UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

HEMILIN DO AMARAL PASESNY

ESTIMATIVA DE PROBABILIDADES ESPORTIVAS VIA MODELOS DE POISSON E DE DIXON-COLES

CURITIBA

2026

HEMILIN DO AMARAL PASESNY

ESTIMATIVA DE PROBABILIDADES ESPORTIVAS VIA MODELOS DE POISSON E DE DIXON-COLES

Estimating Sports Probabilities using Poisson and Dixon-Coles Models

Dissertação apresentada como requisito para obtenção do título de Mestra em Matemática. Área de Concentração: Matemática na Educação Básica, no Programa Mestrado Profissional em Matemática em Rede Nacional - PROFMAT da Universidade Tecnológica Federal do Paraná - UTFPR.
Linha de pesquisa: Formação de Professores de Matemática da Educação Básica.
Orientador: Dr. João Luis Gonçalves
Coorientador: Dr. Ronie Peterson Dario

CURITIBA

2026

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Campus Curitiba

HEMILIN DO AMARAL PASESNY

ESTIMATIVA DE PROBABILIDADES ESPORTIVAS VIA MODELOS DE POISSON E DE DIXON-COLES

Trabalho de pesquisa de mestrado apresentado como requisito para obtenção do título de Mestre Em Matemática da Universidade Tecnológica Federal do Paraná (UTFPR). Área de concentração: Matemática Na Educação Básica.

Data de aprovação: 04 de Fevereiro de 2026

Dr. Joao Luis Goncalves, Doutorado - Universidade Tecnológica Federal do Paraná

Dr. Andre Fabiano Steklain Lisboa, Doutorado - Universidade Tecnológica Federal do Paraná

Dr. Roberto Pettres, Doutorado - Universidade Federal do Paraná (Ufpr)

Documento gerado pelo Sistema Acadêmico da UTFPR a partir dos dados da Ata de Defesa em 05/02/2026.

Dedico este trabalho a Deus, que em Sua infinita bondade me concedeu forças e tornou possível a conclusão desta etapa.

AGRADECIMENTOS

A Deus, que me capacitou e orientou no desenvolvimento deste trabalho, oferecendo-me bênçãos além do que poderia imaginar.

À minha família, especialmente à minha mãe, Marlena, e ao meu pai, Altamir, pelo amor, incentivo e apoio incondicional em minha caminhada acadêmica.

Às minhas irmãs, Kamila e Hingred, pela compreensão constante, e aos meus amigos Marcio e Izabella, pelo incentivo e companhia ao longo desta jornada.

Ao Prof. Dr. João Luis Gonçalves, que aceitou me orientar e o fez de forma exemplar, agradeço profundamente por sua paciência, dedicação, vida e exemplo.

Ao meu coorientador, Prof. Dr. Ronie Peterson Dario, por toda ajuda e apoio oferecido.

À Sociedade Brasileira de Matemática que, na busca da melhoria do ensino de matemática na Educação Básica, viabilizou a implementação do PROFMAT.

À CAPES, pela recomendação do PROFMAT por meio do parecer do Conselho Técnico Científico da Educação Superior.

RESUMO

PASESNY, Hemilin do Amaral. **Estimativa de Probabilidades Esportivas via Modelos de Poisson e de Dixon-Coles**. 62 f. Dissertação - Programa de Mestrado Profissional em Matemática em Rede Nacional - PROFMAT, Universidade Tecnológica Federal do Paraná. Curitiba, 2026.

O presente trabalho explora os modelos matemáticos de Poisson e de Dixon-Coles, utilizados para predizer probabilidades de resultados de partidas entre times de futebol em um campeonato. Para compreendermos esses modelos, realizamos uma fundamentação estatística consistente, abordando temas fundamentais, como Função de Massa de Probabilidade, Distribuição Normal, Distribuição de Bernoulli, Distribuição de Poisson e Função de Verossimilhança que formam a base dos Modelos Lineares e Modelos Lineares Generalizados. Realizamos a implementação dos modelos de Poisson e de Dixon-Coles em Python. Utilizamos o modelo de Dixon-Coles para criar estimativas referentes a partidas do Campeonato Brasileiro de Futebol Masculino, Série A, 2025. Isso permite comparar tais estimativas com os resultados reais de uma partida e com as *odds* das casas de apostas, e possibilita conclusões. Uma sequência de atividades foi sugerida para contribuir com o desenvolvimento da habilidade de tomada de decisão, a prevenção de comportamentos de risco e para auxiliar os professores de Matemática da Educação Básica a abordar o tema em sala de aula de forma coerente, a fim de mostrar a realidade nociva das apostas.

Palavras-chave: distribuição; Poisson; Dixon-Coles; estimativas; casas de apostas.

ABSTRACT

PASESNY, Hemilin do Amaral. **Estimating Sports Probabilities using Poisson and Dixon-Coles Models**. 62 pg. Dissertation - Programa de Mestrado Profissional em Matemática em Rede Nacional - PROFMAT, Universidade Tecnológica Federal do Paraná. Curitiba, 2026.

This work explores the Poisson and Dixon-Coles mathematical models used to predict the probabilities of match results between soccer teams in a championship. To understand these models, we provide a consistent statistical foundation, addressing fundamental topics such as Probability Mass Function, Normal Distribution, Bernoulli Distribution, Poisson Distribution, and Likelihood Function, which form the basis of Linear Models and Generalized Linear Models. We implemented the Poisson and Dixon-Coles models in Python. We chose to use the Dixon-Coles model to create estimates for matches in the 2025 Brazilian Men's Soccer Championship, Series A. This allows us to compare these estimates with the actual results of a match and with the odds offered by bookmakers, enabling us to make relevant observations about these comparisons. The sequence of suggested activities aims to contribute to the development of decision-making skills and the prevention of risky behavior, as well as to help elementary school math teachers address the topic in the classroom in a coherent manner, in order to show the harmful reality of gambling.

Keywords: distribution; Poisson; Dixon-Coles; estimates; bookmakers.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A regulamentação do setor de apostas, Lei nº14.790 de 2023, alavancou o já grande mercado de apostas. Esse mercado movimentou cerca de R$ 22 bilhões em 2025 no Brasil, o que colocou o país na quinta posição entre os maiores mercados mundiais do segmento segundo (MOTA, 2025).

As *bets*, termo popular para as apostas esportivas, caíram no gosto dos brasileiros. É comum ouvirmos relatos como “meu vizinho vive de apostas”, “eu entendo de futebol, sei quem vai ganhar”, “estudando dá para ganhar” e “é fácil adivinhar o que vai acontecer”.

Para fazer a rotina de apostas superavitária é necessário, além de grande conhecimento das caracteríticas do esporte em questão, amplo conhecimento matemático, em particular, um grande domínio sobre o cálculo de probabilidades e modelagem computacional, para tentar construir modelos matemáticos superiores aos das casas de apostas.

Porém, veremos que ganhar das casas de apostas, ou seja, ser lucrativo no conjunto de suas apostas, é uma tarefa extremamente difícil, ao contrário do que muitos podem imaginar. Por isso, vamos resumidamente compreender a dinâmica das casas de apostas. Um termo muito utilizado nesse meio é *odd* que, traduzido do inglês para o português, significa “chance”.

As *odds* podem ser expressas de várias formas, mas no Brasil a forma mais comum é a *odd* decimal. Outros tipos comuns de *odds* são as fracionária e a americana. A *odd* decimal é um número com até duas casas decimais, que representa o fator pelo qual o valor apostado será multiplicado caso a aposta seja vencedora. As *odds* decimais são aproximadamente o inverso da probabilidade.

As casas de apostas propõem *odds* de forma a garantir ganhos na ocorrência de qualquer resultado e as *odds*, via de regra, são ajustadas dinâmicamente ao longo do tempo para equilibrar os volumes de apostas, corrigir estimativas e limitar perdas, evitando prejuízos.

Ao observarmos as *odds* conseguimos analisar a expectativa das casas de apostas, pois elas são inversamente proporcionais a probabilidade do evento se concretizar, ou seja, quanto maior a *odd*, menor a probabilidade do evento ocorrer e inversamente uma *odd* pequena indica grande probalidade do evento ocorrer. Obviamente as *odds* são limitadas inferiormente por 1, pois caso contrário mesmo ganhando a aposta o apostador teria prejuízo.

Eventualmente pode-se ganhar da casa de apostas, entretanto para ganhar assiduamente é necessário ter informação ou estimativas mais precisas sobre as probabilidades do evento em questão e sem a ajuda de um bom modelo matemático essa tarefa é quase impossível. Um modelo matemático é um algoritmo que utiliza informações sobre o jogo para tentar prever os resultados.

Como disse David Sumpter, professor de matemática na Universidade de Uppsala, Suécia

(GALLAS, 2024), caso uma pessoa crie um modelo matemático mais eficiente do que o utilizado pelas casas de apostas, seria mais vantajoso vendê-lo as próprias casas do que utilizá-lo para apostar e obter lucro esporádico.

O objetivo das casas de apostas é lucrar sobre a totalidade das apostas e não sobre uma aposta em particular, como fazem os apostadores. Para cada aposta, a casa de apostas determina a cobrança de uma taxa, da qual obtem seu lucro. Essa taxa é chamada de "*vigorish*" e está imbutida no valor das *odds*.

Outro grande desafio para os apostadores é a lei dos grandes números. A lei afirma que a medida que um experimento é realizado repetidamente, ele tende a ter resultados mais próximos das estimativas.

Por exemplo, ao lançar uma moeda justa temos 50% de probabilidade tanto para cara quanto para coroa. Lançando essa moeda 1000 ou 10000 vezes, e analisando os resultados, o percentual de caras e coroas devem estar bem próximos de 50%. Essa lei pode ser usada para estimar a probabilidade de um determinado evento com precisão controlada, desde que o evento possa ser repetido muitas vezes.

Contudo, no contexto de apostas esportivas os eventos dificilmente se repetem e tão pouco a quantidade de apostas disponíveis é grande o suficiente para validar a qualidade das estimativas de probabilidade dos apostadores.

A escolha desse tema deve-se a sua atualidade e grande impacto no cotidiano da população, inclusive de muitos estudantes. Ao trabalhar essa temática em sala de aula, proporcionamos um aprendizado que pode melhorar a vida das pessoas, evitando perdas financeiras, por exemplo, além de proteger os alunos de possíveis danos causados por comportamentos de risco, como o vício em jogos de azar, logo o assunto se mostra relevante para a formação dos estudantes.

Segundo a Base Nacional Comum Curricular, BNCC (BRASIL, 2018), de acordo com a primeira competência específica da matemática e suas tecnologias no Ensino Médio, desenvolver a habilidade de tomada de decisão e de prevenção de comportamentos de risco se enquadraria na formação de cidadãos críticos e reflexivos. E, trabalhar o comportamento das casas de apostas e a análise de riscos do apostador se enquadra na unidade de conhecimento Probabilidade e Estatística, contemplando as seguintes habilidades: EM13MAT106 – "Identificar situações da vida cotidiana nas quais seja necessário fazer escolhas levando-se em conta os riscos probabilístico" (Brasil, 2018, p. 546) e EM13MAT312 – "Resolver e elaborar problemas que envolvem o cálculo de probabilidade de eventos em experimentos aleatórios sucessivos" (Brasil, 2018, p. 546).

O objetivo geral deste trabalho será apresentar o modelo Dixon-Coles e utilizá-lo para estimar probabilidades esportivas, visando evidenciar que, de modo geral, as apostas não são uma prática vantajosa.

Tendo como objetivos específicos: assimilar os conteúdos trabalhados, de modo a realizar estimativas e fazer conjecturas sobre dados e informações. Compreender o uso da matemática

financeira aplicada a situações-problema da vida cotidiana. Resolver problemas que envolvam o cálculos de *odds* via probabilidade. E, desenvolver atividades didáticas sobre a temática, baseada nas competências gerais e específicas da BNCC.

Defendemos que nosso objetivo não é incentivar as apostas, mas sim mostrar como é difícil ganhar das casas de apostas. Sendo assim, desaconselhável realizar apostas esportivas ou de qualquer outra forma.

O modelo de Dixon-Coles (DIXON; COLES, 1997) apresentado por Mark Dixon e Stuart Coles, em 1997, melhora o modelo de Poisson ao introduzir um fator para tratar placares menos elásticos, como 0x0, 1x0, 0x1 e 1x1, que são os mais frequentes.

Esse modelo foi utilizado por muito tempo para estimar o resultado (números de gols) de partidas de futebol. O modelo se baseia na quantidade de gols que os times sofrem e marcam, para estabelecer a força ofensiva e defensiva de cada equipe, tentando desta forma prever a quantidade de gols que a partida teria. O modelo de Dixon-Coles, bem como o modelo de Poisson (POISSON, 1837) que o antecede serão apresentados no Capítulo 4.

Utilizamos o Capítulo 2 como suporte para os Capítulos 3 e 4, nele abordaremos brevemente a distribuição de Poisson, que integra a base dos modelos Poisson e Dixon-Coles. Também discutimos a distribuição de Bernoulli que modela eventos binários e a distribuição de Poisson que pode ser vista como uma situação limite da distribuição de Bernoulli. As funções massa de probabilidade são responsáveis por conectar as distribuições ao modelo Dixon-Coles. Por fim, discutimos a função de verossimilhança, utilizada para estimar parâmetros.

No Capítulo 3, estudaremos os modelos lineares e sua extensão, os modelos lineares generalizados. Para isso, utilizamos as distribuições e as funções de massa de probabilidade apresentadas no Capítulo 2, que servem como base para parametrizar e estimar os modelos lineares generalizados. O modelo Dixon–Coles, é um modelo linear generalizado, com função de massa de probabilidade corrigida.

Com a fundamentação matemática necessária estabelecemos o modelo de Poisson e de Dixon–Coles no Capítulo 4. A implementação dos modelos na linguagem Python ocorre no Capítulo 5 onde, posteriormente, utilizamos o modelo de Dixon-Coles para estabelecer um *ranking* da força de ataque e outro da força de defesa de cada equipe do Campeonato Brasileiro de Futebol Masculino, Série A, de 2025. Também apresentamos as estimativas do modelo comparadas aos resultados reais e às odds das casas de apostas. Considerando uma rodada completa do campeonato, comparamos as probabilidades estimadas pelo modelo a todos os jogos da rodada e os resultados observados, destacando as análises mais relevantes.

Por fim, o Capítulo 6 traz uma sequência de atividades direcionadas a estudantes do Ensino Médio. Dessa forma, o estudo propõe-se a explicar o modelo e a evidenciar que as dificuldades em estimar probabilidades são enormes. Este trabalho visa contribuir para a formação e o aperfeiçoamento de professores da educação básica.

2 FUNDAMENTAÇÃO ESTATÍSTICA

Neste capítulo apresentamos e discutimos alguns conceitos que utilizaremos no decorrer do trabalho.

2.1 FUNÇÃO DE MASSA DE PROBABILIDADE

A Função de Massa de Probabilidade (FMP), em inglês PMF (*Probability Mass Function*), estudada em estatística, é utilizada na análise de variáveis aleatórias discretas.

Uma variável aleatória é discreta quando seus valores possíveis formam um conjunto discreto, ou seja, um conjunto enumerável, que pode ser finito ou infinito, espaço amostral.

Apresentamos duas situações como exemplo, para esclarecer a diferença entre espaço amostral finito e infinito.

A primeira é o lançamento de um dado honesto de seis faces, sendo o experimento "lançar um dado de seis faces uma única vez", que fornece uma variável aleatória discreta com conjunto finito de valores, a saber o espaço amostral é $\{1, 2, 3, 4, 5, 6\}$.

A segunda situação é considerar o evento "lançar uma moeda honesta (cara(K) ou coroa(C)) repetidamente até que o resultado **cara** apareça pela primeira vez", que é representado por uma variável aleatória discreta com conjunto infinito de valores, a saber o espaço amostral é $\{K, CK, CCK, CCCK, CCCCK, \ldots\}$ ou $\{1, 2, 3, 4, 5, \ldots\}$ sendo que o segundo conjunto representa a quantidade de lançamentos até obter cara.

A cada um dos exemplos associa-se uma função de probabilidade e é possível calcular o valor esperado. A intencionalidade desses exemplos é mostrar que uma variável aleatória é uma função, que relaciona os elementos do espaço amostral com números reais.

A FMP para uma variável aleatória discreta X, é uma função que associa a cada possível valor x a sua probabilidade exata de ocorrência.

A FMP é denotada por $p(x)$ e definida como, $p(x) = P(X = x)$. Lê-se: "A função p aplicada a um valor x é igual a probabilidade de que a variável aleatória X assuma exatamente o valor x".

