs_binomial, levels=1:114)
            fo_binom<-table(d_obs_bin_fator)

            if(is.na(chisq.test(fo_binom,p=fe_binom,rescale.p=FALSE)$
               statistic))
            {
              cont2 = 0
            }else
            {
              cont2 = 1
              ac_qui_1[i,1]<-if(chisq.test(fo_binom,p=fe_binom,rescale
                 .p=FALSE)[[3]]>=0.05){1}else{0}
            }
          }
          normal_1<-numeric(0)
          for(k in 2:length(suporte)){
          normal_1<-c(normal_1, pnorm(suporte[k], mean(d_obs),  sqrt(
             mean_sigma_2_obs)) - pnorm(suporte[k-1], mean(d_obs),
             sqrt(mean_sigma_2_obs)))}
          hist_d_obs<-hist(d_obs,breaks=suporte,plot=FALSE)

          if(is.na(chisq.test(hist_d_obs$counts,p=normal_1,rescale.p=
             TRUE)$statistic))
          {
            cont = 0
          }else{
            cont = 1
            ac_qui_1[i,2]<-if(chisq.test(hist_d_obs$counts,p=normal_1,
               rescale.p=TRUE)[[3]]>=0.05){1}else{0}
          }
        }
      }
      ad_qui_1[j,1]<-sum(ac_qui_1[,1])/n_simu
      ad_qui_1[j,2]<-sum(ac_qui_1[,2])/n_simu
    }
    return(ad_qui_1)
}
```

```
#Probabilidades fixadas como mandante (prob_c) e como visitante (
   prob_f) em 10 situacoes hipoteticas
prob_c_1<-c(0.5, 0.3, 0.2);  prob_f_1<-c( 0.4, 0.4, 0.2)
prob_c_2<-c(0.8, 0.1, 0.1);  prob_f_2<-c( 0.5, 0.1, 0.4)
prob_c_3<-c(0.8, 0.1, 0.1);  prob_f_3<-c( 0.33,0.33,0.34)
prob_c_4<-c(0.8, 0.1, 0.1);  prob_f_4<-c( 0.1, 0.1, 0.8)
prob_c_5<-c(0.3, 0.1, 0.6);  prob_f_5<-c( 0.5, 0.3, 0.2)
prob_c_6<-c(0.3, 0.3, 0.4);  prob_f_6<-c( 0.5, 0.5, 0)
prob_c_7<-c(0.33,0.33,0.34); prob_f_7<-c( 0.7, 0.2, 0.1)
prob_c_8<-c(0.5, 0.3, 0.2);  prob_f_8<-c( 0.8, 0.2, 0)
prob_c_9<-c(0.2, 0.3, 0.5);  prob_f_9<-c( 0.2, 0.6, 0.2)
prob_c_10<-c(0.2,0.3, 0.5);  prob_f_10<-c(0.2, 0.8, 0)

ad1<-uma_aderencia(prob_c_1, prob_f_1)
ad2<-uma_aderencia(prob_c_2, prob_f_2)
ad3<-uma_aderencia(prob_c_3, prob_f_3)
ad4<-uma_aderencia(prob_c_4, prob_f_4)
ad5<-uma_aderencia(prob_c_5, prob_f_5)
ad6<-uma_aderencia(prob_c_6, prob_f_6)
ad7<-uma_aderencia(prob_c_7, prob_f_7)
ad8<-uma_aderencia(prob_c_8, prob_f_8)
ad9<-uma_aderencia(prob_c_9, prob_f_9)
ad10<-uma_aderencia(prob_c_10, prob_f_10)
```

6 CONSIDERAÇÕES FINAIS DA DISSERTAÇÃO

A presente dissertação desenvolveu uma métrica para obtenção da vantagem de casa no futebol. Como ponto de partida, a dissertação se baseou na métrica utilizada no artigo de Pollard e colaboradores (2008) e, com base em três aspectos destacados aqui como fragilidades dessa métrica, propôs-se uma nova métrica. A nova métrica foi denominada de diferença de pontos relativa ou d, que foi apresentada, discutida e exemplificada com a sua aplicação no Campeonato Brasileiro de Futebol Série A de 2003 a 2020. Como a primeira das principais características da métrica d, podemos citar que d não foi afetada pelo número de pontos conquistados pelo time (no sentido de ser inflacionada pela habilidade do time). Uma vez que ela apenas exprime quantos pontos um time ganhou a mais em casa do que fora em relação ao máximo possível de diferença de pontos. Isto é, $d = -100\%$ indica que o time ganhou todos pontos fora e nenhum em casa. Já os $d = 0\%$ indica que o time ganhou a mesma quantia de pontos em casa e fora e $d = 100\%$ indica que o time ganhou todos os pontos que disputou em casa.

