



Universidade Federal do Maranhão
Centro de Ciências Exatas e Tecnologia
Programa de Mestrado Profissional em Matemática em Rede
Nacional - PROFMAT



Rodrigo dos Santos Neri da Silva

Modelagem Probabilística em Apostas Esportivas com Abordagem no
Ensino Médio

São Luís - MA
2025

Universidade Federal do Maranhão
Centro de Ciências Exatas e Tecnologia
Programa de Mestrado Profissional em Matemática
em Rede Nacional - PROFMAT

Rodrigo dos Santos Neri da Silva

Modelagem Probabilística em Apostas Esportivas com Abordagem no
Ensino Médio

Dissertação apresentada ao Programa de Mestrado
Profissional em Matemática (PROFMAT) da UFMA
como requisito parcial para a obtenção do título de
Mestre em Matemática.

São Luís - MA
2025

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Diretoria Integrada de Bibliotecas/UFMA

dos Santos Neri da Silva, Rodrigo.

Modelagem Probabilística em Apostas Esportivas com
Abordagem no Ensino Médio / Rodrigo dos Santos Neri da
Silva. - 2025.

88 f.

Orientador(a): Josenildo de Souza Chaves.

Dissertação (Mestrado) - Programa de Pós-graduação em
Rede - Matemática em Rede Nacional/ccet, Universidade
Federal do Maranhão, São Luís - Ma, 2025.

1. Ensino de Probabilidade. 2. Distribuição de
Poisson. 3. Apostas Esportivas. 4. Sequência Didática.
5. Ensino Médio. I. de Souza Chaves, Josenildo. II.
Título.

Universidade Federal do Maranhão
Centro de Ciências Exatas e Tecnologia
Programa de Mestrado Profissional em Matemática
em Rede Nacional - PROFMAT

Dissertação apresentada ao Programa de Mestrado Profissional em Matemática (PROFMAT) da UFMA como requisito parcial para a obtenção do título de Mestre em Matemática.

Área de Concentração: Probabilidade e matemática Crítica.

Aprovada em:

Prof. Dr. Josenildo de Souza Chaves
Orientador

Universidade Federal do Maranhão (UFMA)

Prof. Dra. Valeska Martins de Souza
Universidade Federal do Maranhão (UFMA)

Prof. Dr. Adecarlos Costa Carvalho
Universidade Federal do Maranhão (UFMA)

À minha filha Aurora

Agradecimentos

Os dois anos do Mestrado Profissional em Matemática em Rede Nacional (PROFMAT) representaram um período repleto de aprendizado, desafios e superação. Sem o amparo de Deus e sem o apoio das pessoas e instituições que menciono a seguir, a conclusão deste trabalho estaria seriamente comprometida.

Aos meus pais, Almir e Aurene, que, apesar das dificuldades, sempre me incentivaram e encorajaram, transmitindo valiosos ensinamentos que guardarei para toda a vida.

À minha esposa Lucimara e à minha filha Aurora, que nasceu durante o primeiro semestre do curso, pela compreensão e paciência durante os períodos em que precisei estar ausente. Vocês renovam diariamente em mim o sentido de cada esforço e dedicação.

Ao meu orientador, professor Dr. Josenildo Souza Chaves, cuja dedicação, orientação e valiosos conselhos foram cruciais para a realização deste trabalho. Sua paciência e compromisso em compartilhar seu conhecimento contribuíram de forma significativa para minha formação, e sou imensamente grato por tê-lo encontrado em minha trajetória.

Aos colegas e professores do PROFMAT-UFMA, pelas experiências, pelo aprendizado compartilhado e pelo companheirismo desde o primeiro dia do mestrado.

Estendo meus agradecimentos à Universidade Federal do Maranhão (UFMA) e ao PROFMAT, que representam um divisor de águas em minha trajetória acadêmica, qualificando professores da educação básica e fortalecendo a formação profissional.

Por fim, agradeço a todos que, direta ou indiretamente, contribuíram para a realização deste trabalho, seja por meio de apoio, incentivo ou amizade.

A beleza da matemática só se mostra
aos seguidores mais pacientes

Maryam Mirzakhani

Resumo

Este trabalho discute o uso da distribuição de Poisson para prever resultados de partidas de futebol, com o objetivo de desenvolver uma sequência didática para alunos do Ensino Médio. A abordagem é justificada pelo crescimento exponencial do número de casas de apostas no Brasil, especialmente em um cenário ainda carente de legislação regulatória. O objetivo principal é familiarizar alunos do Ensino Médio, por meio da educação crítica, com o uso de ferramentas computacionais e probabilísticas empregadas por casas de apostas para gerar lucros, como os overrounds. A base da sequência didática depende de uma revisão de tópicos, incluindo conceitos estatísticos e de probabilidade que o professor deve dominar. A metodologia é exemplificada utilizando dados de campeonatos brasileiros de futebol, especificamente o Campeonato Paulista e o Campeonato Brasileiro Série A.

Palavras-chave: Ensino de Probabilidade. Distribuição de Poisson. Apostas Esportivas. Sequência Didática. Ensino Médio.

Abstract

This work discusses the use of the Poisson distribution to predict soccer match outcomes, with the purpose of developing a teaching sequence for high school students. The approach is justified by the exponential growth in the number of bookmakers in Brazil, especially in a scenario still lacking regulatory legislation. The main purpose is to familiarize high school students, through critical education, with the use of computational and probabilistic tools employed by bookmakers to generate profits, such as overrounds. The basis of the teaching sequence relies on a review of topics, including statistical and probability concepts that the teacher must master. The methodology is exemplified using data from Brazilian soccer championships, specifically the São Paulo State Championship and the Brazilian Serie A.

Key-words: Probability Teaching. Poisson Distribution. Sports Betting. Didactic Sequence. High School.

Sumário

1	INTRODUÇÃO	1
1.1	A Ascensão das Apostas Esportivas no Brasil: Regulamentação, Impactos Econômicos e Desafios Sociais	2
1.2	Probabilidade e Estatística sobre o Viés da Base Nacional Comum Curricular (BNCC)	3
1.3	Organização da Dissertação	4
2	DESCREVENDO OS DADOS	6
2.1	Classificação das Variáveis	7
2.2	Distribuição de Frequências de uma Variável	7
2.3	Descrição Gráfica de uma Variável	11
2.3.1	Gráficos de Barras e Gráficos de Setores (Pizza)	13
2.3.2	Gráficos de Dispersão	14
2.3.3	Histogramas	15
2.4	Descrição Numérica de uma Variável	17
2.4.1	Média Aritmética	17
2.4.2	Mediana	18
2.4.3	Moda	18
2.4.4	Medidas de Dispersão	19
2.4.5	Quartis	21
2.4.6	<i>Boxplot</i>	22
3	PROBABILIDADE E VARIÁVEIS ALEATÓRIAS	23
3.1	Fundamentos e Propriedades da Probabilidade	23
3.2	Propriedades Fundamentais de Probabilidade	25
3.3	Probabilidade Condicional	26
3.4	Variáveis aleatórias	28
3.5	Função de Probabilidade de uma Variável Aleatória X	29
3.6	Função de Distribuição Acumulada	30
3.7	Valor Esperado de uma Variável Aleatória	31
3.8	Variância de uma Variável Aleatória Discreta	32
3.9	Desvio Padrão de uma Variável Aleatória	33
3.10	Covariância entre duas Variáveis Aleatórias	33
3.11	Modelos Probabilísticos de Variáveis Aleatórias Discretas	33
3.11.1	Distribuição de Bernoulli	34
3.11.2	Distribuição Binomial	35
3.11.3	Binomial Negativa	35
3.11.4	Distribuição Hipergeométrica	36
3.11.5	Distribuição de Poisson	36
3.11.6	Distribuição Conjunta de Duas Variáveis Aleatórias	38

3.11.7	Distribuição de Poisson Bivariada	40
4	PROBABILIDADE NAS APOSTAS ESPORTIVAS	41
4.1	Odds	41
4.1.1	Odds Decimais	41
4.1.2	Odds Fracionárias	41
4.1.3	Odds Americana	42
4.2	Lucro das Casas de Apostas	42
4.3	Cálculo de Odds Usando a Distribuição de Poisson	43
4.3.1	Campeonato Paulista 2025	44
4.3.2	Brasileirão Série A	48
5	ENSINO DE PROBABILIDADE COM FOCO EM APOSTAS ESPORTIVAS	50
5.1	GeoGebra: Cálculo da Distribuição de Poisson	50
5.1.1	Passo 1: Acesso à Calculadora de Probabilidade	51
5.1.2	Passo 2: Seleção da Distribuição de Poisson	51
5.1.3	Passo 3: Inserção de Parâmetros e Cálculo de Probabilidades	52
5.2	Público-alvo	53
5.3	Objetivo	54
5.4	Habilidades e Competências	54
5.5	Roteiro para a Sequência Didática	55
5.6	Planejamento das Aulas	55
5.6.1	Aula 1 – Probabilidade no Universo das Apostas Esportivas	55
5.6.2	Aula 2 – Como as Casas de Aposta Transformam <i>Odds</i> em Lucro?	58
5.6.3	Aula 3 – O Modelo Probabilístico de Poisson a Partir do Modelo Binomial	60
5.6.4	Aula 4 – Uso da Calculadora de Probabilidades do GeoGebra (Distribuição de Poisson)	62
5.6.5	Aula 5 – Aplicação, Análise Crítica e Reflexões	63
6	CONSIDERAÇÕES FINAIS	66
	Referências	67
A	Códigos em R da Análise de Dados do Campeonato Paulista de Futebol 2025	70
B	Códigos em R da Análise de Dados do Campeonato Brasileiro de Futebol – Série A 2025	73
C	Estimação de Parâmetros	77
C.1	Introdução	77
C.2	Estimador de Máxima Verossimilhança	77
C.3	Estimando o Parâmetro λ de uma Poisson	78

INTRODUÇÃO

As *bets* (termo em inglês para apostas) têm ganhado notoriedade, ultrapassando o limite entre entretenimento e temas como saúde pública, endividamento e tributação. Impulsionadas por propagandas — em partidas de futebol ou divulgadas por influenciadores digitais — e pelo fácil acesso a smartphones e à internet, essa modalidade tem atraído principalmente jovens. Segundo Póvoa et al. (2023), a ausência de barreiras decisórias, chamadas *decision points*, tende a favorecer comportamentos irracionais. A Figura 1.1 apresenta o número de empresas de apostas que operavam no Brasil, com sede no exterior, entre 2021 e 2024.