Para que uma função seja considerada uma função massa de probabilidade válida, ela deve satisfazer duas condições:

(i) Não-negatividade: A probabilidade de qualquer evento nunca pode ser negativa. Portanto, para todo valor possível x vale,

$$p(x) \geq 0.$$

(ii) Soma igual a 1: A soma das probabilidades de todos os resultados possíveis deve ser exatamente 1, isto é,

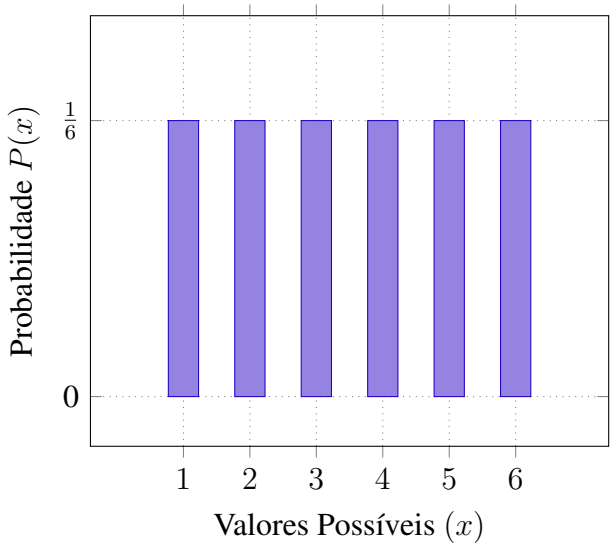
$$\sum_x p(x) = 1.$$

Dessa forma, temos que $0 \leq p(x) \leq 1$ para qualquer valor x.

A FMP é normalmente representada por um gráfico de barras, onde o eixo horizontal lista os valores possíveis da variável aleatória, enquanto o eixo vertical mostra a probabilidade de cada valor.

Para a primeira situação (variável aleatória discreta com conjunto finito de valores) considerada acima, o gráfico teria 6 barras, todas com a mesma altura de $1/6$:

Figura 2.1 – Gráfico de barras da função de probabilidade de um lançamento de um dado justo (distribuição uniforme discreta).



Fonte: Autora.

A FMP é utilizada para determinar não apenas o valor esperado, mas também a variância, que mede a dispersão dos valores que a variável aleatória pode assumir em torno da média, que podem ser calculadas por meio da fórmula:

$$Var(X) = E[(X - \mu)^2] = \sum_x (x - \mu)^2 \cdot p_X(x)$$

Sendo μ a média e $E[(X - \mu)^2]$ o valor esperado dos desvios quadráticos.

Um exemplo de uma distribuição de probabilidade é a distribuição de Poisson, que modela o número de eventos em um intervalo fixo de tempo.

A extensão desses conceitos para espaços amostrais contínuos é bastante direta. Essencialmente, a distribuição de probabilidades passa a ser representada por uma curva contínua e a condição de soma igual a 1 para as probabilidades é feita usando integração.

2.2 DISTRIBUIÇÃO NORMAL

A distribuição normal é uma distribuição de probabilidade absolutamente contínua parametrizada pelo seu valor esperado μ e pelo desvio padrão σ, com $\mu \in \mathbb{R}$ e $\sigma \in \mathbb{R}^+$. A função de densidade (ou de massa) de probabilidade da distribuição normal é denotada como

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}, \tag{2.1}$$

para $x \in \mathbb{R}$.

Uma distribuição é dita normal quando a média é nula e o desvio padrão é unitário, ela é muito útil em diversas situações. Essa curva também é conhecida como gaussiana. Ela possui dois parâmetros, a média μ, que indica onde a curva está centralizada e a variância σ^2, positiva e não nula, que é responsável por descrever o seu grau de dispersão. Dessa forma, temos que X possui uma distribuição normal e escreve-se $X \sim N(\mu, \sigma^2)$.

A curva de densidade associada a essa distribuição tem forma de sino. A variável X se distribui de forma contínua (variável contínua) no intervalo $\{-\infty < x < +\infty\}$ e possui área total unitária sob a curva do modelo. A média μ é chamada de valor esperado, $E[X]$, e a variância σ^2 é representada por $Var(X)$.

A função densidade de probabilidade definida na Equação 2.1 , devidamente ajustada pertence a uma família exponencial, conforme a Equação 3.3.

2.3 DISTRIBUIÇÃO DE BERNOULLI

Um ensaio de Bernoulli é um experimento aleatório com exatamente dois resultados possíveis, sucesso ou fracasso. São exemplos de ensaios de Bernoulli os seguintes eventos:

- O aluno passa ou não passa na matéria de Matemática;
- O resultado para o teste de covid-19 é positivo ou negativo;
- No lançamento de um dado ocorre ou não ocorre a face 1 voltada para cima;
- No lançamento de uma moeda ocorre ou não ocorre a face cara voltada para cima.

Nas situações descritas, os eventos em questão têm apenas duas possibilidades, que poderiam ser representadas por sucesso e fracasso. Onde podemos associar numericamente ao sucesso o número 1 e ao fracasso o número 0. A probabilidade de obter sucesso no evento em questão é denotada por p e o seu complemento $1 - p$ é a probabilidade de obter fracasso no evento. Cada ensaio de Bernoulli é independente, então o resultado de um ensaio não interfere na probabilidade de sucesso de ensaios futuros.

Uma distribuição de Bernoulli é uma distribuição de probabilidade discreta, que descreve/modela uma variável aleatória X que pode assumir apenas dois valores distintos, 1 e 0, sendo $X = 1$ a ocorrência de sucesso e $X = 0$ a ocorrência de fracasso. Dessa forma podemos descrever a função de probabilidade da distribuição de Bernoulli da seguinte forma,

$$P(X = x) = p(x) = \begin{cases} p, & \text{para } X = 1 \\ q = 1 - p, & \text{para } X = 0. \end{cases}$$

Neste caso a variável aleatória X tem distribuição de Bernoulli com parâmetro $\mathbf{p}$. Utilizando a Equação 2.1 para distribuição de Bernoulli, temos,

$$f(x) = f(x, p) = \begin{cases} p, & \text{para } X = 1 \\ q = 1 - p, & \text{para } X = 0, \end{cases}$$

$$f(x) = f(x, p) = p^x(1 - p)^{1-x}, x \in \{0, 1\} \tag{2.2}$$

Na distribuição de Bernoulli o valor esperado e a variância são dados por:

$$E(X) = p,$$
$$Var(X) = p(1 - p).$$

Vamos considerar um experimento em que lançamos uma moeda e a face que fica virada para cima depois que a moeda para é cara ou coroa, denotadas respectivamente por K e C. Assumimos que a probabilidade de a face da moeda virada pra cima ser cara é $p \in [0, 1]$, consequentemente a probabilidade de tal face ser coroa é q, então $q = 1 - p$, haja vista que os eventos são mutuamente exclusivos. Esse teste é chamado teste de Bernoulli com probabilidade de cara igual a p e probabilidade de coroa igual a q. Definimos o espaço das variáveis aleatórias de Bernoulli $S = \{K, C\}$, isto é, os valores possíveis para a variável aleatória do teste de Bernoulli são K ou C. Ainda, definimos $X : S \to \{0, 1\}$ com $X(K) = 1$ e $X(C) = 0$.

Então,

$$P_X(0) := P(X = 0) = q = 1 - p,$$

ou seja, a probabilidade de a função X ter valor 0 é q. De forma complementar

$$P_X(1) := P(X = 1) = p.$$

Assim, X é a variável aleatória de Bernoulli com probabilidade de ocorrência de K igual a p, e função de massa de probabilidade P_X.

Agora que definimos a variável aleatória de Bernoulli, vamos supor que sejam realizados n testes de Bernoulli independentes, com o número de sucessos (face da moeda igual K) sendo

representado por Y. Sabendo que a probabilidade de obter K é p, então para $0 \leq x \leq n$, com x pertencente aos inteiros ($x \in \mathbb{Z}$), temos:

$$P_Y(x) = P(Y = x) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{n}{x} p^x q^{n-x}. \tag{2.3}$$

Na Equação 2.3, o fator p^x é a probabilidade de se obter **cara** x vezes. Visto que são n lançamentos constatamos que **coroa** tem que ocorrer $n - x$ vezes e consequentemente com probabilidade q^{n-x} ou $(1-p)^{n-x}$. O número de sequências possíveis com x caras e $n - x$ coroas é representado pelo fator $\binom{n}{x}$.

A variável Y é chamada de variável aleatória binomial, com função massa de probabilidade $P_Y(x)$ e parâmetros n e p.

Note que o valor esperado da variável Y é igual a np e a variância é igual a npq.

Calcular probabilidades para uma variável aleatória binomial Y é fácil para um número relativamente pequeno de testes (n pequeno). Por exemplo, imagine que deseja-se saber o número de carros que passam em uma rua específica em um dia, e depois saber quantos carros passam em uma hora. Então, pode-se tentar determinar a probabilidade de uma certa quantidade de carros passar naquela rua em um segundo, e isso torna-se um acréscimo no número de testes feitos. Assumindo que $X \approx \mathcal{B}(n, \lambda/n)$, com $\lambda > 0$, e mantendo a definição de p e q, temos:

$$P_n(x) := \frac{n!}{x!(n-x)!}(p_n)^x (q_n)^{n-x} \tag{2.4}$$

$$= \frac{n!}{x!(n-x)!}\Big(\frac{\lambda}{n}\Big)^x \Big(1 - \frac{\lambda}{n}\Big)^{n-x} \tag{2.5}$$

$$= \frac{n(n-1)(n-2)\ldots(n-x+2)(n-x+1)}{x!}\Big(\frac{\lambda}{n}\Big)^x \Big(1 - \frac{\lambda}{n}\Big)^{n-x} \tag{2.6}$$

$$= \frac{n(n-1)(n-2)\ldots(n-x+2)(n-x+1)}{n^x}\frac{\lambda^x}{x!}\frac{\Big(1 - \frac{\lambda}{n}\Big)^n}{\Big(1 - \frac{\lambda}{n}\Big)^x}. \tag{2.7}$$

Se considerarmos o número de lançamentos n aumentando indefinidamente, ou seja, o limite quando $n \to \infty$, temos:

$$\lim_{n\to\infty} P_n(x) = \lim_{n\to\infty} \frac{n(n-1)(n-2)\ldots(n-x+2)(n-x+1)}{n^x}\frac{\lambda^x}{x!}\frac{\Big(1 - \frac{\lambda}{n}\Big)^n}{\Big(1 - \frac{\lambda}{n}\Big)^x}$$

$$= e^{-\lambda}\frac{\lambda^x}{x!} \tag{2.8}$$

e desse limite segue a definição para distribuição de Poisson:

$$f(x; \lambda) := e^{-\lambda}\frac{\lambda^x}{x!}. \tag{2.9}$$

Se n é grande o suficiente e consequentemente λ/n é pequeno o suficiente, então a Equação 2.9 pode ser considerada para aproximar a Equação 2.3. Além disso, a variável aleatória

cuja função de massa de probabilidade pode ser expressa como na Equação 2.9 é chamada de variável aleatória de Poisson com parâmetro λ.

2.4 DISTRIBUIÇÃO DE POISSON

A distribuição de Poisson, nomeada dessa forma em homenagem a seu autor, o matemático francês Siméon Denis Poisson, é uma distribuição de probabilidade discreta e não-negativa e foi originalmente publicada em Recherches sur la probabilité des jugements en matières criminelles e matière civile (POISSON, 1837).

A distribuição de Poisson mede a probabilidade de um certo número de eventos ocorrer em uma determinada quantidade de tempo fixada, com a condição de que tais eventos ocorram de forma independe entre si e com taxa média constante λ. O parâmetro λ é o parâmetro de taxa da distribuição e é igual a média do número de eventos acontecendo em um determinado intervalo de tempo fixo.

Considerando a definição da distribuição de Poisson, diz-se que a variável aleatória X segue a distribuição de Poisson com o parâmetro de taxa $\lambda > 0$, se para $x = 0, 1, 2, \ldots$, a função de massa de probabilidade de X pode ser expressa como,

$$f(x; \lambda) = P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}. \tag{2.10}$$

Em referência a Equação 3.3, podemos mostrar que a distribuição de Poisson pertence a família exponencial de distribuições no contexto de modelos lineares generalizados, pois se assumimos $\phi = 1; \theta = \log(\lambda) \iff \lambda = e^{\theta}$; $b(\theta) = \lambda = e^{\theta}$ e $c(x; \phi) = -log(x!)$, assim temos,

$$\begin{aligned} f(x; \theta; \phi) = \frac{e^{-\lambda}\lambda^x}{x!} &= \exp(\log(e^{-\lambda}) + \log(\lambda^x) - log(x!)) \\ &= \exp(x \log(\lambda) - \lambda - \log(x!)) \end{aligned} \tag{2.11}$$

No Lema 3.1, veremos que a distribuição de Poisson tem valor esperado

$$E[X] = b'(\theta) = e^{\theta} = \lambda$$

e variância

$$Var[X] = b''(\theta)\phi = e^{\theta} = \lambda.$$

Assim, fica claro que para a distribuição de Poisson o valor esperado é igual a variância. Além disso, como $\theta = \log(\lambda)$ é a função de ligação canônica, ou seja, a função de ligação que mapeia o valor esperado ao preditor linear é logarítmica, então

$$\theta = \log(b'(\theta)) = \eta = \log(\lambda) = X_i^T \beta. \tag{2.12}$$

2.5 FUNÇÃO DE VEROSSIMILHANÇA

A palavra Verossimilhança, em inglês *likelihood*, designa a semelhança entre o que é descrito, narrado ou modelado e a realidade, amostras, resultados e banco de dados reais.

Muitas distribuições de probabilidade possuem parâmetros com valores desconhecidos, os quais são estimados usando dados amostrais ou reais. A função de verossimilhança nos dá uma medida de quão bem os dados reais (resultados observados) são descritos por esses parâmetros.

A função de verossimilhança é uma função que mede a razoabilidade dos parâmetros previstos, em um conjunto de dados observados. Essa medida é o primeiro passo para se encontrar parâmetros que melhor descrevam os resultados observados.

Seja $L(\theta, y)$ a função de verossimilhança, com θ um escalar ou um vetor de parâmetros e y os valores realizados de uma variável aleatória Y, ou seja, os valores obtidos de uma amostra ou observação.

Ao maximizarmos a função de verossimilhança, para um conjunto de dados, determinamos os valores de parâmetros que melhor ajustam o modelo a esses dados.

É chamado de estimador de máxima verossimilhança o número ou vetor $\hat{\theta}$ que maximiza a função $L(\theta, y)$, de acordo com (HERZOG, 2022).

Ao dividirmos a função de verossimilhança pelo número que maximiza a função obtemos a função de verossimilhança normalizada, denotada por $R(\theta, y)$, variando de $[0, 1]$:

$$R(\theta, y) = \frac{L(\theta, y)}{L(\hat{\theta}, y)}. \tag{2.13}$$

A função de log-verossimilhança, denotada por $l(\theta, y)$, é a composição da função logarítmica com a função de verossimilhança, assim:

$$l(\theta, y) = log(L(\theta, y))$$

A função de log-verossimilhança é mais apropriada para cálculos computacionais, pois seus valores são analisados aditivamente e não multiplicativamente como é, em geral, a forma da função de verossimilhança. Por exemplo,

$$l(\theta, y) = log(\prod_{i=1}^{n} f(y_i, \theta)) = \sum_{i=1}^{n} log(f(y_i, \theta)). \tag{2.14}$$

Maximizar a função de verossimilhança L é equivalente a maximizar o logaritmo natural da função L de verossimilhança, pois a função logarítmica é crescente.

Se a função $L(\theta, y)$ é diferenciável em relação a θ, o valor de θ que maximiza a função é dado por:

$$\frac{d}{d\theta} log(L(\theta, y)) = 0$$

De acordo com (MARTINEZ et al., 2019), se a função de verossimilhança depende de mais de um parâmetro, ou seja, se θ é um vetor de parâmetros, para obter os estimadores de máxima verossimilhança devemos resolver o sistema de equações obtido por,

$$\frac{\partial}{\partial\theta_j} lnL(\theta, y) = 0, \text{ com } j = 1, 2, ..., k.$$

O conceito de verossimilhança é comumente confundido com o de probabilidade. As funções de verossimilhança e de probabilidade têm essencialmente a mesma definição. A diferença entre elas está em quem é variável independente e quem é variável dependente.

Na probabilidade, os parâmetros do modelo são fixos, e calculamos a probabilidade de ocorrência de diferentes resultados. Enquanto na verossimilhança, os resultados (os dados) são fixos e avaliamos a plausibilidade de diferentes parâmetros do modelo. Para ilustrar a relação entre probabilidade e verossimilhança vamos analisar dois exemplos.

Exemplo 2.1. *Suponha que uma moeda tem três resultados possíveis após seu lançamento, cara, coroa e borda, cujas probabilidades são respectivamente,*

$$\theta_1, \quad \theta_2 \quad e \quad \theta_3,$$

com a restrição,

$$\theta_1 + \theta_2 + \theta_3 = 1.$$

A probabilidade de em 4 lançamentos observarmos x vezes cara, y vezes coroa e z vezes borda, com $x + y + z = 4$ e x, y e z no conjunto $\{0, 1, 2, 3, 4\}$, é descrita pela função,

$$P(x, y, z) = \theta_1^x \cdot \theta_2^y \cdot \theta_3^z. \tag{2.15}$$

Note que, θ_1, θ_2 e θ_3 são parâmetros da função probabilidade e que x, y e z são as variáveis independentes dessa função.

Por exemplo, se de alguma forma sabemos que $\theta_1 = 0.475$, $\theta_2 = 0.475$ e $\theta_3 = 0.05$, então a probabilidade de em 4 lançamentos termos duas caras, uma coroa e uma borda é dada por

$$P(2, 1, 1) = \theta_1^2 \cdot \theta_2^1 \cdot \theta_3^1 = 0.475^2 \cdot 0.475^1 \cdot 0.05^1 = 0.0053586. \tag{2.16}$$

Note que o evento de que se quer saber a probabilidade é descrito por $(x, y, z) = (2, 1, 1)$.