Foi estimado por ponto e intervalo, o efeito de casa populacional para os times que participaram do Campeonato Brasileiro de 2003 à 2020. Ainda, foi proposto um teste para fazer inferência em uma única participação de um time em uma edição do campeonato. Isto é, responde a seguinte pergunta: a diferença de pontos obtida em uma única participação em uma competição pode ser apenas fruto do acaso (de algum fator aleatório ou não relacionado), ou se há evidências suficientes para dizer que existiu mais pontos conquistados em casa do que fora na participação de um time em uma edição. Ainda, uma novidade do teste é que se a diferença de pontos é pouco expressiva, o teste retorna três possíveis saídas: existência de efeito de casa positivo (vantagem de casa); inexistência de efeito de casa ou; efeito de casa negativo (desvantagem de casa). Sendo que os três resultados foram encontrados no Campeonato Brasileiro de 2003 à 2020, que do total de 370 participações, uma apresentou efeito de casa negativo, 259 tiveram efeito de casa positivo e 110 não apresentaram efeito de casa. Uma informação de vantagem de casa por participação pode contribuir para melhorar as explicações do que causa a vantagem de casa. Especialmente pelo fato de que o efeito de casa para um mesmo clube não foi constante ao longo do tempo.

Além dessas características supracitadas, podemos mencionar que a métrica d é relativamente simples de ser obtida, porém, o teste para a métrica não é tão simples de ser obtido, sendo necessário o uso de recursos computacionais. Um código que realiza o cálculo da mé-

trica e do teste foi construído em linguagem R e disponibilizado na apêndice do Capítulo 4. Cabe ressaltar ainda que a métrica mede a diferença de pontos entre casa e fora, e não diz respeito aos times que ganharam mais pontos como mandante.

No Capítulo 5, a métrica foi abordada sob a perspectiva da estatística e, por isso, a métrica foi modelada como uma v.a. (que passou a ser denominada de D ao invés de d) e buscou-se estudar sua distribuição. Pode-se elencar que, no futebol, D consiste em uma v.a. composta por outras duas v.a. Y_c e Y_f, que foram assumidas como independentes. Ainda, tem-se que Y_c e Y_f, cada uma é constituída da combinação linear de duas v.a. dependentes assumidamente com distribuições binomiais.

Uma vez que a distribuição da combinação linear de duas variáveis binomiais dependentes não é trivial, não foi possível obter a forma fechada da distribuição exata de D. E por isso foram obtidas duas aproximações, uma pela distribuição normal e outra pela distribuição binomial. Como principal resultado do estudo, a aproximação que teve melhor aderência aos dados no estudo de simulação foi a aproximação normal. E, portanto, essa foi a aproximação escolhida para representar D e ser utilizada nas aplicações. Ainda como resultado do estudo de simulação, a aderência dos dados à aproximação normal foi boa na maioria dos tamanhos de amostra, sendo que em algumas combinações dos parâmetros $\mathbf{p}_c$ e $\mathbf{p}_f$, a normal apresentou menor aderência nos maiores tamanhos amostrais. Pode-se mencionar que em vetores de probabilidade obtidos no Campeonato Brasileiro, a aderência da normal foi superior àquelas situações com parâmetros $\mathbf{p}_c$ e $\mathbf{p}_f$ hipotéticos.

Como conclusão, foi proposta uma v.a. (ou métrica) e encontrada uma aproximação da sua distribuição, o que permitiu realizar inferências. Pode-se assumir que a métrica é aproximadamente normalmente distribuída, com média e variâncias explicitadas no Capítulo 5. Como D é aproximadamente normal, há certas vantagens, isto é, o $\bar{X}$ e S^2 podem ser utilizados como estimadores dos parâmetros populacionais. Características essa que possibilita a utilização de testes já amplamente difundidos e conhecidos.

REFERÊNCIAS

ALMEIDA, L. G.; OLIVEIRA, M. L.; SILVA, C. D. Uma análise da vantagem de jogar em casa nas duas principais divisões do futebol profissional brasileiro. **Revista Brasileira de Educação Física do Esporte**, São Paulo, v. 25, n. 1, p. 49–54, 2011.