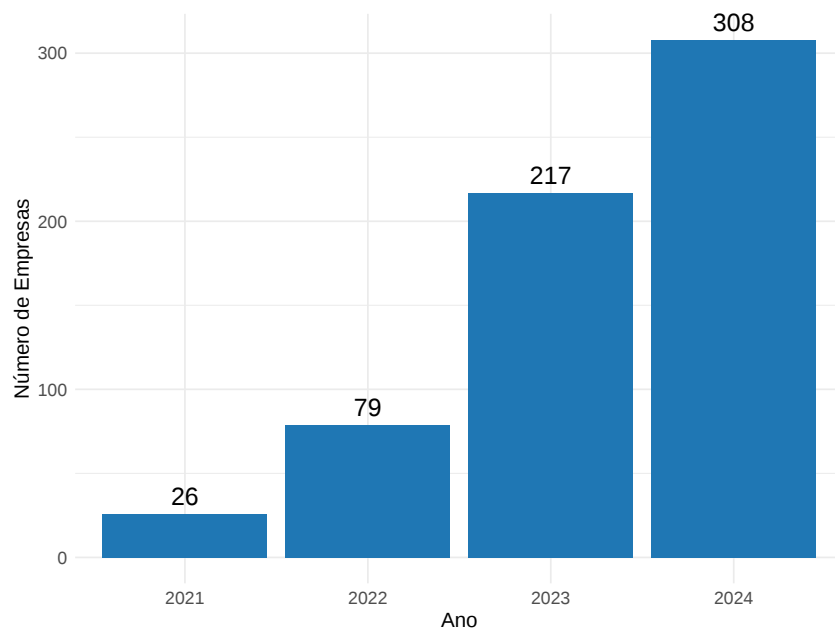


Figura 1.1. Empresas de apostas registradas no Brasil, 2021–2024

Fonte: CNN Brasil (2024)

Observe que o número de casas de apostas aumentou cerca de $\left(\frac{308-26}{26}\right) \times 100 \approx 1\,085\%$ nesse período.

1.1 A Ascensão das Apostas Esportivas no Brasil: Regulamentação, Impactos Econômicos e Desafios Sociais

Embora não seja uma temática inteiramente nova, as apostas esportivas receberam enfoque legal apenas nos últimos anos. Em 2018, o então governo Temer abordou a modalidade lotérica denominada “apostas de quota fixa”, por meio da Lei nº 13.756, de 12 de dezembro de 2018 (BRASIL, 2018b). Essa legislação legalizou as apostas esportivas de caráter prognóstico, definidas no inciso IV do Art. 14 como uma modalidade de loteria na qual o apostador busca prever os resultados de eventos esportivos (BRASIL, 2018b).

Apesar de a Lei nº 13.756/2018 autorizar as apostas esportivas, elas ainda são realizadas predominantemente em sites de empresas situadas no exterior. Não há empresas físicas no Brasil, pois o processo regulatório é recente e teve início somente em 2023, com a promulgação da Lei nº 14.790/2023 (BRASIL, 2023). A partir disso, o Ministério da Fazenda criou, em 2024, a Secretaria de Prêmios e Apostas do Ministério da Fazenda (SPA-MF), com o objetivo de regulamentar e cadastrar as empresas de apostas esportivas no país.

No Brasil, o termo *bets* é utilizado de forma genérica para se referir a apostas online, incluindo apostas esportivas e jogos de cassino. Cassinos físicos são proibidos no país desde 1946, com exceção de bingos e jogos de habilidade em situações específicas (como concursos culturais ou eventos filantrópicos). Apesar dessa proibição, os cassinos online têm ganhado popularidade e operam legalmente quando sediados no exterior, aproveitando a ausência de regulamentação nacional específica. Já as apostas esportivas (*bets* esportivas) enquadram-se como uma categoria de jogos de azar, mas com foco em resultados esportivos. Com a aprovação da Lei nº 13.756/2018, as apostas de quota fixa passaram a ser regulamentadas no Brasil, embora a regulamentação prática ainda esteja em fase de implementação (BRASIL, 2018b).

Atualmente, o apostador pode consultar o Sistema de Gestão de Apostas – SIGAP, que, em seu sítio eletrônico, afirma:

O Sistema de Gestão de Apostas – SIGAP é uma solução tecnológica desenvolvida pelo Ministério da Fazenda com o objetivo de permitir a regulação, o monitoramento e a fiscalização do mercado de apostas no Brasil, a partir do disposto na Lei nº 13.756/2018 e na Lei nº 14.790/2023 (FAZENDA, 2024).

Com a regulamentação das apostas esportivas, espera-se a arrecadação de uma quantidade significativa de tributos, além da geração de novos postos de trabalho. Entretanto, essa regulamentação deve ser minuciosa para evitar o uso indevido do setor em práticas ilícitas, como lavagem de dinheiro, manipulação de resultados e fraudes. Além disso, a população estará legalmente amparada em situações de bloqueio de contas, recusa de pagamentos, dificuldade de contato com as empresas, entre outras. Para além das questões econômicas, é necessário observar também os impactos sociais dessas práticas.

Alguns estudos procuraram quantificar os custos do jogo em comparação com os benefícios — estes medidos em termos de receitas tributárias — e concluíram que os custos são significativamente maiores, inclusive porque, cada vez mais, têm sido necessárias políticas públicas sistemáticas para mitigar danos onde prevalece o jogo (PÓVOA et al., 2023).

A dimensão social, diferentemente da econômica, apresenta impactos majoritariamente negativos, uma vez que as apostas esportivas estão frequentemente associadas a problemas

como endividamento excessivo, comportamentos compulsivos, queda no rendimento profissional e conflitos familiares. Dessa forma, as apostas esportivas podem causar impactos significativos, surgindo, portanto, a necessidade de implementação de políticas públicas para seu controle e regulamentação.

Segundo Browne et al. (2017), em 2014–2015, o total de gastos com apostas em Victoria, estado da Austrália, alcançou AUD 5,8 bilhões, dos quais AUD 1,6 bilhão foi direcionado ao Estado como receita tributária. No entanto, os autores estimam que os custos sociais decorrentes do jogo — englobando saúde, produtividade, criminalidade e apoio social — chegaram a aproximadamente AUD 7 bilhões, gerando um impacto fiscal negativo de cerca de AUD 5,4 bilhões naquele ano. Esses efeitos são intensificados por campanhas publicitárias que utilizam inteligência artificial (IA) e atingem, com maior intensidade, a população mais vulnerável, especialmente crianças e adolescentes.

Segundo Póvoa et al. (2023), em 2022, entre os vinte clubes que integravam a Série A do Campeonato Brasileiro de Futebol, apenas o Palmeiras não contava com o patrocínio de alguma casa de apostas esportivas. Assim, além da proteção aos consumidores, a regulamentação deve considerar também os impactos sociais e de saúde pública decorrentes da expansão dessas práticas.

1.2 Probabilidade e Estatística sobre o Viés da Base Nacional Comum Curricular (BNCC)

A Constituição Federal de 1988, em seu artigo 210, trata dos conteúdos mínimos para a educação: "Serão fixados conteúdos mínimos para o ensino fundamental, de maneira a assegurar formação básica comum e respeito aos valores culturais e artísticos, nacionais e regionais." (BRASIL, 1988, Art. 210). Este artigo da constituição fundamenta a criação de uma diretriz curricular nacional, a qual culminou na criação da Base Nacional Comum Curricular (BNCC). Ao encontro da Constituição Federal, a Lei nº 9.394/96, Lei de Diretrizes e Bases da Educação Nacional (LDB), é a principal norma que detalha o funcionamento da educação infantil ao ensino superior. Em seu artigo 26, a LDB diz que:

Segundo a Lei de Diretrizes e Bases da Educação Nacional, os currículos da educação infantil, do ensino fundamental e do ensino médio devem ter uma base nacional comum, a ser complementada, em cada sistema de ensino e em cada estabelecimento escolar, por uma parte diversificada, exigida pelas características regionais e locais da sociedade, da cultura, da economia e dos educandos (BRASIL, 1996, Art. 26).

Dessa forma a LDB reforça a necessidade da fixação de conteúdo mínimos, de tal forma que em todo território nacional esses conteúdos sejam ensinados; garantindo uma equidade educacional em âmbito nacional. assim surge a BNCC, que em seu site é definida como:

Conforme definido na Lei de Diretrizes e Bases da Educação Nacional (LDB, Lei n.º 9.394/1996), a Base deve nortear os currículos dos sistemas e redes de ensino das Unidades Federativas, bem como as propostas pedagógicas de todas as escolas públicas e privadas de Educação Infantil, Ensino Fundamental e Ensino Médio, em todo o Brasil.