A função de verossimilhança para o evento (x, y, z), que em 4 lançamentos tem x vezes cara, y vezes coroa e z vezes borda é

$$L(\theta_1, \theta_2, \theta_3) = \theta_1^x \cdot \theta_2^y \cdot \theta_3^z. \tag{2.17}$$

Note que, as variáveis independentes da função de verossimilhança são θ_1, θ_2 e θ_3, as probabilidades de cara, coroa e borda, respectivamente. Ainda, o evento considerado está fixo, (x, y, z), e representa os parâmetros da função de verossimilhança.

A semelhança nas definições das funções probabilidade, dada pela Equação 2.15, e verossimilhança, representada pela Equação 2.17, pode causar confusão entre os dois conceitos.

Para o evento $(2, 1, 1)$ (duas vezes cara, uma vez coroa e uma vez borda) a função de verossimilhança é definida por,

$$L(\theta_1, \theta_2, \theta_3) = \theta_1^2 \cdot \theta_2^1 \cdot \theta_3^1. \tag{2.18}$$

E, teríamos,

$$L(0.475, 0.475, 0.05) = 0.475^2 \cdot 0.475^1 \cdot 0.05^1 = 0.0053586,$$

que coincide com a $P(2, 1, 1)$ quando as mesmas probabilidades θ_1, θ_2 e θ_3 são consideradas.

Costuma ser complexo definir os parâmetros que compõem a função probabilidade, no Exemplo 2.1 seriam θ_1, θ_2 e θ_3.

Uma abordagem para determinar esses parâmetros é a seguinte, realizamos o evento uma vez (4 lançamentos no Exemplo 2.1) e assumimos o resultado obtido como o mais provável. Por exemplo, se no Exemplo 2.1 em 4 lançamentos obtivermos 2 caras, 1 coroa e 1 borda, supomos que esse é o resultado mais provável. Usando esse resultado como parâmetros, (x, y, z), para a função de verossimilhança, determinamos os parâmetros desta que a maximizam.

Assim como, ao maximizar a função de verossimilhança do evento $(2, 1, 1)$, no Exemplo 2.1, encontramos as probabilidades de θ_1, θ_2 e θ_3 que tornam o evento $(2, 1, 1)$ o mais provável possível.

No caso de eventos que podem ser facilmente repetidos, como no Exemplo 2.1, ao invés de um único evento poderíamos realizar uma grande quantidade de eventos e considerar o mais frequente, ou algum tipo de média. Contudo, em muitas situações a repetição de eventos é inviável. Por exemplo, se o evento é uma partida de futebol, a mesma dificilmente será repetida mais de uma vez em condições minimamente similares.

Para o Exemplo 2.1, considerando o evento $(2, 1, 1)$ a maximização da função de verossimilhança é feita como a seguir:

Substituindo $\theta_3 = 1 - \theta_1 - \theta_2$, obtemos,

$$L(\theta_1, \theta_2) = \theta_1^2 \cdot \theta_2^1 \cdot (1 - \theta_1 - \theta_2)^1,$$

com as restrições,

$$0 \leq \theta_1 \leq 1, \quad 0 \leq \theta_2 \leq 1, \quad \theta_1 + \theta_2 \leq 1.$$

Maximizar a verossimilhança é equivalente a maximizar o logarítmo da verossimilhança. E maximizar o logarítmo da verossimilhança tem a vantagem de que potências tornam-se produtos e produtos tornam-se somas, reduzindo assim situações de *underflow* numérico.

Log-verossimilhança:

$$\ell(\theta_1, \theta_2) = \ln(L(\theta_1, \theta_2)) = 2\ln\theta_1 + \ln\theta_2 + \ln(1 - \theta_1 - \theta_2).$$

Derivadas parciais:

$$\frac{\partial \ell}{\partial \theta_1} = \frac{2}{\theta_1} - \frac{1}{1 - \theta_1 - \theta_2} = 0,$$
$$\frac{\partial \ell}{\partial \theta_2} = \frac{1}{\theta_2} - \frac{1}{1 - \theta_1 - \theta_2} = 0.$$

Resolvendo as equações:

Da primeira derivada:

$$\frac{2}{\theta_1} = \frac{1}{1 - \theta_1 - \theta_2} \implies 2(1 - \theta_1 - \theta_2) = \theta_1,$$
$$2 - 2\theta_1 - 2\theta_2 = \theta_1 \implies 2 - 2\theta_2 = 3\theta_1,$$
$$3\theta_1 + 2\theta_2 = 2. \quad (1)$$

Da segunda derivada:

$$\frac{1}{\theta_2} = \frac{1}{1 - \theta_1 - \theta_2} \implies \theta_2 = 1 - \theta_1 - \theta_2,$$
$$1 - \theta_1 - 2\theta_2 = 0 \implies \theta_1 + 2\theta_2 = 1. \quad (2)$$

Resolvendo o sistema linear:

$$\begin{cases} 3\theta_1 + 2\theta_2 = 2 \\ \theta_1 + 2\theta_2 = 1 \end{cases}$$

Subtraindo a segunda equação da primeira:

$$(3\theta_1 - \theta_1) + (2\theta_2 - 2\theta_2) = 2 - 1 \implies 2\theta_1 = 1 \implies \theta_1 = \frac{1}{2}.$$

Substituindo em (2):

$$\frac{1}{2} + 2\theta_2 = 1 \implies 2\theta_2 = \frac{1}{2} \implies \theta_2 = \frac{1}{4}.$$

Calculando θ_3:

$$\theta_3 = 1 - \theta_1 - \theta_2 = 1 - \frac{1}{2} - \frac{1}{4} = \frac{1}{4}.$$

Conclusão:

O estimador de máxima verossimilhança é:

$$\boxed{\hat{\theta}_1 = 0.5, \quad \hat{\theta}_2 = 0.25, \quad \hat{\theta}_3 = 0.25.}$$

Exemplo 2.2. *Suponha que a moeda do exemplo anterior é tal que o evento 2 caras, 1 coroa e 1 borda é o mais provável possível, ou seja, pelo resultado obtido $\theta_1 = 0.5$, $\theta_2 = 0.25$ e $\theta_3 = 0.25$. Qual a probabilidade de o resultado em 3 lançamentos ter mais borda do que caras e/ou coroas?*

Solução:

$$\begin{aligned} P_{B>}(\theta_1, \theta_2, \theta_3) &= \underbrace{\theta_3^3}_{3\ bordas} + \underbrace{\theta_3^2\theta_1}_{2\ bordas\ e\ 1\ cara} + \underbrace{\theta_3^2\theta_2}_{2\ bordas\ e\ 1\ coroa} \\ &= \frac{1}{4^3} + \frac{1}{4^2}\frac{1}{2} + \frac{1}{4^2}\frac{1}{4} \\ &= \frac{3}{64} \\ &= 0,046875 \end{aligned}$$

Com isso fica mais claro que a Verossimilhança gera uma função de parâmetros a partir de dados aferidos, enquanto a Probabilidade gera uma função de dados a partir de parâmetros determinados.

3 MODELOS

Um modelo linear generalizado é uma flexibilização da regressão linear que permite variáveis de resposta com distribuição de erros não necessariamente normal. Essa abordagem foi proposta por Nelder e Wedderburn (NELDER; WEDDERBURN, 1972).

Neste capítulo, apresentamos rapidamente o que é um modelo linear, ou seja, uma regressão linear, e na sequência detalhamos como os modelos lineares generalizados são construídos. Em particular, por ser o tipo de modelo que utilizaremos nos próximos capítulos, vamos abordar a regressão de Poisson. Grande parte dos conceitos foram baseados no material de Kubrusly (KUBRUSLY, 2014).

3.1 MODELOS LINEARES

Um modelo linear estabelece uma relação linear entre uma variável dependente (variável resposta), Y_i, e variáveis independentes (variáveis preditoras), $x_{1i} \ldots, x_{pi}$, em que as primeiras são determinadas pelas segundas. No modelo linear, a média da distribuição de Y_i varia de forma linear com as variáveis $x_{1i} \ldots, x_{pi}$. Matematicamente, um modelo linear é determinado por relações da forma

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i, \text{ para } i = 1, \ldots N,$$

em que β_j, com $j = 0, 1, \ldots, p$, são os coeficientes que determinam a relação (parâmetros) e ϵ_i representam o erro entre o valor predito para a variável independente (valor observado real). Também, ϵ_i representa um erro aleatório, com distribuição normal, média zero e variância constante ($N(0, \sigma^2)$). Com σ^2, a variância, também sendo um parâmetro. Matricialmente equivale a

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon, \tag{3.1}$$

$$\text{com } \mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix}_{N\times 1}, \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1,p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2,p} \\ 1 & x_{31} & x_{32} & \ldots & x_{3,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & \ldots & x_{N,p} \end{bmatrix}_{N\times(p+1)}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}_{(p+1)\times 1} \text{ e}$$

$$\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix}_{N\times 1}.$$

Em outras palavras, a regressão linear prevê o valor esperado de uma determinada quantidade desconhecida (variável de resposta, variável aleatória) como uma combinação linear de um conjunto de valores observados (preditores).

Desejamos encontrar os parâmetros β_j que são os coeficientes da combinação linear e que determinam as relações entre as variáveis. Uma vez que temos os parâmetros β_j, podemos predizer/estimar os valores de Y a partir de valores de x.

Os parâmetros (coeficientes) são obtidos utilizando métodos como Mínimos Quadrados e Máxima Verossimilhança. Ambos os métodos produzem a mesma resposta para β em 3.1, a saber

$$\beta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \tag{3.2}$$

Essa solução é obtida desconsiderando o erro ϵ e pré-multiplicando a Equação 3.1 por $\mathbf{X}^T$, o que resulta no sistema de equações chamado normal, que é quadrado. Resolvendo esse sistema obtemos a Equação 3.2.

3.2 MODELOS LINEARES GENERALIZADOS

Embora sejam muito úteis e amplamente aplicados, os modelos lineares não são capazes de modelar as variáveis de interesse em todas as circunstâncias. Nos modelos lineares supõem-se que a variável dependente $\mathbf{Y}$ é normalmente distribuída com média μ e variância σ^2, hipótese essa que não é razoável em muitas situações, por exemplo quando a variável de interesse é binária ou uma variável de contagem.

Os modelos lineares generalizados atendem essas situações, permitindo que as variáveis de resposta tenham distribuições arbitrárias, em vez de apenas distribuições normais, e que uma função arbitrária da variável de resposta, a função de ligação, varie linearmente com os preditores, ao invés de assumir que a própria resposta deve variar linearmente.

Esta generalização do modelo linear depende de três partes principais, a saber: a **componente aleatória**, a **componente sistemática** e a **função de ligação**.

A **componente aleatória** refere-se a escolha de um tipo de distribuição para variável de resposta. Assumimos que a variável de resposta $\mathbf{Y}$, uma variável aleatória, é tal que sua distribuição depende apenas de um parâmetro θ e que a função de massa de probabilidade da distribuição pode ser expressa como

$$f(y;\theta;\phi) = \exp\Big(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\Big), \tag{3.3}$$

então a distribuição pertence a **família exponencial**. Na família exponencial representada na Equação 3.3, θ é o parâmetro canônico de localização que depende do modelo linear preditor, $b()$ é uma função real duas vezes diferenciável de θ, ϕ é chamado de parâmetro de dispersão e, em geral, é conhecido, a função $c()$ também é conhecida e independe de θ.

O Lema 3.1 apresentado em (AGRESTI, 2002) mostra quem são o valor esperado e a variância para uma distribuição da família exponencial, apresentado na Equação 3.3.

Lema 3.1. *Se uma distribuição da variável aleatória* $\mathbf{Y}$ *pertence a família exponencial na Equação 3.3, vale que:*

- *O valor esperado de* $\mathbf{Y}$ *é igual a primeira derivada de* b *com respeito a* θ*, isto é,* $E[\mathbf{Y}] = b'(\theta)$.
- *A variância de* $\mathbf{Y}$ *é o produto da segunda derivada de* b *com o parâmetro de dispersão,* ϕ*, isto é,* $Var[\mathbf{Y}] = \phi b''(\theta)$.

A **componente sistemática** identifica o conjunto de variáveis explicativas do modelo e o preditor linear é função das variáveis explicativas em combinação linear com β, e representada por η, ou seja,

$$\eta = \mathbf{X}_j^T \beta. \tag{3.4}$$

A **função de ligação**, g relaciona o valor esperado da variável dependente $\mathbf{Y}$ com o preditor linear η, assim

$$g(E[\mathbf{Y}]) = \eta. \tag{3.5}$$

A função g deve ser monótona e diferenciável.

Definição 3.1. *Uma função de ligação é dita canônica se relaciona o parâmetro canônico* θ *diretamente com o preditor linear* η*. Ou seja, se*

$$\theta = g(b'(\theta)) = \eta = \mathbf{X}_j^T \beta. \tag{3.6}$$

Um modelo linear generalizado pode ser usado para modelar variáveis de contagem ou binárias se, e somente se, sua distribuição é membro de uma família exponencial de distribuições.

3.3 REGRESSÃO DE POISSON

O modelo de regressão de Poisson é usado para modelar dados de contagem, no nosso contexto esses dados são os números de gols, por exemplo. A regressão de Poisson deriva da distribuição de Poisson com o parâmetro λ dependendo das variáveis exploratórias, que em nosso contexto referem-se aos times.

Os dados usados no modelo de regressão de Poisson consistem em uma amostra de N observações com variável de resposta Y_i e variáveis exploratórias x_i para $i = 1, \ldots, N$.

A variável de resposta Y_i é o número de ocorrências de dado evento, enquanto x_i são os vetores de variáveis exploratórias linearmente independentes que supomos serem determinados pelas variáveis de resposta.

Construímos o modelo de regressão condicionando Y_i a um vetor p-dimensional $x_i^T = [x_{1i}, x_{2i}, \ldots, x_{pi}]$ e a um vetor de parâmetros, coeficientes $\beta = [\beta_1, \beta_2, \ldots, \beta_p]$, tais que $E[y_i|x_i] = \lambda_i(x_i, \beta)$. Na Equação 2.12, já foi proposto que a função de ligação, que mapeia o valor esperado da variável de Poisson aleatória no preditor linear, η_i, seja logarítmica. Portanto, definimos a equação da regressão de Poisson como

$$f(y_i|x_i) = \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}, \quad i = 1, 2, \ldots, N, \tag{3.7}$$

com

$$\lambda_i = e^{x_i^T \beta}.$$

Agora resta estimar β. Neste ponto notamos que como conhecemos a função de massa da distribuição de Poisson é razoável confiar no método de máxima verossimilhança para encontrar estimativas para β.

Vamos considerar as variáveis aleatórias independentes Y_i, as quais tem distribuição de Poisson com parâmetro λ_i, para $i = 1, 2, \ldots, N$. Da Equação 3.7 podemos garantir que a distribuição de Poisson satisfaz as propriedades do modelo linear generalizados. Para estimar o vetor de parâmetros β precisamos aplicar o método da máxima verossimilhança considerando as N observações.

No nosso contexto, os valores de β e as variáveis de resposta Y_i estão relacionadas com os valores λ_i, e sabemos que $E[Y_i] = \lambda_i$. Mais especificamente, é uma função logarítmica que mapeia os valores de λ_i no preditor linear, ou seja, como vale $\lambda_i = e^{x_i^T \beta}$, podemos dizer que $g(E[Y_i]) = \log(\lambda_i) = x_i^T \beta$, em que $x_i = [x_{1i}, x_{2i}, \ldots, x_{pi}]$ é o vetor das variáveis explicativas.

Dessa forma a função de verossimilhança para cada Y_i é dada por

$$L(\theta_i, y_i) = f(y_i; \theta_i; \phi) = \exp\Big(\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\Big). \tag{3.8}$$

Aplicando logarítmo de base e em ambos os lados da Equação 3.8 temos a função de log-verossimilhança

$$l_i(\theta_i) = \log(L(\theta_i; y_i)) = \frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi), \tag{3.9}$$

em que $b()$ e $c()$ são funções que vamos substituir conforme a família exponencial da distribuição de Poisson, isto é, $\phi = 1$, $\theta_i = \log(\lambda_i)$, $b(\theta_i) = \lambda_i$ e $c(y_i; \phi) = -\log(y_i!)$. Dessa forma resulta que

$$l_i(\lambda_i, y_i) = y_i \log(\lambda_i) - \lambda_i - \log(y_i!). \tag{3.10}$$

Como a Equação 3.10 representa a log-verossimilhança de apenas uma variável de resposta e as variáveis Y_i são independentes, a log-verossimilhança da amostra inteira para $y_1, \ldots, y_N$ pode ser representada como,

$$l(\lambda; \mathbf{y}) = \sum_{i=1}^{N} l_i(\lambda_i; y_i) = \sum_{i=1}^{N} y_i \log(\lambda_i) - \sum_{i=1}^{N} \lambda_i - \sum_{i=1}^{N} \log(y_i!). \tag{3.11}$$

Neste ponto o parâmetro β está implícito na função log-verossimilhança, mais especificamente nas variáveis λ_i. Conforme o que se deseja modelar e o método utilizado para a maximização da verossimilhança pode ser necessário explicitar os coeficientes β_i, contudo, isso não será necessário em nosso caso.