BOLFARINE, H.; SANDOVAL, M. C. **Introdução à inferência estatística**. Rio de Janeiro: SBM, 2001.

CLARKE, S. R.; NORMAN, J. M. Home ground advantage of individual clubs in english soccer. **Journal of the Royal Statistical Society: Series D (The Statistician)**, [*s. l.*], v. 44, n. 4, p. 509–521, 1995.

COURNEYA, K. S.; CARRON, A. V. The home advantage in sport competitions: A literature review. **Journal of Sport and Exercise Psychology**, [*s. l.*], v. 14, n. 1, p. 13–27, 1992.

D'AGOSTINO, R.B; STEPHENS, M.A. Overview. *In*: D'AGOSTINO, R.B; STEPHENS, M.A. **Goodness-of-Fit Techniques**. New York: Marcel Dekker. 1986.

Eclesiástico. Português. *In*: **BÍBLIA Sagrada**: Católica Apostólica. [*S. l.*]: Editora Autch, 2012. p.1566. Bíblia A. T.

FAJARDO, L. et al. A vantagem de jogar em casa em relação às séries do campeonato brasileiro de futebol. **Revista Brasileira de Futebol**, Viçosa, v. 10, n. 2, p. 25–34, 2019.

GOUMAS, C. Modelling home advantage for individual teams in uefa champions league football. **Journal of sport and health science**, Xangai, v. 6, n. 3, p. 321–326, 2017.

JOHNSON, N. L.; KOTZ, S.; BALAKRISHNAN, N. **Discrete multivariate distributions**. New York: Wiley Interscience, 1997.

JOHNSON, N. L.; KEMP, A. W.; KOTZ, S. **Univariate Discrete Distributions**. New York: Wiley Interscience, 2005.

JOHNSON, N. L.; KOTZ, S.; BALAKRISHNAN, N. **Continuous univariate distributions**. 2. ed. Vol 1. New York: Wiley Interscience, 1994.

LEITE, W. S. S. Home advantage: Comparison between the major european football leagues. **Athens Journal of Sports**, Atenas, v. 4, n. 1, p. 65–74, 2017.

MAGALHÃES, M. N. **Probabilidade e Variáveis Aleatórias**. 3a ed. São Paulo: Edusp. 2011.

MAREK, P.; VÁVRA, F. Comparison of home advantage in european football leagues. **Risks**, [*s. l.*], v. 8, n. 3, p. 87, 2020.

MOOD, A. M.; GRAYBILL, F. A.; BOES, D. C. **Introduction to the Theory of Statistics** 1974. Singapura: McGraw-Hill Kogakusha, 1974.

MOORE, D. S. Tests of Chi-Squared Type. *In*: D'AGOSTINO, R.B; STEPHENS, M.A. **Goodness-of-Fit Techniques**. New York: Marcel Dekker. 1986.

OLIVEIRA, P. V. S. R. *et al*. Vantagem de jogar em casa na Série A do campeonato brasileiro e na copa do brasil. **Revista Brasileira de Futsal e Futebol**, São Paulo, v. 12, n. 48, p. 180–186, 2020.

POLLARD, R.; SILVA, C. D.; MEDEIROS, N. C. Home advantage in football in brazil: Differences between teams and the effects of distance traveled. **Revista Brasileira de Futebol**, v. 1, n. 1, p. 3–10, 2008.

RESNICK, S. **A probability path**. Boston: Birkhauser, 2005.

RIBEIRO, L. de C. *et al*. Did the absence of crowd support during the covid-19 pandemic affect the home advantage in brazilian elite soccer? **Journal of Human Kinetics**, Polônia, v. 81, p. 251-258, 2022.

SCHADER, M.; SCHMID, F. Two rules of thumb for the approximation of the binomial distribution by the normal distribution. **The American Statistician**, [*s. l.*], v. 43, n. 1, p. 23–24, 1989.

STEFANI, R. Measurement and interpretation of home advantage. **In: Statistical Thinking in Sports**. New York: Chapman and Hall/CRC, 2007. p. 215–228.

ZAR, J. H. **Biostatistical Analysis**. 5 ed. Upper Saddle River: Pearson, 2010.