A Base estabelece conhecimentos, competências e habilidades que se espera que todos os estudantes desenvolvam ao longo da escolaridade básica. Orientada pelos princípios éticos, políticos e estéticos traçados

pelas Diretrizes Curriculares Nacionais da Educação Básica, a Base soma-se aos propósitos que direcionam a educação brasileira para a formação humana integral e para a construção de uma sociedade justa, democrática e inclusiva. (BRASIL, 2018)

Observa-se, portanto, que a BNCC não se limita a apresentar uma lista de conteúdos, mas define competências e habilidades essenciais para o desenvolvimento integral dos estudantes. No âmbito da Matemática, a probabilidade e a estatística ganham destaque como ferramentas fundamentais para compreender e analisar fenômenos do cotidiano, favorecendo a formação de um pensamento crítico e a tomada de decisões baseadas em dados. Essa perspectiva vai ao encontro do objetivo principal deste trabalho, que é a exploração e aplicação de ferramentas probabilísticas e estatísticas em contextos reais, como as apostas esportivas, aproximando o ensino da realidade vivenciada pelos alunos. O trabalho foi desenvolvido seguindo os seguintes objetivos específicos:

- Analisar a abordagem da BNCC (Base Nacional Comum Curricular) e PCN (Parâmetros Curriculares Nacionais) para o ensino de probabilidade;
- Propor modelos básicos de probabilidade para estimar *odds* em apostas de campeonatos de futebol;
- Apresentação de uma sequência didática para o ensino de probabilidade com ênfase em apostas esportivas.

1.3 Organização da Dissertação

Este trabalho está estruturado em seis capítulos e três apêndices. O Capítulo 1 introduz o tema, abordando a ascensão das apostas esportivas no Brasil — regulamentação, impactos econômicos e desafios sociais — e apresenta a importância, para os estudantes da educação básica, dos fundamentos de probabilidade e estatística sob o viés da Base Nacional Comum Curricular (BNCC).

No Capítulo 2 são tratados os conceitos fundamentais de estatística descritiva. São apresentados os tipos de variáveis, as distribuições de frequências e as principais formas de representação gráfica, como gráficos de barras, setores, dispersão e histogramas, além de medidas descritivas.

O Capítulo 3 trata de uma introdução à teoria das probabilidades, variáveis aleatórias e principais distribuições de probabilidade discretas e distribuições conjuntas de variáveis aleatórias.

O Capítulo 4 mostra como a distribuição de Poisson pode ser aplicada na predição de placares em partidas de futebol, tomando por base as estatísticas dos times. O Capítulo 5 apresenta a sequência didática voltada ao Ensino Médio para aplicação do modelo de Poisson na predição de placares. Essa proposta é contextualizada pelo universo das apostas esportivas e da análise de *odds*, incluindo o roteiro de uso do GeoGebra, a definição do público-alvo, os objetivos, as habilidades e competências previstas pela BNCC e o planejamento detalhado das aulas.

Por fim, o Capítulo 6 reúne as considerações finais, apresentando as principais conclusões do trabalho e sugestões para pesquisas futuras.

Complementando o texto principal, o Apêndice A.1 apresenta a análise do Campeonato Paulista de 2025; o Apêndice B.1 traz a análise do Brasileirão Série A de 2025 como estudo de caso, ambas realizadas no software R. O Apêndice C reúne fundamentos de

inferência estatística, com ênfase na estimação de parâmetros e no método da máxima verossimilhança, incluindo a dedução do estimador para o parâmetro λ da distribuição de Poisson.

DESCREVENDO OS DADOS

Suponhamos que um torcedor, interessado em prognosticar o desempenho de seu time em uma partida próxima, siga os seguintes passos:

- Selecionou partidas anteriores em que o atual elenco estava presente. Levantou alguns dados, como gols e vitórias em casa e fora de casa, gols por jogador, número de gols por partida, número de cartões, etc.
- Organizou os dados obtidos em tabelas, construiu gráficos e também determinou as medidas de tendência central, bem como a variabilidade dos dados.
- Por fim, confirmou se o que havia previsto realmente ocorreu e verificou a distância entre o que ocorreu e o que foi previsto.

Na primeira etapa, foi coletada uma **Amostra** das partidas anteriores. Na segunda etapa, realizou-se uma descrição dos dados, caracterizando a **Estatística Descritiva**. Por fim, na terceira etapa, foi realizada uma **Inferência Estatística**.

A inferência estatística é uma das partes da Estatística. Esta é a parte da metodologia da Ciência que tem por objetivo a coleta, redução, análise e modelagem dos dados, a partir do que, finalmente, faz-se a inferência para uma população da qual os dados (a amostra) foram obtidos. Um aspecto importante da modelagem dos dados é fazer previsões, a partir das quais se podem tomar decisões (MORETTIN; BUSSAB, 2010, p. 1).

Assim, segundo Morettin e Bussab (2010) partindo de um conjunto de dados relacionados a partidas anteriores, podemos cogitar o comportamento de eventos futuros. É importante salientar que, a inferência se dá a partir de um subconjunto (amostra) do universo observado (população) e que, tratando-se de eventos esportivos, para encontrar a probabilidade de um evento, não nos interessa todas as partidas já jogadas, mas em especial as últimas em que o atual elenco (atletas), estava presente. Enquanto a inferência estatística trabalha com dados coletados e tira conclusões a partir de uma amostra da população, a **Probabilidade** lida com a incerteza da ocorrência de eventos futuros. Dessa forma, a probabilidade, aplicada às apostas esportivas, permite calcular as chances de um determinado resultado em um evento esportivo (como empate, número de gols, vitória ou derrota).

Este capítulo tem por objetivo apresentar conceitos relacionados à descrição numérica e gráfica de um conjunto de dados. A análise desses dados dependerá do tipo de variável. A coleta de dados pode ocorrer de duas formas:

- coleta de todos os elementos da população (censo);

- coleta de dados de apenas uma parte da população (amostra).

Já as medidas que descrevem as propriedades de uma amostra e de uma população recebem, respectivamente, as denominações de *estatística* e *parâmetro*. Em termos práticos, quando analisamos numericamente todas as partidas de um campeonato — por exemplo, calculando a média de gols — estamos diante de um parâmetro. Quando essa mesma descrição numérica se baseia apenas em algumas partidas, o resultado obtido é classificado como estatística.

2.1 Classificação das Variáveis

A coleta e a análise dependem do tipo de variável. Em termos gerais, classificadas como:

- **variáveis qualitativas:** apresentam como resposta um atributo (por exemplo, a posição de um jogador: atacante, meio-campista, defensor, goleiro; ou o tipo de cartão: amarelo, vermelho); podem ser subclassificadas em *nominais*, quando não há ordem entre as categorias, e *ordinais*, quando há uma ordem estabelecida (por exemplo, a classificação em uma liga: 1^o, 2^o, 3^o);
- **variáveis quantitativas:** são obtidas por mensuração ou contagem e, portanto, representadas por números; podem ser *contínuas*, quando podem assumir qualquer valor em um intervalo dos números reais (por exemplo, a altura de um jogador), ou *discretas*, quando assumem valores de um conjunto finito ou enumerável, típicos de contagens (por exemplo, número de faltas, cartões amarelos, escanteios).

De forma geral, as variáveis podem ser classificadas de acordo com a Figura 2.1.

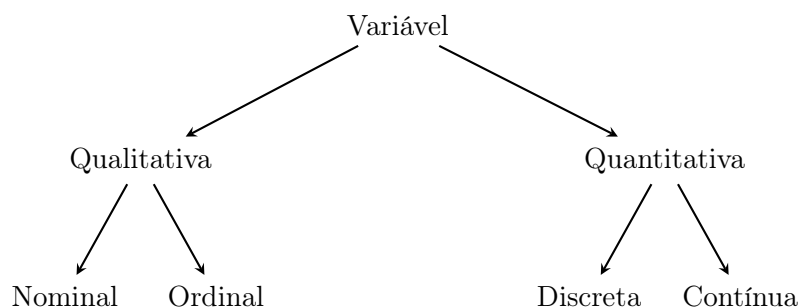


Figura 2.1. Classificação das variáveis estatísticas.

Fonte: Elaborado pelo autor.

2.2 Distribuição de Frequências de uma Variável

Quando se coletam dados, uma ferramenta estatística de grande valor é a distribuição de frequências, a qual nos diz o comportamento da variável. Essa distribuição permite organizar e resumir grandes volumes de dados, além de facilitar a identificação de padrões e tendências nas variáveis estudadas.

A Tabela 2.1 apresenta o placar das últimas 380 partidas do Campeonato Brasileiro Série A. Os dados foram coletados do site Windrawwin (2025).

Tabela 2.1. Brasileirão Série A: Placares Mais Comuns

Placar Final	Número de Jogos	Frequência Relativa (%)
1-0	56	14,74
2-1	43	11,32
0-0	39	10,26
2-0	35	9,21
1-1	33	8,68
3-0	28	7,37
0-1	27	7,11
1-2	21	5,53
0-2	16	4,21
2-2	16	4,21
3-1	15	3,95
1-3	11	2,89
4-1	6	1,58
3-2	5	1,32
0-3	5	1,32
1-4	4	1,05
2-3	4	1,05
5-1	3	0,79
3-3	3	0,79
4-0	2	0,53
4-2	2	0,53
4-3	1	0,26
5-0	1	0,26
0-4	1	0,26
5-2	1	0,26
6-0	1	0,26
6-1	1	0,26
Total	380	100,00

Fonte: Windrawwin (2025). Adaptado pelo autor.