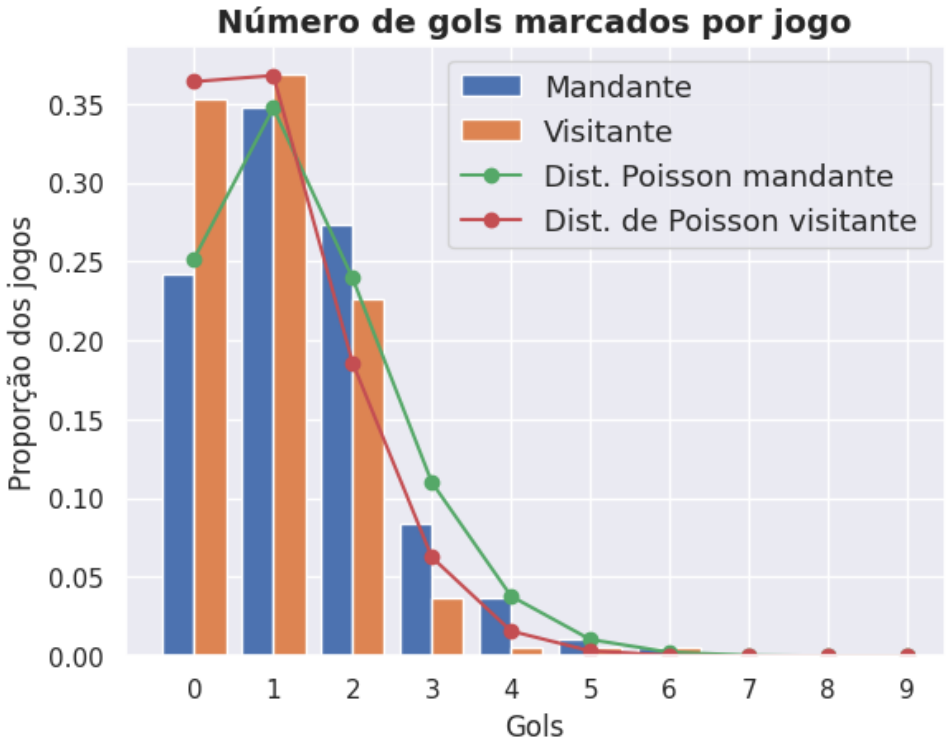
4 MODELOS DE POISSON E DIXON-COLES

Os modelos aqui propostos consideram três observações que são fundamentais para suas características e propriedades. A primeira observação é que as distribuições de gols marcados e sofridos por um time se comportam como distribuições de Poisson, ou seja, tem como características serem eventos raros, independentes e não simultâneos e com taxa média constante.

A segunda é que em um campeonato, em geral, os times têm desempenhos bastante distintos quanto a quantidade de gols marcados e de gols sofridos. Um time forte é aquele que consegue marcar muitos gols e sofre poucos gols. De forma simplista, o perfil de um time é determinado por suas forças ofensiva e defensiva.

E a terceira observação é que há uma taxa maior de vitórias dos times mandantes de seus jogos do que dos times visitantes. Isso justifica o *fator casa*, que representa a vantagem de jogar em seu próprio campo. Essa vantagem pode ter várias razões, como apoio da torcida ou o menor deslocamento para a partida, contudo para o modelo o que importa é a razão entre vitórias de mandantes e vitórias de visitantes.

Figura 4.1 – Número de gols marcados por jogo e as distribuíções de Poisson respectivas, no Campeonato Brasileiro de 2025.



Fonte: Autora.

A grande adequação da distribuição de Poisson, obtida considerando a média de gols, ao

de números de gols marcados fica clara ao analisar dados de um campeonato, como por exemplo para o Campeonato Brasileiro, Série A, Masculino, de 2025, mostrado na Figura 4.1.

Também fica latente da Figura 4.1 que a maioria dos jogos tem placares pouco elásticos, o que justifica os gols serem considerados eventos raros.

4.1 MODELO DE POISSON

O modelo de Poisson é essencialmente uma regressão de Poisson, como apresentada no Capítulo 3, ou seja, um modelo linear generalizado baseado na distribuição de Poisson. A abordagem adotada dialoga com os estudos de Dixon e Coles (DIXON; COLES, 1997).

Nesta descrição do modelo de Poisson vamos ajustar suas características e descrições para que ele atenda o objetivo de estimar o número de gols marcados pelas equipes i e j no confronto das mesmas. Sem perda de generalidade, vamos supor que a equipe i é a equipe mandante e, consequentemente, a equipe j é a equipe visitante. Denotaremos a estimativa de gols marcados pela equipe mandante i, no confronto com a equipe j por U_{ij}. A estimativa de gols da equipe visitante será denotado por V_{ij}.

Neste modelo, as quantidades U_{ij} e V_{ij} serão estimadas por uma função de Poisson e dependerão dos parâmetros α_i e α_j, que representam as forças ofensivas dos times i e j, assim como dos parâmetros β_i e β_j, que representam as forças defensivas dos times i e j, e, por fim, do parâmetro γ, que representa a vantagem de o time mandante jogar em seu estádio.

Quanto maior a força ofensiva do mandante, α_i, menor a força defensiva do visitante, β_j, e maior a vantagem por ser mandante, γ, assumimos que maior será a estimativa de gols U_{ij}. Dessa forma, vamos impor que U_{ij} depende do produto de α_i, β_j e γ. Claro que $i \neq j$ pois um time não enfrentará a si próprio.

Para o time visitante a hipótese é análoga, com diferença de não haver o parâmetro associado a vantagem do mandante, γ. Dessa forma, temos que

$$U_{ij} \approx \text{Poisson}(\alpha_i \beta_j \gamma)$$
$$V_{ij} \approx \text{Poisson}(\alpha_j \beta_i).$$

Em resumo, embora a função de Poisson receba apenas um parâmetro, esse parâmetro é o produto de dois ou três outros parâmetros.

Os parâmetros α_k, β_k e γ, com $k = 1, \ldots, N$, serão estimados a partir de dados de jogos anteriores entre os diversos times de um mesmo campeonato, ou seja, da observação de resultados anteriores. Essas estimativas serão feitas através do método de máxima verossimilhança, isto é, corresponderão aos parâmetros que melhor ajustam as estimativas de gols à quantidade de gols de fato marcados pelas equipes nas partidas.

Uma vez de posse dos parâmetros α_k, β_k e γ, a função de probabilidade de Poisson resulta que a probabilidade de o time i, jogando como mandante, fazer u gols é

$$P(U_{ij} = u) = \frac{\lambda^u e^{-\lambda}}{u!}, \tag{4.1}$$

com $\lambda = \alpha_i \beta_j \gamma$.

E a probabilidade no mesmo confronto do time visitante, j, fazer v gols é dada por

$$P(V_{ij} = v) = \frac{\mu^u e^{-\mu}}{v!}, \tag{4.2}$$

com $\mu = \alpha_j \beta_i$. Note que u e v são números naturais. Como consequência das Equações 4.1 e 4.2, a probabilidade do placar do jogo entre os times i e j ser $u \times v$ é dada por

$$P(U_{ij} = u; V_{ij} = v) = \frac{\lambda^u e^{-\lambda}}{u!} \frac{\mu^u e^{-\mu}}{v!}. \tag{4.3}$$

De posse dessas probabilidades, diversos cenários podem ser estimados.

4.2 MODELO DE DIXON-COLES

O modelo de Poisson, sintetizado na Equação 4.3, tem como hipótese a independência entre as quantidades U_{ij} e V_{ij}, ou seja, supõe que a quantidade de gols marcados por um time não interfere na quantidade de gols do outro time. Contudo, essa hipótese pode não ser a melhor. Em (DIXON; COLES, 1997), os autores provaram que para placares 0×0, 0×1, 1×0 e 1×1, isto é, com poucos gols marcados, a hipótese de independência é falha e existe uma correlação entre o número de gols marcados pelo mandante e pelo visitante.

Para corrigir essa falha, Dixon e Coles propuseram uma alternativa ao modelo de Poisson, na Equação 4.3, que leva o nome de seus autores. O modelo de Dixon-Coles define a probabilidade de que o placar do jogo entre as equipes i e j seja dado por meio da função

$$P(U_{ij} = u; V_{ij} = v) = \tau_{\lambda,\mu}(u, v) \frac{\lambda^u e^{-\lambda}}{u!} \frac{\mu^u e^{-\mu}}{v!}, \tag{4.4}$$

com $\lambda = \alpha_i \beta_j \gamma$, $\mu = \alpha_j \beta_i$ e

$$\tau_{\lambda,\mu}(u, v) = \begin{cases} 1 - \lambda\mu\rho, & \text{se } u = v = 0 \\ 1 + \lambda\rho, & \text{se } u = 0, v = 1 \\ 1 + \mu\rho, & \text{se } u = 1, v = 0 \\ 1 - \rho, & \text{se } u = 1, v = 1 \\ 1, & \text{nos demais casos.} \end{cases} \tag{4.5}$$

Em que $\max(-1/\lambda, -\mu) \leq \rho \leq \min(1/\lambda\mu, 1)$.

O parâmetro ρ é dito de correção. Se $\rho = 0$ o modelo de Dixon-Coles torna-se o Modelo de Poisson.

Precisamos mostrar que esse modelo é uma distribuição de probabilidade válida. Para isso devemos provar que,

$$\sum_{u=0}^{\infty} \frac{\lambda^u e^{-\lambda}}{u!} \sum_{v=0}^{\infty} \frac{\mu^u e^{-\mu}}{v!} \tau_{\lambda,\mu}(u,v) = 1. \tag{4.6}$$

Da conhecida série da exponencial temos,

$$\sum_{u=0}^{\infty} \frac{\lambda^u}{u!} = 1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \cdots = e^{\lambda} \tag{4.7}$$

e portanto,

$$\sum_{u=0}^{\infty} \frac{\lambda^u e^{-\lambda}}{u!} = e^{-\lambda} \sum_{u=0}^{\infty} \frac{\lambda^u}{u!} = e^{\lambda} e^{-\lambda} = e^0 = 1. \tag{4.8}$$

Analogamente,

$$\sum_{v=0}^{\infty} \frac{\mu^v e^{-\mu}}{v!} = 1 \tag{4.9}$$

e, consequentemente,

$$\sum_{u=0}^{\infty} \frac{\lambda^u e^{-\lambda}}{u!} \sum_{v=0}^{\infty} \frac{\mu^u e^{-\mu}}{v!} = 1. \tag{4.10}$$

Note da definição de $\tau_{\lambda,\mu}(u,v)$ que seu valor só é diferente de 1 se $u \leq 1$ ou $v \leq 1$. Portanto,

$$\begin{aligned}
&\sum_{u=0}^{\infty} \frac{\lambda^u e^{-\lambda}}{u!} \sum_{v=0}^{\infty} \frac{\mu^u e^{-\mu}}{v!} \tau_{\lambda,\mu}(u,v) = e^{-\lambda} e^{-\mu} \sum_{u=0}^{\infty} \frac{\lambda^u}{u!} \Big(\sum_{v=0}^{\infty} (\frac{\mu^u}{v!} \tau_{\lambda,\mu}(u,v)) \Big) \\
&= e^{-\lambda} e^{-\mu} (1 - \lambda\mu\rho + \mu + \lambda\mu\rho + \lambda + \lambda\mu\rho + \lambda\mu - \lambda\mu\rho + \frac{\lambda^2}{2!} + \frac{\mu^2}{2!} + \frac{\lambda\mu^2}{2!} + \frac{\lambda^2\mu}{2!} + \cdots) \\
&= e^{-\lambda} e^{-\mu} (1 + \mu + \lambda + \lambda\mu + \frac{\lambda^2}{2!} + \frac{\mu^2}{2!} + \frac{\lambda\mu^2}{2!} + \frac{\lambda^2\mu}{2!} + \cdots) \\
&= e^{-\lambda} e^{-\mu} \sum_{u=0}^{\infty} \frac{\lambda^u}{u!} \sum_{v=0}^{\infty} \frac{\mu^u}{v!} = 1.
\end{aligned} \tag{4.11}$$

Um aspecto importante é que mais um parâmetro, ρ, é adicionado em comparação ao modelo de Poisson. Esse parâmetro adicional também deve ser estimado pelo método da máxima verossimilhança. Para evitar problemas de superparametrização adicionamos as restrições

$$\frac{1}{n} \sum_{i=1}^{n} \alpha_i = 1 \text{ e } \alpha_i \geq 0.$$

Dadas essas restrições, os $2n + 2$ parâmetros (α_i, β_i,γ e ρ) são determinados como os que maximizam a seguinte função de verossimilhança

$$L(\alpha_i; \beta_i; \rho; \gamma; i = 1, \ldots, n) = \prod_{i,j,i \neq j}^{n} \tau_{\lambda_{ij},\mu_{ij}}(u_{i,j}, v_{i,j}) \frac{\lambda_{ij}^{u_{ij}} e^{-\lambda_{ij}}}{u_{ij}!} \frac{\mu_{ij}^{v_{ij}} e^{-\mu_{ij}}}{v_{ij}!} \tag{4.12}$$

com

$$\begin{aligned}\lambda_{ij} &= \alpha_i \beta_j \gamma, \\ \mu_{ij} &= \alpha_j \beta_i.\end{aligned} \tag{4.13}$$

Embora nessa função de verossimilhança sejam considerados dois confrontos entre cada par de times, os índices da função devem ser ajustados conforme os dados de que se dispõem. Por exemplo, se forem considerados dados de mais de uma temporada será preciso considerar mais do que 2 jogos para os pares de times ou se os dados forem de uma temporada incompleta, será necessário remover alguns confrontos.

Aplicando a função logarítmica na Equação 4.12 obtemos a função log-verossimilhança,

$$\begin{aligned}l(\alpha_i; \beta_i; \rho; \gamma; i = 1, \ldots, n) = \sum_{i,j,i \neq j}^{n} \Big[& \log(\tau_{\lambda_{ij},\mu_{ij}}(u_{i,j}, v_{i,j})) \\ & + u_{ij} \log(\lambda_{ij}) - \lambda_{ij} - \log(u_{ij}!) \\ & + v_{ij} \log(\mu_{ij}) - \mu_{ij} - \log(v_{ij}!) \Big].\end{aligned} \tag{4.14}$$

Em termos práticos, usaremos a função *minimize* do pacote *scipy.optimize* e minimizaremos menos a função log-verossimilhança, que é equivalente a maximizar a log-verossimilhança. A opção por log-verossimilhança ao invés de verossimilhança deve-se a sua maior estabilidade numérica, uma vez que permite uma comparação aditiva ao invés de multiplicativa, que é problemática com valores pequenos.

Para que as expressões de λ_{ij} e μ_{ij} atendam o modelo de regressão de Poisson descrito na Equação 3.7, reescrevemos a Equação 4.13 como

$$\begin{aligned}\lambda_{ij} &= \alpha_i \beta_j \gamma = e^{\overline{\alpha_i} + \overline{\beta_j} + \overline{\gamma}}, \\ \mu_{ij} &= \alpha_j \beta_i = e^{\overline{\alpha_j} + \overline{\beta_i}}.\end{aligned} \tag{4.15}$$

Dessa forma, como consequência de a função de ligação ser logarítmica temos que

$$\begin{aligned}\overline{\alpha_i} &= \log(\alpha_i), \\ \overline{\beta_i} &= \log(\beta_i), \\ \overline{\gamma} &= \log(\gamma),\end{aligned} \tag{4.16}$$

são os parâmetros que a maximização da verossimilhança nos retornarão.

A maneira que os dois modelos tratam a força de ataque, de defesa e a vantagem de jogar em casa difere, principalmente, em virtude das premissas implícitas de cada modelo. Esses ajustes fazem do modelo de Dixon-Coles um modelo mais rebuscado que o de Poisson.

5 RESULTADOS NUMÉRICOS

Uma vez apresentados os modelos de Poisson e de Dixon-Coles, aplicá-los requer coletar dados, implementar as funções e analisar os resultados. Essa é uma parte de característica mais prática deste trabalho, mas que é bastante desafiadora, pois situações que a teoria ignora precisam ser tratadas na prática.

Escolhemos aplicar os modelos aos dados do Campeonato Brasileiro de Futebol, da série A, masculino, do ano de 2025. Essa escolha permite que tenhamos maior familiaridade com as equipes, regras e outras situações específicas como jogos atrasados e impactos de jogos de outros campeonatos disputados em paralelo, como a Copa do Brasil e a Copa Libertadores da América.

5.1 IMPLEMENTAÇÃO

Tentamos obter os dados de fontes como o site da Confederação Brasileira de Futebol, porém devido ao grande volume (380 jogos), optamos por tomar os dados do site *football-data.co.uk* (FOOTBALL-DATA, 2025) que fornece os dados de diversos campeonatos já em formato separado por vírgulas (csv).

Nossa implementação é baseada nos trabalhos do blog *pena.lt/y* (EASTWOOD, 2021), que explora diversas análises estatísticas sobre futebol, inclusive os modelos de Poisson e Dixon-Coles, considerando principalmente dados da liga inglesa (Premier League). Neste blog a implementação é feita em *python* e por essa razão utilizamos a mesma linguagem de programação. Também nos baseamos em no blog *dashee87.github.io* (SHEEHAN, 2017).

A implementação se divide em 4 partes essenciais. A primeira parte é a leitura e o tratamento dos dados, que como já mencionamos são obtidos do site *football-data.co.uk* (FOOTBALL-DATA, 2025). Na Figura 5.1 exemplificamos uma implementação típica da leitura e tratamento dos dados, inclusive com algumas possibilidades como remoção de últimos jogos para que eles não fossem considerados no modelo, salvamento dos dados para uma planilha local e print de alguns resultados imediatos como média de gols.

A segunda parte é a implementação da função de verossimilhança. Neste ponto, os modelos de Poisson e de Dixon-Coles se diferenciam pelo fator de correção introduzido por Dixon e Coles. Na Figura 5.2 está a implementação da função de verossimilhança para o Modelo de Poisson.

Para o Modelo de Dixon-Coles o fator de correção é implementado como na Figura 5.3 e esse fator é utilizado na implementação da função de verossimilhança para o Modelo de Dixon-Coles, apresentada na Figura 5.4.

Figura 5.1 – Implementação da leitura de dados.

```python
import os
import pandas as pd
import matplotlib.pyplot as plt


df = pd.read_csv("https://www.football-data.co.uk/new/BRA.csv")
df_2025 = df[df["Season"] == 2025]
df_2025 = df_2025[:-90]# a cada 10 últimas linhas retiradas tira-se uma rodada
df_2025[["Date", "Home", "Away", "HG", "AG"]].head()

#diretorio_atual = os.getcwd()
#print(f"O arquivo será salvo em: {diretorio_atual}")
#df_2025.to_csv('saidas\brasileirao2025.csv')

#HG = Full Time Home Goals
#AG = Full Time Away Goals
#print(df_2025)

#imprime a média de gols marcados em casa e fora
print(df_2025[["HG", "AG"]].mean())
print(df_2025[["Home", "Away", "HG", "AG"]].values)
# essas médias mostram que os times marcam mais gols em casa (Home adavantage)
##print(df)
```

```
HG    1.472414
AG    0.962069
dtype: float64
[['Cruzeiro' 'Mirassol' 2.0 1.0]
 ['Fortaleza' 'Fluminense' 2.0 0.0]
 ['Gremio' 'Atletico-MG' 2.0 1.0]
 ...
 ['Vitoria' 'Corinthians' 0.0 1.0]
 ['Fluminense' 'Internacional' 1.0 0.0]
 ['Sport Recife' 'Mirassol' 1.0 2.0]]
```

Fonte: Autora.

Figura 5.2 – Implementação da função de verossimilhança para o modelo de Poisson.

```python
def log_likelihood(
    goals_home_observed,
    goals_away_observed,
    home_attack,
    home_defence,
    away_attack,
    away_defence,
    home_advantage,
):
    goal_expectation_home = np.exp(home_attack + away_defence + home_advantage)
    goal_expectation_away = np.exp(away_attack + home_defence)

    if goal_expectation_home < 0 or goal_expectation_away < 0:
        return 10000

    home_llk = poisson.logpmf(goals_home_observed, goal_expectation_home)
    away_llk = poisson.logpmf(goals_away_observed, goal_expectation_away)

    log_llk = home_llk + away_llk

    return -log_llk
```

Fonte: Autora.

Figura 5.3 – Implementação da função de correção do modelo de Dixon-Coles.

```python
def rho_correction(goals_home, goals_away, home_exp, away_exp, rho):
    if goals_home == 0 and goals_away == 0:
        return 1 - (home_exp * away_exp * rho)
    elif goals_home == 0 and goals_away == 1:
        return 1 + (home_exp * rho)
    elif goals_home == 1 and goals_away == 0:
        return 1 + (away_exp * rho)
    elif goals_home == 1 and goals_away == 1:
        return 1 - rho
    else:
        return 1.0
```

Fonte: Autora.

Figura 5.4 – Implementação da função de verossimilhança para o modelo de Dixon-Coles.

```python
def log_likelihood(
    goals_home_observed,
    goals_away_observed,
    home_attack,
    home_defence,
    away_attack,
    away_defence,
    home_advantage,
    rho,
):
    goal_expectation_home = np.exp(home_attack + away_defence + home_advantage)
    goal_expectation_away = np.exp(away_attack + home_defence)

    if goal_expectation_home < 0 or goal_expectation_away < 0:
        return 10000

    home_llk = poisson.logpmf(goals_home_observed, goal_expectation_home)
    away_llk = poisson.logpmf(goals_away_observed, goal_expectation_away)
    adj_llk = rho_correction(
        goals_home_observed,
        goals_away_observed,
        goal_expectation_home,
        goal_expectation_away,
        rho
    )

    if goal_expectation_home < 0 or goal_expectation_away < 0 or adj_llk < 0:
        return 10000

    log_llk = home_llk + away_llk + np.log(adj_llk)

    return -log_llk
```

Fonte: Autora.

A terceira parte é o coração dos modelos e consiste na maximização da verossimilhança (otimização) e retorna os parâmetros que descrevem as forças de ataque e defesa das equipes, assim como o parâmetro da vantagem do mandante. Observamos que, na prática utilizamos a função *Minimize* do pacote *scipy.optimize*, bastando ajustar os dados de entrada dessa função, como mostramos na Figura 5.5. Ainda que estejamos maximizando a verossimilhança, na prática estamos minimizando a verossimilhança com o sinal oposto, isso justifica o sinal de menos no valor retornado pela função de verossimilhança, conforme Figuras 5.2 e 5.4.

Figura 5.5 – Implementação da maximização da verossimilhança.

```
def fit_poisson_model():
    teams = np.sort(np.unique(np.concatenate([df_2025["Home"], df_2025["Away"]])))
    n_teams = len(teams)

    params = np.concatenate((np.random.uniform(0.5, 1.5, (n_teams)),  # ataque
            np.random.uniform(0, -1, (n_teams)),  # defesa
            [0.25],  # vantagem mandante
            [-0.1], #rho
        ))

    def _fit(params, df, teams):
        attack_params = dict(zip(teams, params[:n_teams]))
        defence_params = dict(zip(teams, params[n_teams : (2 * n_teams)]))
        home_advantage = params[-2]
        rho = params[-1]

        llk = list()
        for idx, row in df_2025.iterrows():
            tmp = log_likelihood(row["HG"],row["AG"],attack_params[row["Home"]],
                defence_params[row["Home"]],attack_params[row["Away"]],
                defence_params[row["Away"]],home_advantage,rho
            )
            llk.append(tmp)
        return np.sum(llk)
    options = {"maxiter": 100,"disp": False,}

    constraints = [{"type": "eq", "fun": lambda x: sum(x[:n_teams]) - n_teams}]

    res = minimize(_fit,params,args=(df_2025, teams),constraints=constraints,
        options=options,)

    model_params = dict(zip(["attack_" + team for team in teams]
                  + ["defence_" + team for team in teams]+ ["home_adv","rho"],res["x"],))
    print("Log Likelihood: ", res["fun"])
    return model_params

model_params = fit_poisson_model()
```

Fonte: Autora.

A quarta parte consiste em utilizar os parâmetros para obter estimativas e outras informações sobre eventuais confrontos. Essa etapa é mais simples e pode ser feita para as finalidades desejadas. Por exemplo, na Figura 5.6 está implementada uma função que estima as probabilidades de possíveis placares para um jogo específico.

Figura 5.6 – Implementação da estimativa de placares para um jogo específico.

```python
def predict(home_team, away_team, params, max_goals=10):
    home_attack = params["a_" + home_team]
    home_defence = params["d_" + home_team]
    away_attack = params["a_" + away_team]
    away_defence = params["d_" + away_team]
    home_advantage = params["home_adv"]

    home_goal_expectation = np.exp(home_attack + away_defence + home_advantage)
    away_goal_expectation = np.exp(away_attack + home_defence)

    home_probs = poisson.pmf(list(range(max_goals + 1)), home_goal_expectation)
    away_probs = poisson.pmf(range(max_goals + 1), away_goal_expectation)

    probability_matrix = np.outer(home_probs, away_probs)

    return probability_matrix
```

```python
probs = predict( "Vasco","Sao Paulo", model_params, 10)
pprint(100*probs)
print(np.sum(np.tril(probs, -1)))
print(np.sum(np.triu(probs, 1)))
print(np.sum(np.diag(probs)))
```

Fonte: Autora.

5.2 RANKINGS

Os modelos apresentados produzem parâmetros de modelagem que nos permitem estabelecer um *ranking* das equipes de um campeonato, inclusive considerando de forma separada as capacidades ofensivas e defensivas.

Para ilustrar vamos considerar o campeonato brasileiro de 2025, com todas as 38 rodadas e 380 jogos. Esse conjunto de dados é grande o bastante para que tendências fiquem evidentes. A classificação final do campeonato é apresentada na Figura 5.7.

Os modelos de Poisson e de Dixon-Coles produziram valores de parâmetros muito próximos, sendo a maior diferença 0.004138. Portanto, vamos considerar apenas os resultados do Modelo de Dixon-Coles, uma vez que os resultados do Modelo de Poisson seriam muito próximos.

Começando pelo aspecto ofensivo das equipes, utilizamos o parâmetro associado à força de ataque que o modelo de Dixon-Coles forneceu para estabelecer o *ranking* apresentado na Figura 5.8.

Analisando o aspecto defensivo das equipes, ou seja, considerando o parâmetro associado à força defensiva das equipes, fornecidos pelo modelo de Dixon-Coles, para fazer um *ranking*

Figura 5.7 – Classificação final do Campeonato Brasileiro de Futebol Masculino, Série A, de 2025.

	Club	MP	W	D	L	GF	GA	GD	Pts
1	Flamengo	38	23	10	5	78	27	51	**79**
2	Palmeiras	38	23	7	8	66	33	33	**76**
3	Cruzeiro	38	19	13	6	55	31	24	**70**
4	Mirassol	38	18	13	7	63	39	24	**67**
5	Fluminense	38	19	7	12	50	39	11	**64**
6	Botafogo	38	17	12	9	58	38	20	**63**
7	Bahia	38	17	9	12	50	46	4	**60**
8	São Paulo	38	14	9	15	43	47	-4	**51**
9	Grêmio	38	13	10	15	47	50	-3	**49**
10	Bragantino	38	14	6	18	45	57	-12	**48**
11	Atlético Mineiro	38	12	12	14	43	44	-1	**48**
12	Santos	38	12	11	15	45	50	-5	**47**
13	Corinthians	38	12	11	15	42	47	-5	**47**
14	Vasco da Gama	38	13	6	19	55	60	-5	**45**
15	Vitória	38	11	12	15	35	52	-17	**45**
16	Internacional	38	11	11	16	44	57	-13	**44**
17	Ceará	38	11	10	17	34	40	-6	**43**
18	Fortaleza	38	11	10	17	43	58	-15	**43**
19	Juventude	38	9	8	21	35	69	-34	**35**
20	Sport	38	2	11	25	28	75	-47	**17**

Fonte: Confederação Brasileira de Futebol (2025).

das equipes, assim temos os resultados apresentados na Figura 5.9.

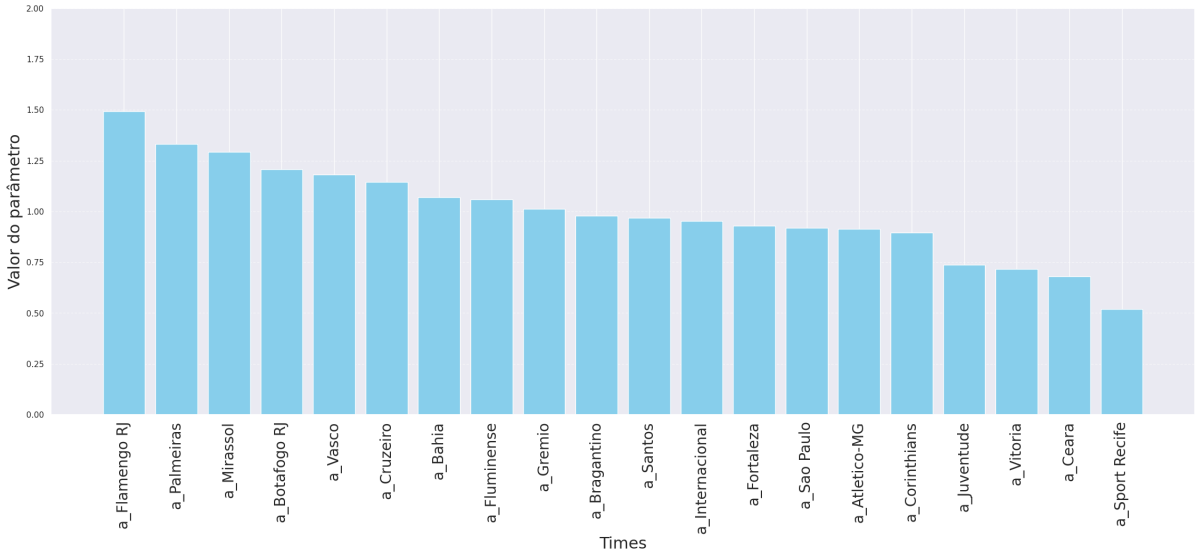
Comparando o resultado dos modelos com o resultado final do campeonato, apresentado na Figura 5.7, observamos que, para a maior parte das equipes, há coerência entre os *rankings* e a classificação final do campeonato.

As situações das equipes Ceará e Vasco nos chamam a atenção, pela grande diferença entre os parâmetros de força ofensiva e defensiva.

Por um lado, temos o Vasco com o terceiro pior parâmetro defensivo, o quinto maior parâmetro ofensivo e classificação final na décima quarta posição.

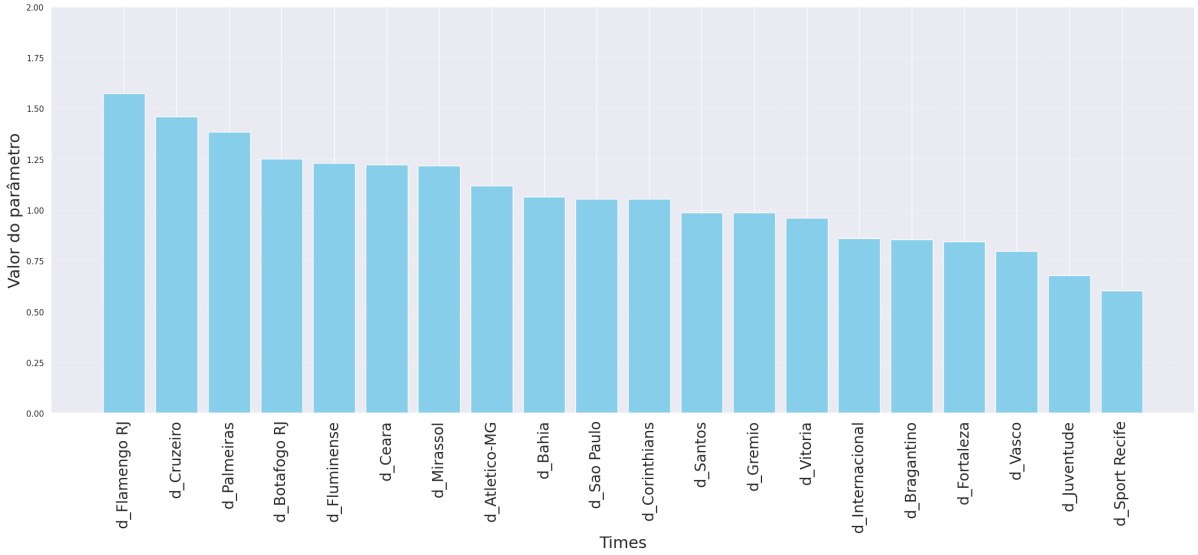
Já a equipe do Ceará, por sua vez, tem o sexto melhor parâmetro defensivo, o segundo

Figura 5.8 – Ranking baseado no parâmetro associado à força de ataque, obtido pelo modelo de Dixon-Coles, dos times do Campeonato Brasileiro de Futebol, na temporada de 2025.



Fonte: Autora.

Figura 5.9 – Parâmetro associado a força de defesa, obtido pelo modelo de Dixon-Coles, dos times do Campeonato Brasileiro de Futebol, na temporada de 2025.



Fonte: Autora.

pior parâmetro ofensivo e terminou o campeonato na décima sétima posição, sendo assim rebaixado para a série B do campeonato.