Na primeira coluna da Tabela 2.1, são apresentados os placares registrados. Na segunda coluna, encontra-se a frequência absoluta. Em um conjunto de dados, a frequência absoluta representa o número de vezes que esse valor ou atributo aparece na amostra. Em outras palavras, é a contagem bruta de ocorrências. Já na última coluna, é exibida a frequência relativa. Em um conjunto de dados, a frequência relativa representa a proporção de vezes que um determinado valor ou atributo aparece na amostra, em relação ao total de observações. Em outras palavras, é a frequência absoluta dividida pelo tamanho total da amostra, geralmente expressa como uma porcentagem ou fração, que cumpre a importante função de permitir a comparação dessa pesquisa (Placares do Brasileirão Série A) com outras pesquisas com um tamanho de amostra diferente.

Vamos considerar a tabela abaixo, que fornece os placares da Copa do Mundo de 2022.

Tabela 2.2. Placares da Copa do Mundo FIFA 2022

Placar Final	Número de Jogos	Frequência Relativa (%)
2-1	11	17,19
1-0	10	15,63
2-0	9	14,06
0-0	8	12,50
1-1	8	12,50
0-1	6	9,38
1-2	6	9,38
0-2	5	7,81
3-1	5	7,81
3-0	4	6,25
2-2	3	4,69
3-2	3	4,69
4-1	3	4,69
0-3	2	3,13
1-3	2	3,13
2-3	2	3,13
3-3	2	3,13
0-4	1	1,56
1-4	1	1,56
2-4	1	1,56
4-0	1	1,56
4-2	1	1,56
4-3	1	1,56
5-0	1	1,56
5-1	1	1,56
5-2	1	1,56
6-0	1	1,56
6-1	1	1,56
Total	100	156,26

Fonte: FIFA (2022). Adaptado pelo autor.

Observe que o placar 1-0 teve uma frequência absoluta de 56 no Brasileirão Série A, enquanto na Copa do Mundo de 2022 essa mesma frequência foi de 10. Dessa forma, uma comparação direta não é adequada, pois as frequências absolutas são fortemente influenciadas pelo tamanho da amostra. No entanto, ao analisar as frequências relativas, torna-se possível comparar a distribuição dos resultados entre diferentes competições, já que essas medidas são padronizadas para um total de 100%.

No Brasileirão Série A, o placar 1-0 representou 14,74% das partidas, sendo o mais comum. Já na Copa do Mundo de 2022, esse placar correspondeu a 15,63%, ocupando a segunda posição entre os mais frequentes.

As tabelas 2.1 e 2.2 referem-se a uma variável quantitativa discreta, pois representam a contagem de gols, que assume apenas valores inteiros. Para variáveis contínuas, como a posse de bola média das equipes, a construção de tabelas de frequência exige um cuidado maior. A posse de bola é uma importante descrição do modo de jogo do time. O domínio

da posse de bola indica que o time está em vantagem, tendo mais oportunidades para atacar e criar jogadas, bem como evitar que o adversário desenvolva seu jogo. A Tabela 2.3 apresenta a posse de bola média dos 20 times que participaram do Brasileirão Série A 2022.

Tabela 2.3. Posse de bola dos times no Campeonato Brasileiro Série A 2022

Equipe	Posse de bola (%)
América-MG	43,1
Coritiba	44,3
Goiás	45,5
Cuiabá	46,8
Vasco	47,4
Cruzeiro	47,9
Santos	48,7
Athletico-PR	49,3
Bragantino	49,8
Fortaleza	50,5
Corinthians	50,9
Grêmio	51,5
Internacional	52,8
Palmeiras	53,7
Botafogo	53,9
Fluminense	55,6
Flamengo	56,0
Bahia	56,1
Atlético-MG	57,4
São Paulo	57,7

Fonte: Marçal et al. (2022). Adaptado pelo autor.

Como cada time obteve diferentes valores de posse de bola, e essa é uma variável contínua, teremos uma tabela com 20 linhas para a análise desses dados. Uma maneira interessante de visualizar melhor o comportamento dessa variável é agrupá-la em classes. No entanto, essas classes não devem ser muito curtas, pois isso não contribuiria para uma melhor visualização; por outro lado, classes muito longas resultariam na perda de informações relevantes.

Dessa forma, considerando que os dados variam entre os valores inteiros 43 e 58%, utilizamos 5 classe e cada classe terá intervalo 3. De modo geral, a quantidade e tamanho das classes quem determina é o analista, atendo-se a uma melhor visualização das informações. O símbolo $a \vdash b$ representa o intervalo real $[a, b)$.

Tabela 2.4. Posse de Bola Média no Campeonato Brasileiro Série A 2022

Posse de Bola (%)	Freq. Absoluta	Freq. Relativa (%)	Equipes
43 ┤ 46	6	30,0	América-MG Coritiba Goiás Cuiabá Vasco Cruzeiro
46 ┤ 49	5	25,0	Santos Athletico-PR Bragantino Fortaleza Corinthians
49 ┤ 52	4	20,0	Grêmio Internacional Palmeiras Botafogo
52 ┤ 55	3	15,0	Fluminense Flamengo Bahia
55 ┤ 58	2	10,0	Atlético-MG São Paulo
Total	20	100,0	

Fonte: Marçal et al. (2022). Adaptado pelo autor.

Observe que não conhecemos as duas posses de bola que estão no intervalo 55 ┤ 58. Em alguns casos, dizemos que a posse de bola desse intervalo é seu valor médio, 56,5.

2.3 Descrição Gráfica de uma Variável

Os gráficos despenham um papel fundamental, pois facilitam a interpretação e comportamento dos dados.

Métodos gráficos têm encontrado um uso cada vez maior devido ao seu forte apelo visual. Normalmente, é mais fácil para qualquer pessoa entender a mensagem de um gráfico do que aquela embutida em tabelas ou sumários numéricos. (MORETTIN; BUSSAB, 2010) (pág. 3)

Há vários tipos de gráficos. Para variáveis qualitativas, os mais comuns são os gráficos de barras e de setores (pizza), embora os gráficos de barras sejam geralmente mais recomendados para facilitar comparações. Para variáveis quantitativas, podemos citar diagramas de ramos e folhas, gráficos de dispersão, histogramas e boxplots. Os histogramas são indicados para variáveis quantitativas contínuas, enquanto os diagramas de ramos e folhas são úteis para pequenas amostras. Os boxplots são ferramentas eficazes para resumir a distribuição dos dados e serão tratados adiante na seção Descrição Numérica de uma Variável.

Consideremos a Tabela 2.5 do Campeonato Brasileiro Série B. Nela, estão presentes as informações de desempenho dos 20.

Tabela 2.5. Classificação de times

Time	P	J	V	E	D	GP	GC	SG	%
Santos	68	36	20	8	8	56	28	28	0,62
Mirassol	64	37	18	10	9	41	26	15	0,57
Novorizontino	64	37	18	10	9	43	30	13	0,57
Sport	63	37	18	9	10	54	36	18	0,56
Ceará	60	36	18	6	12	58	41	17	0,55
Goiás	60	37	17	9	11	55	32	23	0,54
Operário-PR	57	37	16	9	12	33	31	2	0,51
Vila Nova	55	37	16	7	14	41	52	-11	0,49
América-MG	55	36	14	13	9	47	34	13	0,50
Coritiba	50	36	14	8	14	39	39	0	0,46
Amazonas	49	37	13	10	14	30	37	-7	0,44
Avaí	47	36	12	11	13	29	30	-1	0,43
Paysandu	47	37	11	14	12	39	42	-3	0,42
Botafogo-SP	42	36	10	12	14	32	47	-15	0,38
Chapecoense	41	36	10	11	15	32	43	-11	0,37
CRB	39	36	10	9	17	35	44	-9	0,36
Ponte Preta	38	37	10	8	19	36	52	-16	0,34
Ituano	37	37	11	4	22	43	62	-19	0,33
Brusque	33	36	7	12	17	22	40	-18	0,30
Guarani	32	36	8	8	20	32	51	-19	0,29

Fonte: Confederação Brasileira de Futebol (2024). Adaptado pelo autor.

A seguir, descrevem-se as variáveis utilizadas:

- **P (Pontos)**: Total de pontos conquistados pelo time ao longo da competição, calculado considerando três pontos por vitória, um ponto por empate e nenhum ponto por derrota.
- **J (Jogos)**: Número total de partidas disputadas pela equipe no campeonato até o momento da coleta dos dados.
- **V (Vitórias)**: Quantidade de partidas em que o time saiu vencedor.
- **E (Empates)**: Número de jogos que terminaram sem um vencedor.
- **D (Derrotas)**: Total de partidas em que a equipe foi derrotada.
- **GP (Gols Pró)**: Total de gols marcados pelo time ao longo do campeonato.
- **GC (Gols Contra)**: Total de gols sofridos pela equipe.
- **SG (Saldo de Gols)**: Diferença entre gols marcados (GP) e gols sofridos (GC), dada pela expressão:

$$SG = GP - GC$$

O saldo de gols é um critério de desempate no caso de times com a mesma pontuação.

- **% (Aproveitamento)**: Representa a proporção de pontos conquistados pelo time em relação ao total de pontos possíveis, calculada por:

$$\text{Aproveitamento} = \frac{P}{J \times 3}$$

Esse índice é expresso em forma decimal, podendo ser convertido para percentual multiplicando por 100.