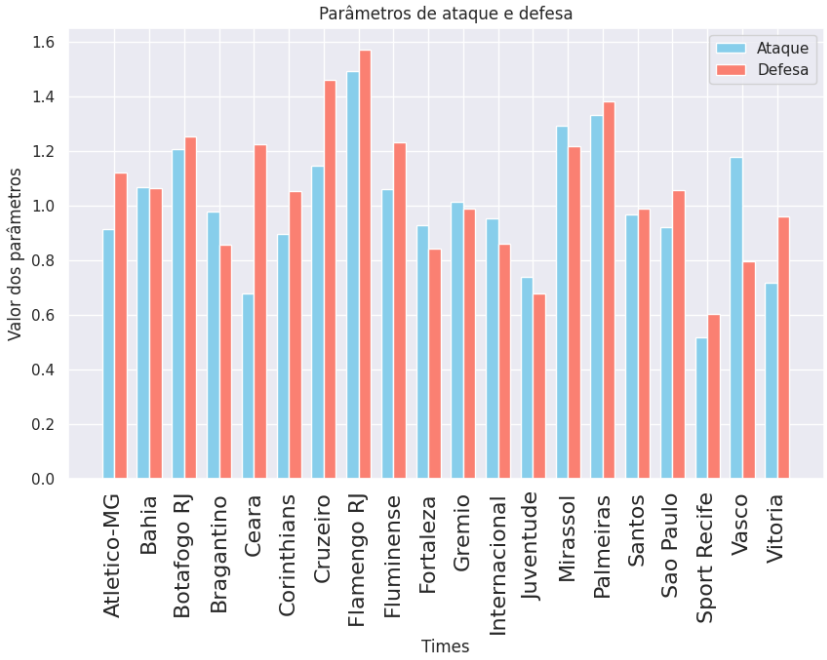
Devido a natureza do modelo, os parâmetros estão bastante relacionados as quantidades de gols marcados e concedidos. Portanto, mesmo que uma equipe marque muitos gols, não necessariamente isso significa que ela vença muitos jogos ou mesmo que o saldo de gols seja muito grande ao fim do campeonato, uma vez que esse quesito também depende da quantidade

de gols tomados.

Para exemplificar, supondo que uma equipe jogue 3 rodadas com os seguintes placares: **5**x0, 2x**0**, **2**x3, destacando o resultado da equipe. Após as 3 partidas, a equipe marcou 7 gols e tomou 5 gols, então possui saldo positivo de 2 gols. Porém, a equipe possui uma vitória e duas derrotas, marcando apenas 3 de 9 pontos possíveis, dessa forma, marcar muitos gols não significa vencer muitos jogos.

Na Figura 5.10 apresentamos, em ordem alfabética, os parâmetros de força de ataque e de defesa das equipes do campeonato em análise.

Figura 5.10 – Parâmetros associados à força de ataque e de defesa, obtido pelo modelo de Dixon-Coles, dos times do Campeonato Brasileiro de Futebol, na temporada de 2025.



Fonte: Autora.

5.3 ESTIMATIVA × RESULTADOS

Na trigéssima sétima rodada do campeonato Brasileiro de 2025, o Cruzeiro e o Botafogo se enfrentaram com mando de campo do Cruzeiro e o placar final foi um empate com dois gols para cada equipe.

Considerando todos os jogos até a trigéssima sexta rodada, o modelo de Dixon-Coles previu probabilidades de 52,96% de vitória do Cruzeiro, 18,46% de vitória do Botafogo e 28,58% de empate.

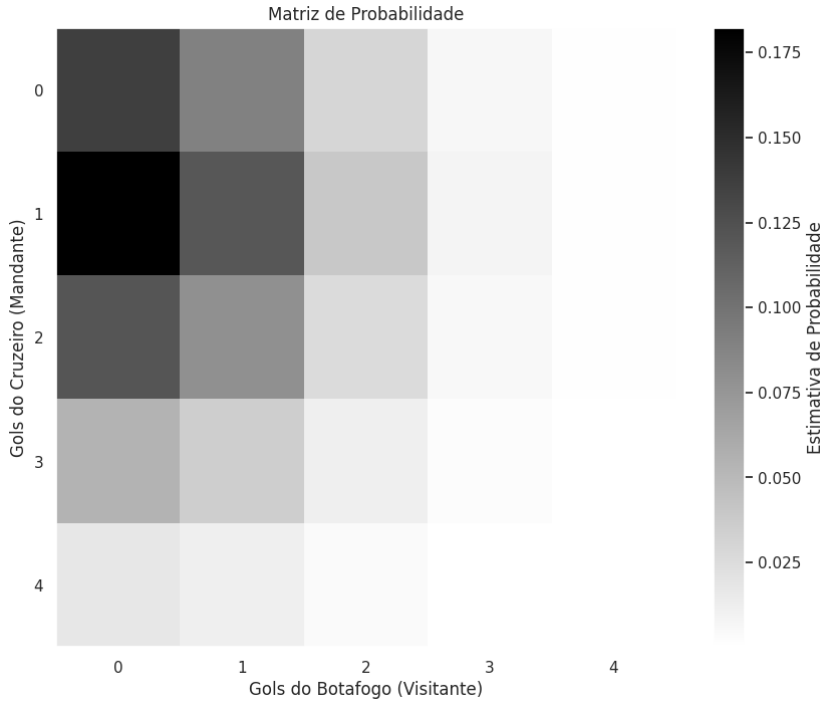
Na Figura 5.11 apresentamos a matriz de probabilidade dos placares, considerando até 4 gols. O placar mais provável, segundo a estimativa do modelo de Dixon-Coles, era a vitória do

Cruzeiro por 1 a 0, com probabilidade de 18,21%. Enquanto que o placar que de fato ocorreu tinha probabilidade de apenas 2,62%.

O resultado ocorrido não está entre os mais prováveis, o que mostra que as estimativas de probabilidade são muitas vezes falhas. Afinal, o evento é bastante complexo e depende de muitas variáveis que não são consideradas nos modelos.

Esse exemplo mostra quão difícil é uma estimativa de probabilidade se realizar. E, serve também, para nos mostrar que estimar placares é ainda mais difícil do que estimar apenas o resultado final de um jogo (vitória, empate ou derrota).

Figura 5.11 – Probabilidade de gols para Cruzeiro X Botafogo RJ, na trigéssima sétima rodada do Campeonato Brasileiro de Futebol, na temporada de 2025.



Fonte: Autora.

Para uma análise mais justa, quanto a qualidade das estimativas de probabilidades, do modelo de Dixon-Coles, vamos aplicá-lo para todos os jogos de uma mesma rodada do campeonato. Escolhemos a trigéssima primeira rodada do Campeonato Brasileiro de Futebol, série A, de 2025. Dessa forma, consideramos todos os jogos anteriores a essa rodada para determinar os parâmetros.

Na Tabela 5.1 está destacado o time vencedor e os placares das rodadas que os times empataram. Nas duas Figuras 5.12 e 5.13, constam as matrizes de probabilidade de cada um dos jogos, em que os resultados mais prováveis estão representados com tons mais escuros e os menos prováveis com tons mais claros. Em cada uma das matrizes de probabilidade há um ponto branco representando o resultado daquela partida.

Tabela 5.1 – Resultados da 31ª rodada do Campeonato Brasileiro de Futebol 2025

Mandante	Visitante	Placar
Santos	Fortaleza	**1x1**
Cruzeiro	Vitória	3x1
Mirassol	Botafogo	**0x0**
Flamengo	Sport	3x0
Bahia	Bragantino	2x1
Ceará	Fluminense	2x0
Corinthians	Grêmio	2x0
Juventude	**Palmeiras**	0x2
Internacional	Atlético-MG	**0x0**
Vasco	**São Paulo**	0x2

Fonte: Autora.

A seguir, algumas constatações sobre os resultados obtidos:

Se observarmos a matriz de probabilidade da Figura 5.12(a), referente a partida Santos x Fortaleza, a estimativa de placar mais provável seria 1 x 0, porém o resultado da partida foi o segundo mais provável, 1 x 1.

Ao analisarmos esses resultados, constatamos que em 3 das 10 partidas, os placares reais estão entre os 2 mais prováveis, segundo as estimativas apresentadas nas matrizes de probabilidade.

Se ampliamos a análise para os 4 placares mais prováveis, conforme a estimativa do modelo, observamos que 70% dos jogos tiveram seus resultados reais entre esses 4 placares. Esse percentual de acerto, aumenta para 80% se consideramos os 5 placares mais prováveis.

A análise comparativa entre as probabilidades estimadas e os resultados reais das partidas representadas nas Figuras 5.12(b) e 5.13(d) servem para nos lembrar que o modelo utilizado leva em consideração apenas o potencial de cada equipe em marcar e conceder gols. Situações atípicas e circunstanciais não são levadas em consideração.

Na partida Vasco x São Paulo, que foi a partida que gerou o resultado mais discrepante de todos os jogos da rodada, o modelo não levou em consideração que nos acréscimos do 1º tempo a equipe do Vasco levou 1 gol de pênalti e isso pode ter abalado o emocional da equipe, impactando no desempenho no 2º tempo, assim como pode ter motivado a equipe do São Paulo a ser mais ofensiva na volta do intervalo, por exemplo.

Vale lembrar que aspectos como clima, altitude, tipo do gramado, emocional dos jogadores, torcidas, partidas com caráter decisivo, jogadores que possuem cartão amarelo, jogadores suspensos ou lesionados, entre outros fatores, podem interferir significativamente nos resultados.

Os modelos mais precisos utilizados pelas casas de apostas certamente levam mais aspectos em consideração, possuindo assim estimativas melhores.

Embora o modelo de Dixon-Coles seja bem mais simples quando comparado com os modelos que as casas de apostas utilizam, os resultados se mostram muito bons para os eventos que analisamos.

5.4 ESTIMATIVAS × ODDS

Não temos acesso aos modelos utilizados pelas casas de apostas, portanto a única forma de comparar o modelo de Dixon-Coles com tais modelos é através das *odds* que são propostas por tais casas de apostas. A variação entre *odds* de diferentes casas de apostas é pequena, pois variações grandes permitiriam a arbitragem.

A arbitragem consiste em realizar apostas em todas as opções possíveis, utilizando as *odds* de diferentes casas de apostas, a fim de obter lucro sem correr grandes riscos.

Desta forma, nesta seção vamos comparar as *odds* estimadas pelo modelo de Dixon-Coles com as praticadas em casas de apostas.

Para a mesma rodada analisada na Seção 5.3, vamos comparar as *odds* observadas em casas de apostas e as *odds* obtidas a partir das probabilidades fornecidas pelo modelo.

Como já mencionamos na Introdução, as *odds* oferecidas nas casas de apostas tem um acréscimo de forma a garantir o lucro das mesmas, contudo isso não deve ter grande impacto na comparação.

Lembramos que a *odd* de um evento é o inverso da probabilidade desse evento.

Na Tabela 5.2 comparamos as *odds* estimadas pelo método de Dixon-Coles com as praticadas pela casa de apostas Bet365 e as disponibilizadas pelo site Agora (ODDSAGORA, 2025), cuja a fonte das *odds* não foi mencionada.

Tabela 5.2 – Comparação entre as *odds* estimadas pelo modelo de Dixon-Coles e as *odds* praticadas a época.

Rodada 31	Estimativas Modelo DC			Odds site Agora			Odds bet365		
Jogo	Mand.	Vis.	Empate	Mand.	Vis.	Empate	Mand.	Vis.	Empate
San. x Fort.	1,91	4,62	3,86	1,70	5,03	3,67	1,72	5,00	3,60
Cru X Vit	1,32	15,22	5,68	1,41	7,92	4,40	1,45	8,00	4,20
Mir. X Bot.	2,07	3,95	3,79	2,32	3,09	3,27	2,25	3,10	3,50
Fla X Sport	1,10	60,71	13,54	1,22	12,65	6,25	1,16	15,00	7,50
Bahia X Brag.	1,48	7,32	5,28	1,65	5,40	3,68	1,66	5,50	3,70
Ceará x Flu	2,28	4,02	3,19	3,02	2,62	2,86	3,10	2,55	3,00
Cor. x Gre.	2,02	4,55	3,51	1,64	5,45	3,72	1,75	4,75	3,80
Juv. x Pal	8,79	1,42	5,51	6,08	1,60	3,69	6,50	1,57	3,75
Int. x Atl	2,15	4,01	3,50	2,03	3,98	3,12	1,95	4,33	3,25
Vas. x SP	1,81	4,64	4,34	2,24	3,56	3,00	2,15	3,50	3,25

Fonte: Autora.

Podemos observar grande compatibilidade entre as *odds* estimadas pelo modelo de Dixon-Coles e as praticadas pelas casas de apostas.

Observamos que exceto pelos jogos Ceará x Fluminense, Juventude x Palmeiras e Vasco x São Paulo, as *odds* do mandante estimadas pelo modelo de Dixon-Coles diferem em no máximo $0,4$ das *odds* consultadas.

Ao analisarmos as *odds* estimadas para os visitantes, aí notamos um aumento signifigativo nas diferenças. Nos jogos Cruzeiro x Vitória, Flamengo x Sport, Bahia x Bragantino, Ceará x Fluminense e Vasco x São Paulo, as estimativas de *odds* dos visitantes do modelo Dixon-Coles diferem acima de $1,0$ das *odds* dos sites comparados.

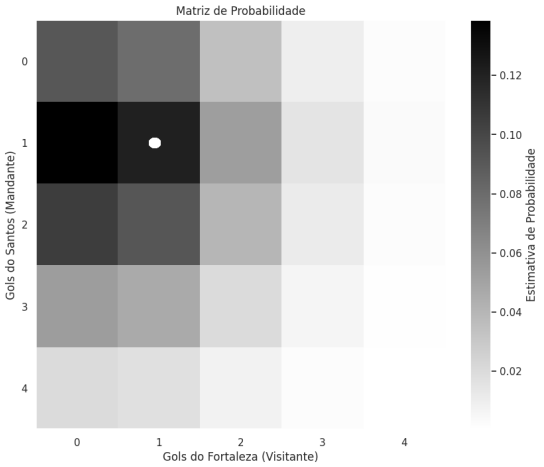
As partidas entre Cruzeiro x Vitória e Flamengo x Sport, têm a maior diferença entre as *odds* estimadas e as das casas de apostas.

Agora, ao analisarmos as *odds* estimadas para empate, aí notamos um aumento signifigativo nas diferenças. Nos jogos Cruzeiro x Vitória, Flamengo x Sport, Bahia x Bragantino, Juventude x Palmeiras e Vasco x São Paulo, as estimativas de *odds* para empate do modelo Dixon-Coles diferem acima de $1,0$ das estimativas dos sites comparados.

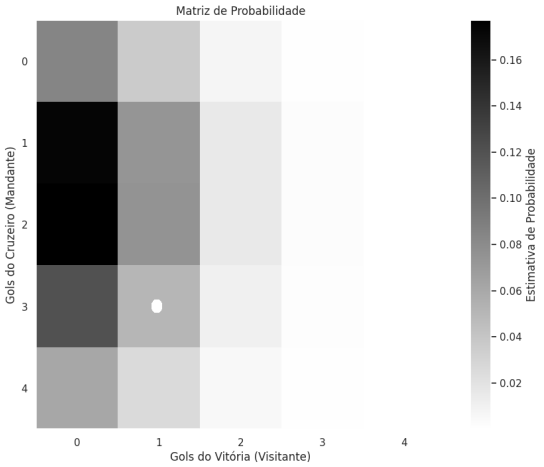
Analisando os resultados das partidas dessa rodada, é perceptível que o mandante tem uma vantagem, tanto que nesta rodada 50% dos jogos foram ganhos pelos mandantes. Isso está de acordo com a utilização de um parâmetro de vantagem do mandante, como prevêm os modelos.

Entretanto, a vantagem do mandante é considerada a mesma para todas as equipes. Isso faz sentido em campeonatos com caracteríticas homogêneas de torcida, condição/tipo de gramado, clima e logística/deslocamento como, por exemplo, ocorre no campeonato Inglês (Premier League). No Campeonato Brasileiro, o mais adequado seria um parâmetro de vantagem do mandante para cada equipe, pois o Brasil tem dimensões continentais que impactam muito na logística de deslocamentos das equipes, os campos não são padronizados, e os tamanhos de estádios e torcidas são bastante diversos.

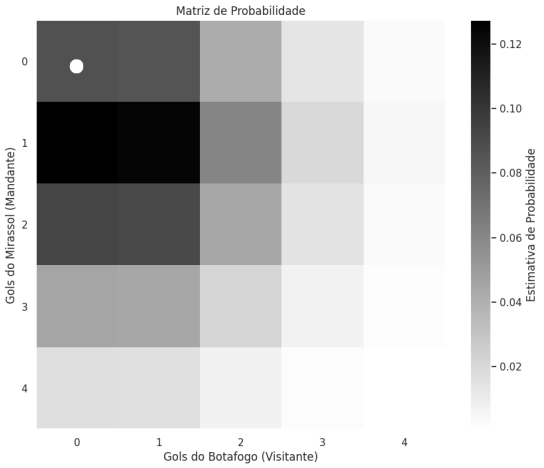
Figura 5.12 – Parte 1 das estimativas de probabilidade e resultados reais da trigésima primeira rodada do Campeonato Brasileiro de 2025, masculino, série A.



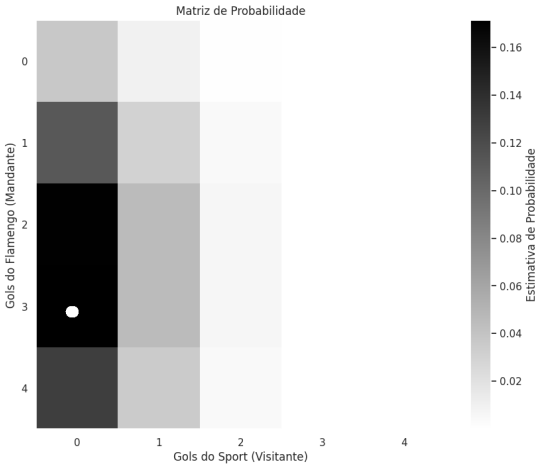
(a) Santos X Fortaleza



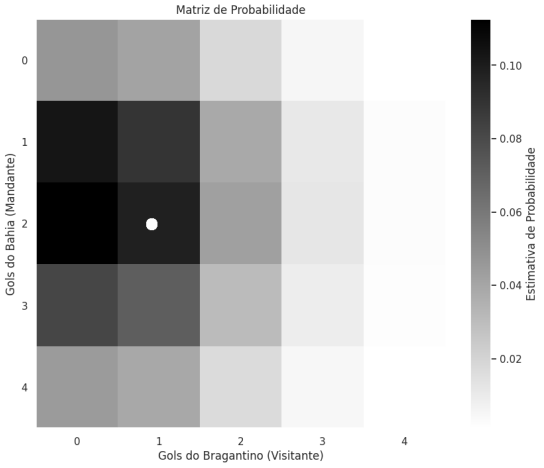
(b) Cruzeiro X Vitória



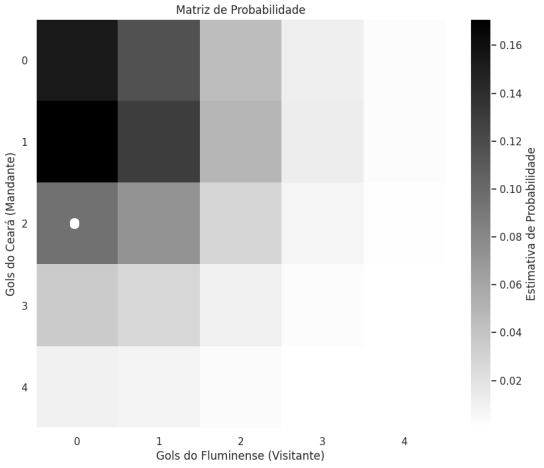
(c) Mirassol X Botafogo



(d) Flamengo X Sport



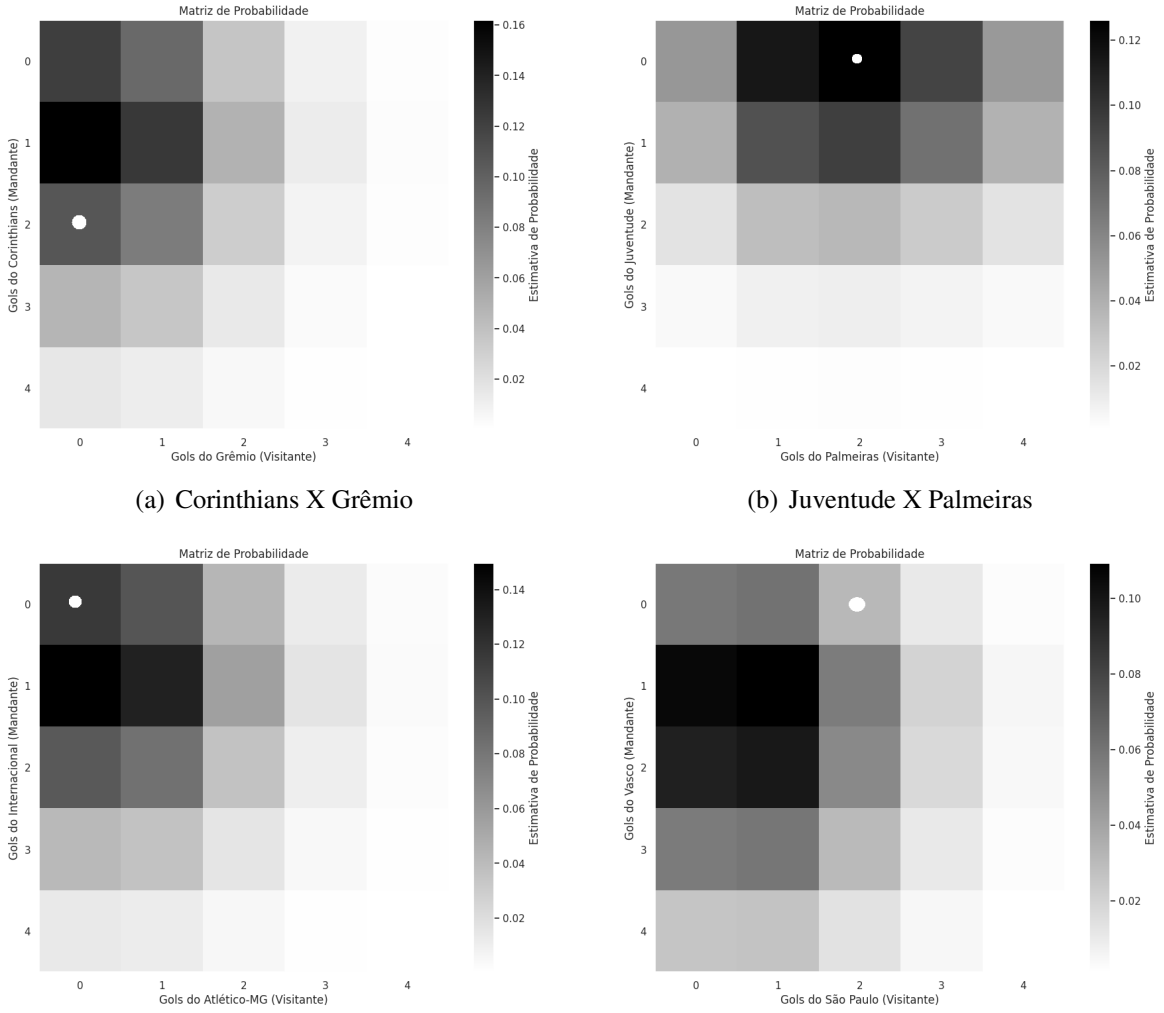
(e) Bahia X Bragantino



(f) Ceará X Fluminense

Fonte: Autora.

Figura 5.13 – Parte 2 das estimativas de probabilidade e resultados reais da trigésima primeira rodada do Campeonato Brasileiro de 2025, masculino, série A.



(a) Corinthians X Grêmio

(b) Juventude X Palmeiras

(c) Internacional X Atlético-MG

(d) Vasco X São Paulo

Fonte: Autora.

6 PROPOSTA DE ATIVIDADE

Este capítulo tem por finalidade apresentar uma sequência de atividades relacionadas com o tema deste trabalho e é voltada para alunos do Ensino Médio. Acreditamos que as atividades podem ser aplicadas e/ou adaptadas para os três anos do Ensino Médio regular, portanto, não indicaremos o ano destinado à cada atividade.

O objetivo é proporcionar aos estudantes um melhor entendimento do tema, possibilitar que realizem cálculos envolvendo as probabilidades de ocorrência de determinados eventos e desenvolvam a capacidade de análise crítica sobre o assunto, prevenindo decisões que lhes sejam prejudiciais.

As quatro primeiras propostas de atividades têm caráter introdutório ao assunto, incluindo a definição do conceito, cálculo e a interpretação das *odds*, cálculo das probabilidades implícitas e o cálculo do valor esperado, seguidos de uma discussão sobre o tema. Por fim, a quinta atividade, aborda a temática de forma avançada, apresentando o modelo de Poisson, sua fórmula e realizando cálculos utilizando a mesma.

6.1 ATIVIDADE 1

A primeira atividade é definir o conceito de *odd* e como ela é apresentada no Brasil, *odds* decimais, mas para isso o professor deve retomar conceitos básicos de porcentagem e probabilidade.

Assim, o primeiro momento da aula é destinado a relembrar como se calcula a porcentagem de uma quantia e a probabilidade de um evento ocorrer, bem como qualquer outro conceito que o professor julgue necessário para dar continuidade a aula. Seguido da explicação de *odd*.

A *odd* decimal é um número com até duas casas decimais, que representa o fator pelo qual o valor apostado será multiplicado caso a aposta seja vencedora. As *odds* decimais são aproximadamente o inverso da probabilidade: $odd = \frac{1}{P}$.

Após a introdução, o estudante será solicitado a responder à seguinte questão:

1) Supondo uma partida de futebol com dois times, A e B, com probabilidade estimada de vitória: time A: 50%, time B: 20% e empate: 30%.

a) Calcule as *odds* decimais utilizando a fórmula $odd = \frac{1}{P}$.

b) Supondo que um torcedor do time A realize uma aposta fictícia de $R\$30,00$. Calcule o retorno esperado para essa aposta.

c) Supondo que um torcedor do time B realize uma aposta fictícia de $R\$15,00$. Calcule o retorno esperado para essa aposta.

d) Supondo que a partida tenha terminado empatada, os apostadores citados em b) e c) teriam lucro ou prejuízo? De quanto?

Essa atividade aborda os conteúdos: probabilidade, frações, porcentagens, razão, proporção e interpretação de resultados.

6.1.1 RESPOSTA COMENTADA

Segue a resposta para a questão proposta.

1) a) Para a vitória do time A, $50\% = 0,5$, assim utilizando a fórmula teríamos:

odd $= \frac{1}{P} = \frac{1}{0,5} = 2$.

Para a vitória do time B, $20\% = 0,2$:

odd $= \frac{1}{P} = \frac{1}{0,2} = 5$.

E, para o empate, $30\% = 0,3$:

odd $= \frac{1}{P} = \frac{1}{0,3} \approx 3,3333$.

Logo, as *odds* para vitória do time A, vitória do time B e empate seriam, respectivamente, $2,00$, $5,00$ e $3,33$.

b) O retorno esperado para a aposta seria dado pelo valor da aposta multiplicado pela *odd*, menos o valor apostado, desta forma:

retorno esperado $= 30,00$ x $2 = 60,00 - 30,00 = 30,00$.

c) O retorno esperado seria:

retorno esperado $= 15,00$ x $5 = 75,00 - 15,00 = 60,00$.

d) Em caso de empate, os dois apostadores teriam prejuízo, o torcedor do time A teria um prejuízo de $R\$30,00$ e o torcedor do time B de $R\$15,00$.

6.2 ATIVIDADE 2

Essa atividade busca trabalhar as *odds* implícitas e a "vantagem do time da casa". Interpretando as *odds* reais e as probabilidades implícitas, buscando desenvolver o pensamento crítico dos alunos.

O primeiro momento da aula é destinado a explicação de *odds* implícitas.

Ao observarmos as *odds* conseguimos analisar a expectativa da casa de apostas, pois elas são inversamente proporcionais a chance/probabilidade do evento se concretizar, ou seja, quanto maior a *odd*, menor a probabilidade do evento ocorrer e inversamente uma *odd* pequena (próxima de 1) indica grande probalidade do evento ocorrer.

Após a apresentação do tema, o estudante será solicitado a responder à seguinte questão:

1) Supondo as *odds* fictícias de uma casa de apostas para uma partida de futebol entre os times A e B: viória do time A: $1,80$, vitória do time B: $4,50$ e empate: $3,60$.

a) Calcule as probabilidades implícitas representadas por essas *odds*.

b) Some todas as probabilidades, essa soma representa quantos por cento?

c) Note que o total de b) é maior que 100%. Agora responda, por que a soma ultrapassou 100% e onde está o lucro da casa de apostas?

Conteúdos desenvolvidos com essa atividade: porcentagem, probabilidade inversa, análise crítica de dados e matemática financeira básica.

6.2.1 RESPOSTA COMENTADA

Segue a resposta para a questão proposta.

1) a) Para a vitória do time A, considerando a *odd* $1,80$, utilizando a fórmula teríamos:

$odd = \frac{1}{P} \implies P = \frac{1}{1,8} \approx 0,56 \approx 56\%$.

Para a vitória do time B, $odd = 4,50$: $odd = \frac{1}{P} \implies P = \frac{1}{4,5} \approx 0,22 \approx 22\%$.

E, para empate, $odd = 3,60$: $odd = \frac{1}{P} \implies P = \frac{1}{3,6} \approx 0,28 \approx 28\%$.

Logo, as probabilidades implícitas de vitória do time A, vitória do time B e empate são, respectivamente, 56%, 22% e 28%.

b) Somando todas as probabilidades: $P_{total} = 56\% + 22\% + 28\% \approx 106\%$, teremos que a soma representa 106%.

c) A soma ultrapassou 100%, pois a casa de apostas obtém lucro de alguma forma e esse lucro encontra-se exatamente na diferença entre as porcentagens, ou seja, a casa de apostas espera obter aproximandamente 6% de lucro.

6.3 ATIVIDADE 3

Essa atividade busca simular apostas e analisar as expectativas matemáticas, tendo como objetivo introduzir o conceito de valor esperado e analisar a tomada de decisão baseada em probabilidade.

O primeiro momento da aula é destinado a explicação e definição de valor esperado de uma variável aleatória, ($E(X)$).

Sugerimos 3 formas de conceituar/definir o valor esperado e o professor decide qual definição utilizar de acordo com o nível da turma que está trabalhando.

Definição 1 (Nível Avançado): Seja X uma variável aleatória discreta que assume valores $x_1, x_2, \ldots, x_n$, com função massa de probabilidade $P(X = x_i)$. O valor esperado de X é

definido por:

$$\mathbb{E}(X) = \sum_{i=1}^{n} x_i \, P(X = x_i).$$

Definição 2 (Nível Intermediário): O valor esperado de uma variável aleatória X é o resultado médio que se espera obter, considerando todas as possibilidades e suas probabilidades. Se X pode assumir os valores $x_1, x_2, \ldots, x_n$ com probabilidades $P(X = x_i)$, então:

$$\mathbb{E}(X) = x_1 \, P(X = x_1) + x_2 \, P(X = x_2) + \cdots + x_n \, P(X = x_n) = \sum_{i=1}^{n} x_i \, P(X = x_i).$$

Definição 3 (Nível Básico): O valor esperado de uma situação aleatória é o resultado médio que se espera obter se a situação fosse repetida muitas vezes. Ele considera todos os resultados possíveis e a probabilidade de cada um e poderia ser calculado da seguinte forma:

$E(X) = P(\text{ganhar}) \cdot \text{ganho} - P(\text{perder}) \cdot \text{perda}$

Essa definição é a mais simples das três, pois omite o sinal do somátio, tornando a aparência da definição mais simples.

Depois de definir o que é o valor esperado, o estudante será solicitado a responder as seguintes questões:

1) Considere o seguinte cenário, um time com probabilidade de vitória de 40%, a *odd* oferecida pela casa de aposta é de $3,00$. Supondo uma aposta de $R\$10,00$.

a) Calcule o valor esperado, $E(X)$, para essa aposta.

b) Discuta com a turma se a aposta é vantajosa ou não, explicando seu ponto de vista.

2) Considere a probabilidade de vitória do time A, $0,4$, de derrota $0,6$, e uma aposta de $R\$10,00$. Preencha a tabela a seguir, sorteando 10 números aleatórios entre 0 e 1, considerando cada número 1 aposta.

Tabela 6.1 – Simulação de apostas

Aposta	**Número sorteado**	**Resultado**	**Lucro/Perda (R$)**
1	0,25	Vitória	+20
2			
3			
4			
5			
6			
7			
8			
9			
10			

Fonte: Autora.

Com base na situação do problema, a regra para determinar o resultado seria:

número $\leq 0,4 \implies$ vitória ($+R\$20$) e número $> 0,4 \implies$ derrota ($-R\$10$).

a) Com base na tabela calcule o lucro total obtido somando os resultados.

b) Calcule o lucro médio por aposta.

c) Comparando o valor esperado (já calculado em 1)a)) e o lucro médio por aposta, o que é possível observar?

O professor pode alterar a *odd* e/ou a probabilidade do Exercício 2 e utilizá-las como uma questão complementar.

Os conteúdos trabalhados foram: probabilidade, operações com números decimais, introdução à estatística e educação financeira.

6.3.1 RESPOSTA COMENTADA

Segue a resposta para a questão proposta.

Por simplicidade, indicaremos a resposta para o valor esperado utilizando apenas a Definição 3.

1) a) Considerando uma aposta de $R\$10,00$ com probabilidade de vitória de 40% e uma *odd* de $3,00$, teríamos:

Lucro esperado em caso de vitória: 10 x $3 - 10 = 30 - 10 = 20$.

Perda esperada, em caso de derrota ou empate, é o valor apostado, $R\$10,00$.

$P(\text{ganhar}) = 40\% = 0,4$ e $P(\text{perder}) = 60\% = 0,6$.

Calculando o valor esperado, por meio da Definição 3:

$E(X) = P(\text{ganhar}) \cdot \text{ganho} - P(\text{perder}) \cdot \text{perda} = 0,4 \cdot 20 - 0,6 \cdot 10 = 8 - 6 = 2$.

Assim, o valor esperado seria de $R\$2,00$, significa que, em média, a cada jogo o apostador ganharia $R\$2,00$.

b) Questão aberta.

2) Completando a tabela com números aleatórios, teríamos:

a) O lucro, L, total obtido seria:

$L = 20 + 20 - 10 - 10 - 10 + 20 - 10 + 20 - 10 - 10 = 80 - 60 = 20$

b) O lucro médio por aposta, L_m, seria: $L_m = \frac{L}{10} = \frac{20}{10} = 2$

c) O valor esperado calculado na Questão 1)a) é $R\$2,00$ e o lucro médio por aposta é $R\$2,00$. Mostrando assim, que o valor esperado representa o comportamento "em média" da aposta.