Essas variáveis têm grande importância para a análise do desempenho das equipes ao longo da competição, permitindo comparações diretas entre os clubes.

2.3.1 Gráficos de Barras e Gráficos de Setores (Pizza)

Os gráficos de barra e setores dão maior clareza as distribuições de frequências, há outro gráficos comumente usados para variáveis qualitativas, porém são variações de um mesmo princípio. Observe a Figura 2.2, que representa o desempenho do goiás no brasileiro série B 2024.

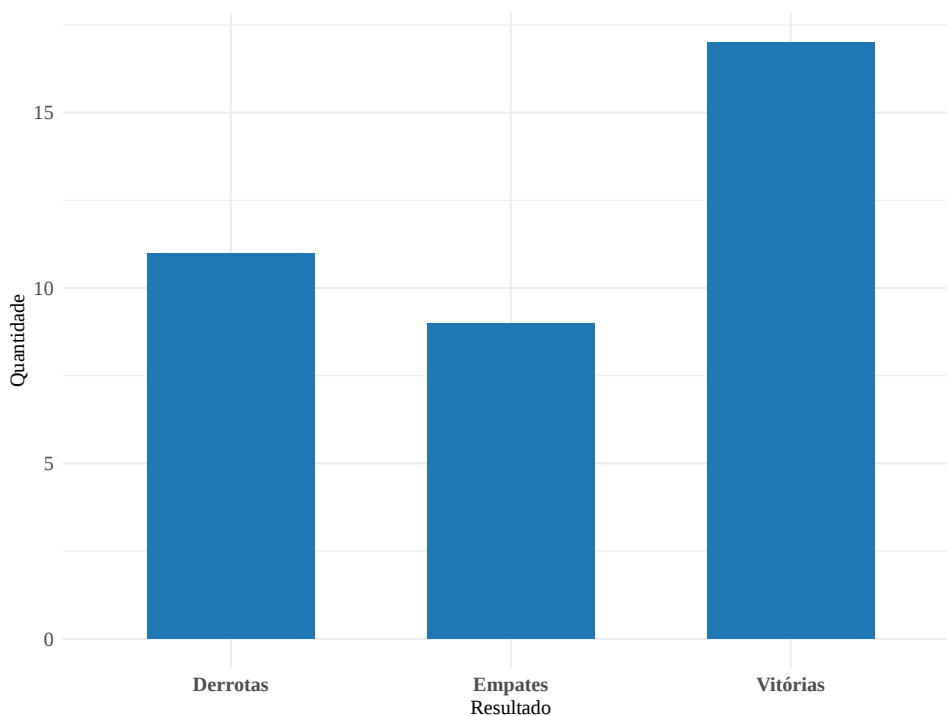


Figura 2.2. Desempenho do Goiás no Brasileiro Série B 2024.

Fonte: Tabela 2.5

Em geral, os gráficos de barras são usualmente empregados para frequências absolutas. Note que a variável desempenho, que assume as categorias derrota, empate é uma variável qualitativa ordinal, já que existe um ranking natural:

- Derrota (pior resultado)
- Empate (Intermediario)
- Vitoria (Melhor resultado)

As variáveis qualitativas nominais possuem categorias que podem ser apresentadas em qualquer ordem. Contudo, em alguns casos, é recomendável organizá-las por ordem decrescente de frequência para facilitar a visualização dos resultados.

O gráfico de setores, popularmente conhecido como gráfico de pizza, é especialmente indicado para representar frequências relativas. Nesse gráfico, cada categoria ocupa um setor circular proporcional à sua participação no total, o que facilita comparações visuais rápidas. Como as frequências relativas totalizam sempre 100%, a interpretação é imediata, pois cada setor corresponde à parcela representada por cada categoria.

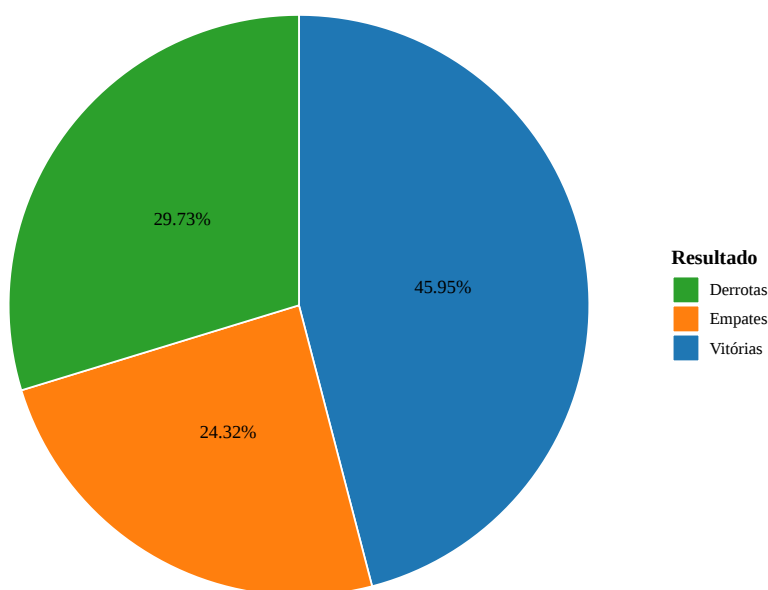


Figura 2.3. Desempenho do Goiás no Brasileirão Série B 2024.

Fonte: Tabela 2.5

2.3.2 Gráficos de Dispersão

O gráfico de dispersão é uma representação gráfica que permite visualizar a relação entre variáveis numéricas. Dependendo da quantidade de variáveis envolvidas, ele pode ser classificado como unidimensional ou bidimensional. No caso do gráfico 2.4, são relacionadas duas variáveis: o saldo de gols, representado no eixo x , e a pontuação, no eixo y , tendo como base os 10 times com maior pontuação da tabela 2.5. Trata-se, portanto, de um gráfico de dispersão bidimensional, utilizado para identificar padrões, tendências ou correlações entre essas duas variáveis.

Observa-se que o Santos lidera tanto em pontos quanto em saldo de gols, ocupando o canto superior direito do gráfico. Esse posicionamento indica consistência e domínio ao longo da competição, com bom desempenho ofensivo e uma defesa sólida.

Outro ponto a ser destacado é que o América-MG e o Novorizontino apresentam o mesmo saldo de gols (13); no entanto, há uma diferença significativa na pontuação total alcançada por essas equipes: 55 pontos para o América-MG e 64 pontos para o Novorizontino.

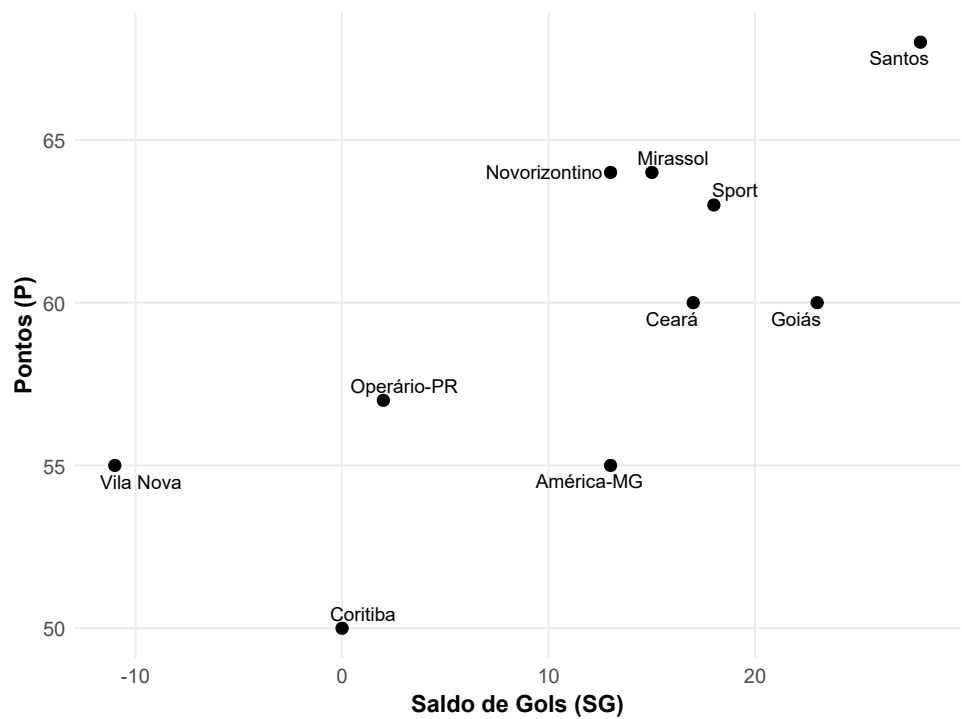


Figura 2.4. Gráfico de Dispersão entre o Saldo de Gols e os Pontos da Série B – 2024

Fonte: Tabela 2.5

2.3.3 Histogramas

Para visualizar a distribuição de uma variável quantitativa, é comum elaborar um *histograma*. O procedimento para construção de um histograma geralmente segue os passos definidos a seguir:

- a) **Definir o número de classes.** Neste exemplo, optou-se por seis classes;
- b) **Determinar a amplitude de cada classe.** Fixamos $\Delta = 0,17$ para todos os intervalos, assegurando bases iguais;
- c) **Estabelecer os limites inferiores e superiores** de cada classe e registrar, em seguida, quantas observações caem em cada uma;
- d) **Construir retângulos** sobre o eixo horizontal: a base coincide com a amplitude da classe e a altura é proporcional à frequência absoluta (ou relativa) das observações contidas no respectivo intervalo.

A Tabela 2.6, derivada da Tabela 2.5, resume esses passos para a variável *Aproveitamento (%)* em seis classes de amplitude 0,17.

Tabela 2.6. Distribuição de frequência da variável Aproveitamento (%)

Aproveitamento (%)	Frequência Absoluta	Frequência Relativa (%)
0,29 ┤ 0,35	4	20,0
0,35 ┤ 0,41	3	15,0
0,41 ┤ 0,47	4	20,0
0,47 ┤ 0,53	3	15,0
0,53 ┤ 0,59	5	25,0
0,59 ┤ 0,65	1	5,0
Total	20	100,0

Fonte: Tabela 2.5. Adaptado pelo autor.

Observa-se que o intervalo com maior frequência absoluta (5) corresponde ao aproveitamento entre 53 % e 59 %, representando 25 % do total. Não há equipes com aproveitamento entre 59 % e 65 %, justificando a frequência mínima nessa faixa.

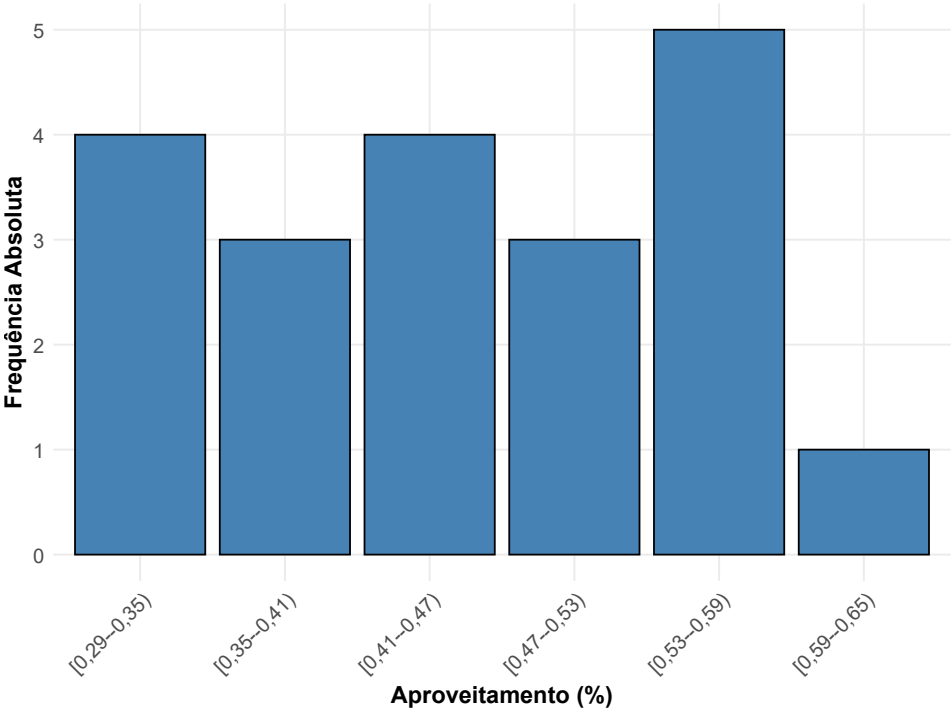


Figura 2.5. Histograma da Distribuição de Frequências do Aproveitamento (%) Gráfico de Colunas

Fonte: Tabela 2.6

Como todas as classes possuem a mesma amplitude, as bases dos retângulos são iguais e as alturas refletem, diretamente, as frequências. Essa homogeneidade facilita a identificação visual de faixas mais ou menos densas em observações. Comparado a um gráfico de dispersão, o histograma destaca de imediato onde os dados se concentram.

Embora o formato em colunas seja o mais divulgado, pode-se converter o topo dos retângulos em segmentos lineares, obtendo-se um *polígono de frequência*, ilustrado na Figura 2.6.

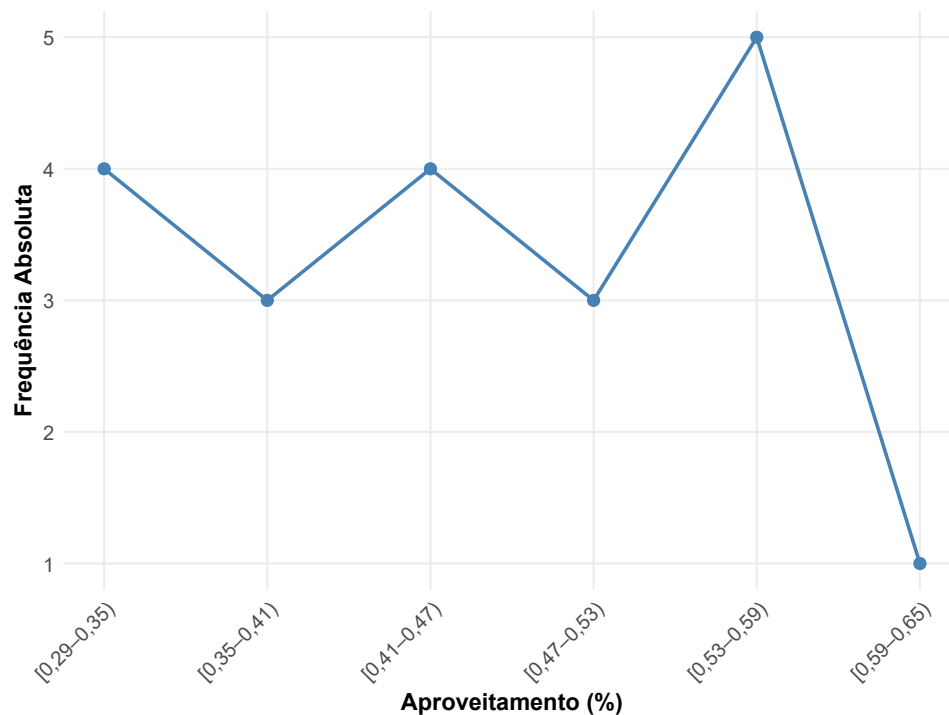


Figura 2.6. Distribuição de Frequências do Aproveitamento (%) Polígono de Frequência

Fonte: Tabela 2.6

2.4 Descrição Numérica de uma Variável

Na seção anterior, conhecemos as tabelas de frequência relativa, que, por meio da coleta e organização dos dados brutos, resumem e nos informam sobre o comportamento de uma variável. Agora, vamos conhecer as medidas de resumo numéricas, que, de forma geral, sintetizam características como tendência central, dispersão e distribuição dos dados.

2.4.1 Média Aritmética

A média aritmética é a mais popular e difundida das médias de tendência central. Se, ao analisar determinado fenômeno, obtivermos as observações $x_1, x_2, x_3, \dots, x_n$, a média aritmética, $\bar{x}$, desse conjunto de valores é dada por:

$$\bar{x} = \frac{x_1 + \dots + x_n}{n} \text{ ou } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i. \quad (2.1)$$

Uma observação importante sobre a média aritmética é que ela é sensível aos extremos. Em dados muito dispersos, pode não indicar uma boa ideia do comportamento da variável, causando uma distorção na percepção da realidade.

Vamos calcular a média aritmética dos pontos dos times da Tabela 2.5, utilizando a Equação 2.1.

Vamos calcular a média aritmética dos pontos (P) dos times da Tabela 2.5, utilizando a Equação 2.1:

$$\bar{P} = \frac{\sum_{i=1}^{20} P_i}{20}.$$

Efetuada o cálculo com os dados da Tabela 2.5:

$$\bar{P} = \frac{1011}{20} \approx 50,55.$$

A média de pontos dos 20 clubes da Tabela 2.5 é $\bar{P} = 50,55$. Se substituirmos o clube de menor pontuação (32 pontos) por um time hipotético invicto ($P_{\text{máx}} = 114$), a nova média torna-se

$$\bar{P} = \frac{\sum_{i=1}^{20} P_i - 32 + 114}{20} = \frac{1001 - 32 + 114}{20} = 54,15,$$

o que evidencia a sensibilidade da média aritmética à presença de valores atípicos.

2.4.2 Mediana

Quando organizados em ordem crescente ou decrescente, a mediana de um conjunto de dados é o valor que ocupa o centro dos dados. Em outras palavras, a mediana é o valor que divide o conjunto de dados ao meio: 50% dos valores ficam abaixo dela, e os outros 50% ficam acima. E, portanto, a mediana é uma medida de tendência central.

Se as observações $x_1, x_2, \dots, x_n$ são tais que $x_1 < x_2 < \dots < x_n$, a mediana é dada por:

$$\text{Mediana} = \begin{cases} x_{(\frac{n+1}{2})}, & \text{se } n \text{ é ímpar} \\ \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}, & \text{se } n \text{ é par} \end{cases}$$

Vamos agora calcular a mediana dos 20 times da Tabela 2.5. Os pontos (P) dos clubes, em ordem crescente, são:

$$32, 33, 37, 38, \dots, 47, 49, 50, 55, \dots, 64, 68.$$

Como o número de observações é par ($n = 20$), a mediana é a média aritmética dos valores que ocupam as posições $\frac{n}{2} = 10$ e $\frac{n}{2} + 1 = 11$:

$$\text{Mediana} = \frac{49 + 50}{2} = 49,5.$$

Mesmo que incluíssemos um time hipotético invicto, com 114 pontos, a mediana se manteria praticamente inalterada, pois considera apenas a posição central dos dados. Essa característica a torna menos sensível a valores extremos, diferentemente da média aritmética. Entretanto, por levar em conta somente os elementos centrais, a mediana ignora a magnitude das demais observações. Por exemplo, no conjunto $\{3, 7, 100\}$, a mediana é 7, não refletindo a presença do valor extremo 100.