É evidente que, para diferentes números sorteados, chegaríamos a diferentes valores de

Tabela 6.2 – Simulação de apostas com números aleatórios - resposta

Aposta	Número sorteado	Resultado	Lucro/Perda (R$)
1	0,25	Vitória	+20
2	0,3	Vitória	+20
3	0,7	Derrota	-10
4	0,5	Derrota	-10
5	0,83	Derrota	-10
6	0,15	Vitória	+20
7	0,49	Derrota	-10
8	0,4	Vitória	+20
9	0,9	Derrota	-10
10	0,41	Derrota	-10

Fonte: Autora.

lucro médio e, possivelmente, não encontraríamos $E(X) = L_m$, entretanto é válido considerar $E(X) \approx L_m$.

6.4 ATIVIDADE 4

Essa atividade possui caráter informativo e visa fomentar a reflexão e o debate sobre o tema, conduzindo à compreensão do impacto potencialmente prejudicial das apostas.

Tendo realizado as atividades 1, 2 e 3, o professor deve propor que os estudantes realizem uma roda de conversa e apresentem os resultados obtidos. Após esse momento inicial, recomenda-se mostrar/citar algumas reportagens, tais como:

- De acordo com Ribbeiro (RIBBEIRO, 2024), “Beneficiários do Bolsa Família gastaram R$ 3 bi com "bets"em agosto, diz BC”;
- Segundo Lima e Cunha (LIMA; CUNHA, 2024), “Bets: 42% dos brasileiros que dizem apostar estão endividados e quase um terço está fora do mercado de trabalho, diz pesquisa do Senado”;
- Para Zorzetto e Orlandi (ZORZETTO; ORLANDI, 2024) "Proliferação das bets aumenta gastos de famílias e risco de problemas com o jogo. Brasileiros já apostam R$ 20 bilhões por mês em plataformas digitais e cresce a procura por tratamento para dependentes";
- Para Bittencourt (BITTENCOURT, 2024), “Apostas on-line atraem crianças e adolescentes, apesar de ilegais”.
- Para Zorzetto e Orlandi (ZORZETTO; ORLANDI, 2024) "Proliferação das bets aumenta gastos de famílias e risco de problemas com o jogo. Brasileiros já apostam R$ 20 bilhões por mês em plataformas digitais e cresce a procura por tratamento para dependentes".

É pertinente que o professor realize questionamentos como:

1. Você conhece alguém que realiza apostas com frequência? Você já apostou?
2. Tem conhecimento de casos em que a pessoa perdeu grandes quantias ou teve a vida prejudicada em função das apostas?
3. Qual é a sua opinião sobre os dados informados/reportagens citadas?

Após uma breve discussão sobre o assunto, o professor deve direcionar a discussão para o encerramento, utilizando perguntas como:

1. Mesmo com valor esperado positivo, todas as apostas geram lucro?
2. O valor esperado mostra o resultado médio ao longo de muitas repetições, isso garante resultado único?
3. Discuta como o risco influencia a tomada de decisão.
4. É facíl ganhar dinheiro apostando?

A intenção é que o professor conduza a discussão de modo a levar os alunos à conclusão de que os riscos são elevados e as chances de obter lucro com apostas são baixas.

6.4.1 RESPOSTA COMENTADA

Para a atividade proposta, não há uma resposta única, no entanto, espera-se que ela conduza os estudantes à conclusão de que as apostas tendem a ser mais prejudiciais do que benéficas.

6.5 ATIVIDADE 5

Esta atividade é sugerida para uma turma mais avançada, serve também como complemento após desenvolver as Atividades $1, 2, 3$ e 4.

O objetivo desta atividade é aplicar Poisson para calcular a probabilidade da ocorrência de um determinado número de gols em um jogo e relacionar probabilidades de Poisson com expectativa matemática em apostas.

No primeiro momento da aula, cabe uma breve explicação do que é o modelo de Poisson e como ele é utilizado para estimar o número de gols de uma partida de futebol.

O professor deve apresentar a fórmula de Poisson: $P(X = x) = \dfrac{\lambda^x e^{-\lambda}}{x!}$.

Vale relembrar alguns conhecimentos adquiridos com as atividades anteriores. Como, a interpretação de que a *odd* é o inverso da probabilidade, então a probabilidade pode ser obtida pelo inverso da *odd*. Logo, $P = \frac{1}{odd}$.

Após a apresentação do tema, o estudante será solicitado a responder as questões:

1) Considere uma partida entre os times A e B, cujo número médio de gols (λ) do time A é $1,8$ e do time B é $1,2$ gols/jogo. Supondo que o número de gols marcados por cada time siga uma distribuição de Poisson, responda:

a) Calcule a probabilidade de cada time marcar 0, 1, 2 ou 3 gols utilizando a fórmula de Poisson: $P(X = x) = \dfrac{\lambda^x e^{-\lambda}}{x!}$.

b) Utilizando as probabilidades do item a) calcule a probabilidade dos resultados 0x0, 1x0 e 1x1.

c) Supondo uma *odd* fictícia de $6,50$ para um placar de 1x0. Calcule a probabilidade implícita presente na *odd*.

d) Analise as *odds* de uma partida aleatória de futebol. Em sua opinião, por que resultados com muitos gols possuem *odds* mais altas?

2) Considere a seguinte situação: "o time A marca exatamente 2 gols", assumindo que o número médio de gols é $\lambda = 1{,}8$, que a casa de apostas oferece uma odd de $3{,}20$ e o valor apostado é de $R\$10{,}00$.

a) Qual é a probabilidade de sucesso da afirmação?

b) Qual é a probabilidade implícita da aposta?

c) Calcule o valor esperado da aposta utilizando a fórmula:

$E(X) = P(\text{ganhar}) \cdot \text{ganho} - P(\text{perder}) \cdot \text{perda}$

d) Discuta quando uma aposta é "favorável" ou "desfavorável" com base no cálculo.

6.5.1 RESPOSTA COMENTADA

Segue resposta para a atividade proposta.

1) a) Utilizando a fórmula de Poisson para calcular a probabilidade do time A, com $\lambda_A = 1,8$, marcar 0, 1, 2 ou 3 gols, teríamos:

$P_A(0) = \frac{1{,}8^0 e^{-1,8}}{0!} = e^{-1,8} \approx 0,1653,$

$P_A(1) = \frac{1{,}8^1 e^{-1,8}}{1!} = 1,8 \cdot e^{-1,8} \approx 0,2975,$

$P_A(2) = \frac{1{,}8^2 e^{-1,8}}{2!} = 1,62 \cdot e^{-1,8} \approx 0,2678$ e

$P_A(3) = \frac{1{,}8^3 e^{-1,8}}{3!} \approx 0,1607.$

Para o time B, com $\lambda_B = 1,2$, teríamos:

$P_B(0) = e^{-1,2} \approx 0,3012$,

$P_B(1) = 1,2 \cdot e^{-1,2} \approx 0,3615$,

$P_B(2) = \frac{1,2^2 e^{-1,2}}{2!} = 0,2169$ e

$P_B(3) = \frac{1,2^3 e^{-1,2}}{3!} \approx 0,0868$.

b) A probabilidade de um placar exato como, por exemplo, 0x0, é dada pelo produto das probabilidades de cada time:

$P(\text{placar 0x0}) = P_A(0) \cdot P_B(0) \approx 0,1653 \cdot 0,3012 \approx 0,0498 \approx 4,98\%$.

A probabilidade de 1x0 para o time A, é:

$P(\text{placar 1x0}) = P_A(1) \cdot P_B(0) \approx 0,2975 \cdot 0,3012 \approx 0,0896 \approx 8,96\%$.

Já a probabilidade de 1x1, é:

$P(\text{placar 1x1}) = P_A(1) \cdot P_B(1) \approx 0,2975 \cdot 0,3615 \approx 0,1075 \approx 10,75\%$.

Desta forma, podemos calcular placares específicos multiplicando as probabilidades de cada time.

c) Podemos calcular a probabilidade implícita, presente em uma *odd*, utilizando a fórmula: $P = \frac{1}{odd}$. Assim,

$P = \frac{1}{odd} = \frac{1}{6,5} \approx 0,1538 \approx 15,38\%$.

A questão pode ser estendida utilizando o mesmo enunciado e sugerindo *odds* diferentes.

d) Espera-se que o estudante perceba que partidas com muitos gols têm menor probabilidade, por isso, as *odds* são maiores. Uma vez que, placares com poucos gols são mais prováveis.

2) a) No evento, sucesso é o time A marcar exatamente 2 gols, probabilidade que já calculamos em 1) a): $P(sucesso) = P(X = 2) \approx 0,2678 \approx 26,78\%$.

b) A probabilidade implícita da aposta é: $P = \frac{1}{odd} = \frac{1}{3,2} = 0,3125 = 31,25\%$.

c) Para calcular o valor esperado da aposta precisamos da $P(\text{perda}) = P(\text{fracasso})$. Como, o fracasso representa qualquer outro número de gols, diferente do time A marcar exatamente 2 gols, podemos escrever: $P(\text{fracasso}) = 1 - P(X = 2) \approx 1 - 0,2678 \approx 0,7322 \approx 73,22\%$.

Supondo a aposta de $R\$10,00$ e a *odd* de $3,20$: o ganho seria de 10 x $3,2 = 32 - 10 = R\$22,00$ e a perda seria de $R\$10,00$.

Assim, o valor esperado da aposta é calculado por:

$E(X) = P(\text{ganhar}) \cdot \text{ganho} - P(\text{perder}) \cdot \text{perda} = 0,2678 \cdot 22 - 0,7322 \cdot 10 \approx 5,89 - 7,32 \approx -1,43$.

d) Como encontramos o valor esperado negativo, podemos concluir que a aposta é

desfavorável, ou seja, a longo prazo, a repetição desse tipo de aposta tende a gerar prejuízo ao apostador, mesmo que o evento possua uma probabilidade razoável de ocorrer.

7 CONCLUSÃO

Apresentamos como os modelos básicos de estimativa de probabilidades para eventos esportivos são construídos. Aplicamos esses modelos e corroboramos que os modelos têm funcionamento bastante adequado, se comparado ao que as casas de apostas propõem e aos resultados reais.

A análise dos resultados também serviu para exemplificar que ganhar das casas de apostas é uma tarefa muito difícil, mesmo possuindo um bom modelo matemático como o modelo de Dixon-Coles. São muitas as variáveis a considerar, probabilidades a calcular e opções de placares prováveis/possíveis a ponderar.

A matemática financeira tem ganho um espaço significativo nas aula do RCO+Aulas (SEED, 2022), tanto para o Ensino Fundamental como o Ensino Médio, uma aula voltada para a prática e desenvolvimento futuro do aluno. Isso justifica trabalhar os riscos das apostas e a dinânima das casas de apostas em sala de aula, por isso, do ponto de vista pedagógico, as propostas de atividade expostas no Capítulo 6 são relevantes.

Cada uma das atividades propostas tem uma intencionalidade, a primeira atividade é responsável por situar o estudante sobre o tema, expondo o conceito de *odd* e ensinando o cálculo. Já a segunda atividade, busca apresentar o "lucro da casa de apostas". A terceira atividade tem como objetivo incorporar o conceito de valor esperado. Já a quarta atividade busca concluir o tema, trazendo a relevância do assunto para a atualidade, onde espera-se que o aluno entenda que não é aconselhável apostar. E por fim, a quinta atividade serve como uma atividade complementar.

Como várias partes do trabalho tem aplicação no Ensino Médio, sugerimos as atividades propostas como material de apoio a professores de Matemática do Ensino Médio, para que possa ser trabalhado em sala de aula a fim de reduzir/desincentivar a participação de jovens e adolescentes na realização de apostas esportivas. Por outro lado, introduzimos uma análise estatística que pode motivar os estudantes a tomar decisões com base em dados, ainda que incompletos.

Ao abordarmos esse tema em sala de aula, estamos trabalhando a matemática financeira, pois de certa forma esperamos que os estudantes desenvolvam senso crítico quanto a gestão de riscos, não só em apostas mas em diversas situações. A ideia é alertá-los para que tomem decisões financeiras fundamentadas.

REFERÊNCIAS

AGRESTI, A. **Categorical Data Analysis**. 2a. ed. New Jersey: John Wiley & Sons, Inc., 2002. 710 p. 27

BITTENCOURT, C. **Apostas on-line atraem crianças e adolescentes, apesar de ilegais**. BBC News Brasil, 2024. Disponível em: <https://lunetas.com.br/apostas-on-line-atraem-criancas-e-adolescentes-apesar-de-ilegais/>. Acesso em: 06 dez. 2024. 55

BRASIL. **Base Nacional Comum Curricular**. Brasília, 2018. Disponível em: <https://www.gov.br/educacao/pt-br/assuntos/base-nacional-comum-curricular>. Acesso em: 14 abr. 2024. 12

DIXON, M.; COLES, S. Modelling association football scores and inefficiencies in the football betting market. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, Oxford University Press, v. 46, n. 2, p. 265–280, 1997. Disponível em: <http://www.jstor.org/stable/2986290>. Acesso em: 10 set. 2025. 13, 31, 32

EASTWOOD, M. **Predicting Football Results Using Python and the Dixon and Coles Model**. Reino Unido: Blog Pena.lt/y, 2021. Disponível em: <https://pena.lt/y/2021/06/24/predicting-football-results-using-python-and-dixon-and-coles/>. Acesso em: 02 abr. 2025. 35

FOOTBALL-DATA. 2025. Disponível em: <https://www.football-data.co.uk>. Acesso em: 03 nov. 2025. 35

GALLAS, D. **Por que você quase sempre vai perder dinheiro com bets, segundo a matemática**. BBC News Brasil, 2024. Disponível em: <https://www.bbc.com/portuguese/articles/c981g2n1dm9o>. Acesso em: 10 mar. 2025. 12

HERZOG, R. S. **Função de Verossimilhança**. Observatório Obstétrico Brasileiro, 2022. Disponível em: <https://observatorioobstetricobr.org/ciencia-de-dados/funcao-de-verossimilhanca/>. Acesso em: 07 out. 2025. 20

KUBRUSLY, J. **Modelos Lineares I**: Notas de aula. Universidade Federal Fluminense, 2014. 128 p. Disponível em: <https://www.professores.uff.br/jessica/wp-content/uploads/sites/137/2017/09/notas_de_aula_modelos_lineares.pdf>. Acesso em: 10 nov. 2025. 25

LIMA, K.; CUNHA, M. **Bets: 42% dos brasileiros que dizem apostar estão endividados e quase um terço está fora do mercado de trabalho, diz pesquisa do Senado**. 2024. Disponível em: <https://abrir.link/RvBbI>. Acesso em: 06 dez. 2024. 55

MARTINEZ, E. et al. **Uma introdução aos métodos bayesianos aplicados à análise de dados**. São Paulo: Cia do Ebook, 2019. 274 p. Disponível em: <https://www.google.com.br/books/edition/Uma_introdu%C3%A7%C3%A3o_aos_m%C3%A9todos_bayesianos/wOqaDwAAQBAJ?hl=pt-BR&gbpv=1>. 21

MOTA, C. V. **Como Brasil se tornou 5º maior mercado de bets no mundo**. BBC News Brasil, 2025. Disponível em: <https://www.bbc.com/portuguese/articles/cp98gn2rpyvo>. Acesso em: 07 nov. 2025. 11

NELDER, J. A.; WEDDERBURN, R. W. M. **Generalized Linear Models**. Journal of the Royal Statistical Society: Blackwell Publishing, 1972. v. 135. 370–384 p. Disponível em: <https://www.jstor.org/stable/2344614?origin=JSTOR-pdf>. 25

ODDSAGORA. **Resultados e Odds Históricas para Brasileirão Betano 2025**. OddsAgora, 2025. Disponível em: <https://www.oddsagora.com.br/football/brazil/brasileirao-betano-2025/results/>. Acesso em: 16 dez. 2025. 46

POISSON, S. D. **Recherches sur la probabilité des jugements en matières criminelles e matière civile.** Paris: Bachelier, 1837. 415 p. 13, 19

RIBBEIRO, L. **Beneficiários do Bolsa Família gastaram R$ 3 bi com "bets" em agosto, diz BC**. BBC News Brasil, 2024. Disponível em: <https://www.bbc.com/portuguese/articles/c981g2n1dm9o>. Acesso em: 06 dez. 2024. 55

SEED, S. de Estado da E. **RCO+Aulas**. Paraná: [s.n.], 2022. Disponível em: <https://professor.escoladigital.pr.gov.br/rco_mais_aulas>. Acesso em: 10 dez. 2025. 60

SHEEHAN, D. **Predicting Football Results With Statistical Modelling**. Londres: Blog dashee87.github.io, 2017. Disponível em: <https://dashee87.github.io/data%20science/football/r/predicting-football-results-with-statistical-modelling/>. Acesso em: 10 mar. 2025. 35

ZORZETTO, R.; ORLANDI, A. P. **Os efeitos nocivos dos jogos on-line**. Revista Pesquisa FAPESP, edição impressa nº 344, 2024. Disponível em: <https://revistapesquisa.fapesp.br/os-efeitos-nocivos-dos-jogos-on-line/>. Acesso em: 05 dez. 2024. 55