2.4.3 Moda

A moda é uma medida que representa a observação com maior frequência, como pode ocorrer de numa certa observação termos todas as observações diferentes, diz nesse caso que temos um conjunto de dados **amodal**. Já quando há 2 valores com a mesma frequência, dizemos que o conjunto de dados é **bimodal**, quando três ou mais observações têm a mesma frequência, dizemos que o conjunto de dados é **Multimodal**. Já quando apenas uma observação tem maior frequência, dizemos que o conjunto de dados é **Unimodal**.

Para calcular a moda dos pontos (P) dos 20 times da Tabela 2.5, identificamos o(s) valor(es) que ocorre(m) com maior frequência no conjunto de dados. Os pontos, em ordem crescente, são:

$$32, 33, 37, 38, \dots, 47, 47, \dots, 60, 60, \dots, 64, 64, 68.$$

A contagem de frequências mostra que:

- 47, 55, 60 e 64 aparecem **2 vezes** cada.
- Todos os demais valores aparecem apenas uma vez.

Portanto, o conjunto é **multimodal**, possuindo como modas:

$$\text{Moda} = 47, 55, 60, 64.$$

2.4.4 Medidas de Dispersão

Considere dois atacantes fictícios de futebol, denominados João e Pedro. Durante um campeonato com dez partidas, ambos tiveram a mesma média aritmética de gols por jogo (1 gol por partida), porém com distribuições diferentes ao longo dessas partidas.

Atacante João

Sequência de gols (10 jogos):

$$1, 1, 1, 1, 1, 1, 1, 1, 1, 1.$$

Média aritmética:

$$\bar{x}_{\text{João}} = \frac{1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1}{10} = 1.$$

Atacante Pedro

Sequência de gols (10 jogos):

$$0, 3, 2, 0, 0, 1, 0, 2, 0, 2.$$

Média aritmética:

$$\bar{x}_{\text{Pedro}} = \frac{0 + 3 + 2 + 0 + 0 + 1 + 0 + 2 + 0 + 2}{10} = 1.$$

Contudo, observa-se que as dispersões são bastante diferentes. João possui um desempenho constante, enquanto Pedro apresenta maior variabilidade no número de gols marcados. Esse exemplo evidencia a importância das medidas de dispersão, mostrando que, apesar de a média aritmética ser útil, ela pode ocultar diferenças relevantes na regularidade dos resultados.

Para avaliar quão distantes os dados estão da média aritmética, consideremos duas medidas de dispersão: o desvio médio e a variância. Seja um conjunto $X = \{x_1, \dots, x_n\}$ com média $\bar{x}$.

Desvio Médio (DM)

Definição:

$$\text{DM}(X) = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|.$$

Variância (Var)

Tratando o conjunto como população:

$$\text{Var}(X) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2.$$

O desvio médio (ou desvio absoluto médio) é menos utilizado que o desvio padrão por conveniência algébrica: o uso de quadrados, em vez de valores absolutos, se encaixa naturalmente em resultados e operações em probabilidade e estatística.

O **desvio padrão** (dp) é a raiz quadrada da variância:

$$dp = \sqrt{\text{Var}(X)} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}.$$

Assim, o desvio padrão mantém a mesma unidade dos dados e é bastante usado para expressar o afastamento típico em torno da média.

Cálculo das medidas no exemplo**Atacante João**

$$\text{DM}_{\text{João}} = \frac{1}{10} \sum_{i=1}^{10} |x_i - \bar{x}_{\text{João}}| = \frac{1}{10} (0 + \dots + 0) = 0.$$

$$\text{Var}_{\text{João}} = \frac{1}{10} \sum_{i=1}^{10} (x_i - \bar{x}_{\text{João}})^2 = \frac{1}{10} (0 + \dots + 0) = 0.$$

$$dp_{\text{João}} = \sqrt{\text{Var}_{\text{João}}} = \sqrt{0} = 0.$$

Atacante Pedro

$$\begin{aligned} \text{DM}_{\text{Pedro}} &= \frac{1}{10} (|0 - 1| + |3 - 1| + |2 - 1| + |0 - 1| + |0 - 1| \\ &\quad + |1 - 1| + |0 - 1| + |2 - 1| + |0 - 1| + |2 - 1|) = \frac{10}{10} = 1. \end{aligned}$$

$$\begin{aligned} \text{Var}_{\text{Pedro}} &= \frac{1}{10} [(0 - 1)^2 + (3 - 1)^2 + (2 - 1)^2 + (0 - 1)^2 + (0 - 1)^2 \\ &\quad + (1 - 1)^2 + (0 - 1)^2 + (2 - 1)^2 + (0 - 1)^2 + (2 - 1)^2] = \frac{12}{10} = 1,2. \end{aligned}$$

$$dp_{\text{Pedro}} = \sqrt{\text{Var}_{\text{Pedro}}} = \sqrt{1,2} \approx 1,095.$$

Síntese

No caso de João, todas as medidas de dispersão resultam em zero, indicando regularidade total (um gol em cada jogo). Já Pedro, embora com a mesma média, apresenta $DM > 0$, $Var > 0$ e $dp > 0$, evidenciando distribuição desigual dos gols e maior inconstância de desempenho.

2.4.5 Quartis

Os quartis separam um conjunto de dados ordenados em quatro partes iguais. A mediana é um quartil, indicada por $q_2 = q_{0,5}$, e divide o conjunto de dados em duas partes, nas quais 50% dos valores são menores que q_2 e os outros 50% são maiores que q_2 . Os principais quartis são:

- **Primeiro quartil** ($q_1 = q_{0,25}$): também chamado de quartil inferior, separa os 25% menores valores dos 75% maiores. Corresponde à mediana da metade inferior dos dados.
- **Segundo quartil** ($q_2 = q_{0,5}$): é a mediana do conjunto de dados, como já mencionado. Divide os dados em duas metades iguais.
- **Terceiro quartil** ($q_3 = q_{0,75}$): também chamado de quartil superior, separa os 75% menores valores dos 25% maiores. Corresponde à mediana da metade superior dos dados.

Em alguns casos, outros quartis podem ser utilizados, como q_i , com $0 < i < 1$, que representa o quartil em que $100\% \cdot i$ dos valores estão abaixo de q_i .

Com base na Tabela 2.5, vamos determinar os três quartis principais em relação à variável P (Pontos), com os dados fornecidos. A verificação fica por conta do leitor:

- **Primeiro quartil:** $q_1 = q_{0,25} = 55$
- **Segundo quartil (mediana):** $q_2 = q_{0,5} = 60$
- **Terceiro quartil:** $q_3 = q_{0,75} = 64$

A análise desses valores permite compreender a distribuição do desempenho das equipes ao longo da competição:

- **Primeiro quartil** ($q_1 = 55$): indica que 25% dos times possuem até 55 pontos. Esses são os times com desempenho relativamente mais baixo dentro do grupo analisado. Podemos dizer que quatro times possuem pontuação inferior ou igual a esse valor.
- **Segundo quartil ou mediana** ($q_2 = 60$): representa o valor central da distribuição. Isso significa que metade dos times possui pontuação igual ou inferior a 60 pontos, e a outra metade possui pontuação igual ou superior. A mediana oferece uma medida de tendência central menos sensível a extremos do que a média.
- **Terceiro quartil** ($q_3 = 64$): indica que 75% dos times têm pontuação igual ou inferior a 64. Apenas 25% dos times ultrapassaram esse valor, estando entre os de melhor desempenho no grupo analisado.

Em resumo, os quartis mostram que a maior parte das equipes está concentrada em uma faixa de pontuação entre 55 e 64 pontos, e apenas uma pequena parte (os 25% superiores) conseguiu ultrapassar esse patamar, com destaque para o time líder, que alcançou 68 pontos.

2.4.6 *Boxplot*

Para visualizar de forma mais clara os dados relacionados aos quartis, utiliza-se um *boxplot*, que consiste em um diagrama de caixas no qual é possível observar a distribuição dos valores e identificar o valor mínimo, o primeiro quartil, o segundo quartil (mediana), o terceiro quartil e o valor máximo. A Figura 2.7 apresenta o *boxplot* da variável **Pontuação** dos times listados na Tabela 2.5.

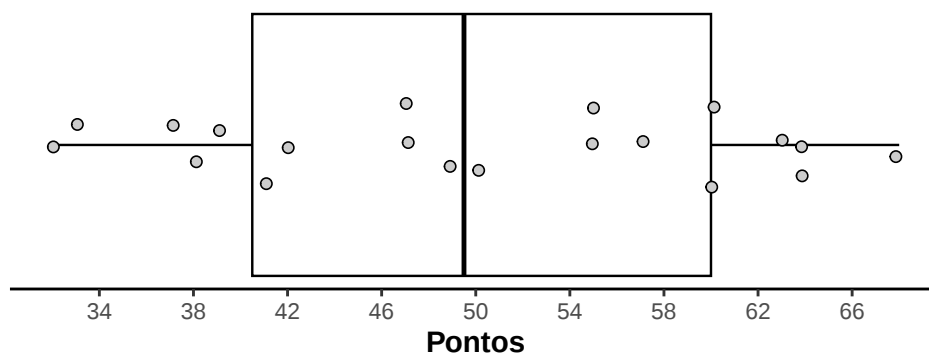


Figura 2.7. Boxplot das pontuações dos times do Brasileirão Série B

Fonte: Tabela 2.5

O *boxplot* mostra a distribuição das pontuações dos 20 times. A caixa abrange os 50% centrais dos dados, com a mediana localizada em cinquenta e sete pontos. As extremidades da “caixa” indicam os valores mínimo (32) e máximo (68). Os pontos sobre o gráfico representam as pontuações individuais de cada equipe, permitindo visualizar a posição de cada uma na distribuição. O time que obteve 68 pontos destaca-se como o de melhor desempenho, enquanto aquele que atingiu 32 pontos aparece na extremidade inferior da escala. A posição no eixo x indica a pontuação de cada time, enquanto a coordenada vertical foi deslocada aleatoriamente para evitar que equipes com pontuações iguais ou semelhantes ficassem sobrepostas.

PROBABILIDADE E VARIÁVEIS ALEATÓRIAS

No capítulo anterior, foram abordadas técnicas de resumo de um conjunto de dados por meio de tabelas de frequências, gráficos e medidas numéricas. Neste capítulo, abordamos conceitos básicos de probabilidade, incluindo os conceitos clássico e frequentista, e alguns modelos probabilísticos, como as distribuições binomial, geométrica, hipergeométrica, Poisson e binomial negativa.

Muitas vezes é impossível ou inviável obter um parâmetro populacional exato. Dessa forma, coleta-se uma amostra da população e deseja-se que as frequências relativas das observações reflitam o comportamento de toda a população. Morettin e Bussab (2010, p. 123) afirmam que “as frequências relativas podem ser interpretadas como estimativas das probabilidades associadas a eventos de interesse”. Com hipóteses apropriadas, é possível construir um modelo teórico que represente de forma razoável a distribuição dessas frequências, mesmo sem observar diretamente o fenômeno aleatório. Esses modelos, conhecidos como modelos probabilísticos, permitem descrever matematicamente a distribuição dos eventos observados.

3.1 Fundamentos e Propriedades da Probabilidade

Definição 3.1. Experimento Aleatório. É o experimento cujos resultados não se pode prever com certeza, mesmo quando repetido em condições idênticas.

Mesmo que dois times joguem repetidamente entre si nas mesmas condições, o placar final pode ser diferente a cada jogo devido a fatores imprevisíveis, como erros individuais, decisões da arbitragem e variações no desempenho dos jogadores.

Definição 3.2. Espaço Amostral. É o conjunto Ω de todos os resultados possíveis de um experimento aleatório.

Num campeonato de futebol, se o experimento aleatório for o resultado de uma partida entre dois times A e B , o espaço amostral pode ser representado por

$$\Omega = \{\text{vitória do time } A, \text{ empate, derrota do time } A\}.$$

Caso o interesse seja o número de gols marcados em uma partida de futebol, o espaço amostral pode ser o conjunto de todos os possíveis placares, $\Omega = \{0 - 0, 1 - 0, 0 - 1, 2 - 1, 1 - 2, \dots\}$.

Definição 3.3. Evento. É qualquer subconjunto do espaço amostral Ω .

Podemos afirmar que um evento $E \subseteq \Omega$ ocorre sempre que o resultado do experimento pertencer a ele. Por exemplo, considerando o espaço amostral dos placares possíveis de um jogo, um evento pode ser “o time A marcar pelo menos dois gols e não perder o jogo”, que pode ser representado pelo subconjunto

$$E = \{2 - 0, 2 - 1, 2 - 2, 3 - 0, 3 - 1, 3 - 2, 3 - 3, \dots\}.$$

Definição 3.4. Probabilidade. Seja $\mathcal{E}$ um experimento aleatório e Ω o espaço amostral associado a $\mathcal{E}$. A cada evento $A \subseteq \Omega$ associamos um número real $P(A)$, denominado *probabilidade de A* , que satisfaça às seguintes propriedades:

1. $0 \leq P(A) \leq 1$;
2. $P(\Omega) = 1$;
3. Se A e B forem eventos mutuamente excludentes, então

$$P(A \cup B) = P(A) + P(B);$$

4. Se $A_1, A_2, \dots$ forem, dois a dois, eventos mutuamente excludentes, então

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

A seguir, apresentamos duas formas de determinar a probabilidade de um evento, a abordagem clássica e a abordagem frequentista.

- **Abordagem Clássica da Probabilidade:** Sob a suposição de que todos os resultados possíveis de um espaço amostral Ω finito são igualmente prováveis, a probabilidade de um evento A , denotada por $P(A)$, é dada por

$$P(A) = \frac{\#A}{\#\Omega},$$

sendo que, $\#A$ e $\#\Omega$ representam o número de elementos dos conjuntos A e Ω , respectivamente.

Por exemplo, considerando os resultados de uma partida de futebol igualmente prováveis, sob o espaço amostral $\Omega = \{\text{derrota}, \text{empate}, \text{vitória}\}$, a probabilidade de ocorrer o evento $A = \{\text{empate}, \text{vitória}\}$ é dada por

$$P(A) = \frac{\#A}{\#\Omega} = \frac{2}{3} \approx 66,6\%.$$

- **Abordagem Frequentista de Probabilidade:** Nesta abordagem a probabilidade é calculada *a posteriori*. Após a realização de n repetições de um experimento aleatório, a probabilidade de um evento A , obtida por meio da observação direta dos resultados é dada por

$$P(A) = \frac{n_A}{n},$$

sendo que, n_A é o número de vezes em o evento A ocorreu em n repetições do experimento. Por exemplo, se após as cobranças de 100 pênaltis o goleiro defendeu 25, então a probabilidade estimada de $A = \{\text{defesa}\}$ é

$$P(A) = \frac{25}{100} = 25\%.$$

De maneira geral, a abordagem clássica se baseia na hipótese de simetria dos resultados possíveis, enquanto a abordagem frequentista utiliza observações repetidas do experimento sob condições idênticas.

3.2 Propriedades Fundamentais de Probabilidade

A seguir, apresentamos algumas propriedades básicas de probabilidade. Todas podem ser demonstradas diretamente a partir dos três axiomas fundamentais de probabilidade abordados na definição 3.4.

1. Monotonicidade:

Se $A \subseteq B$, então $P(A) \leq P(B)$. Em outras palavras, se todo elemento de A também está em B , não é possível que a probabilidade de A seja maior que B .

2. Probabilidade do Conjunto Vazio:

$P(\emptyset) = 0$. Como o conjunto vazio não descreve ocorrência alguma, logo, sua probabilidade é zero.

3. Limite de Valores:

Para qualquer evento A , vale: $0 \leq P(A) \leq 1$. Isto significa que uma probabilidade nunca pode ser negativa nem exceder 1.

4. Complementação:

Para qualquer evento A , o seu complementar é $\bar{A}$. Assim, $P(\bar{A}) = 1 - P(A)$.

5. União de Dois Eventos:

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Ao somar $P(A)$ e $P(B)$, a probabilidade do evento $A \cap B$ é contada duas vezes, devendo ser subtraída uma vez para corrigir o valor.

6. Subaditividade:

Para qualquer coleção finita de eventos $A_1, A_2, \dots, A_n$,

$$P(A_1 \cup A_2 \cup \dots \cup A_n) \leq \sum_{i=1}^n P(A_i).$$

Esta desigualdade reflete o fato de que pode haver intersecções entre os eventos, de modo que a união não precisa aumentar linearmente com a soma das probabilidades.

7. Continuidade de Probabilidade (Crescente e Decrescente):

(a) *Continuidade Crescente*: Se $A_1 \subseteq A_2 \subseteq \dots$ e $A = \bigcup_{n=1}^{\infty} A_n$, então

$$\lim_{n \rightarrow \infty} P(A_n) = P(A).$$

(b) *Continuidade Decrescente*: Se $B_1 \supseteq B_2 \supseteq \dots$ e $B = \bigcap_{n=1}^{\infty} B_n$, com $P(B_1) < \infty$, então

$$\lim_{n \rightarrow \infty} P(B_n) = P(B).$$

8. Adição Finita para Eventos Mutuamente Exclusivos:

Se $A_1, A_2, \dots, A_n$ forem mutuamente exclusivos (ou seja, $A_i \cap A_j = \emptyset$ para $i \neq j$), então

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = \sum_{i=1}^n P(A_i).$$

9. Aditividade Contável (ou σ -aditividade):

Se tivermos uma quantidade enumerável de eventos disjuntos $A_1, A_2, A_3, \dots$, então

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

3.3 Probabilidade Condicional

Definição 3.5. Dado um espaço amostral Ω e os eventos $A, B \subseteq \Omega$ com $P(B) > 0$, a probabilidade condicional de A dado B é definida por

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \quad (3.1)$$

A expressão em (3.1) representa a probabilidade de ocorrência do evento A , sob a condição de que o evento B já tenha ocorrido. Ela quantifica como a ocorrência de B altera nossa expectativa sobre A . Quando $P(A|B) \neq P(A)$, dizemos que B é informativo sobre A .

Exemplo 3.1. Considere um campeonato de futebol em que se sabe, a partir de dados históricos, que:

- A probabilidade de um time jogar em casa e vencer a partida é $P(\text{Vitória} \cap \text{Casa}) = 0,36$;
- A probabilidade de o time jogar em casa é $P(\text{Casa}) = 0,60$.

Deseja-se calcular a probabilidade de o time vencer, dado que ele jogou em casa. Ou seja, queremos $P(\text{Vitória}|\text{Casa})$.

Aplicando a expressão (3.1) de probabilidade condicional:

$$P(\text{Vitória}|\text{Casa}) = \frac{P(\text{Vitória} \cap \text{Casa})}{P(\text{Casa})} = \frac{0,36}{0,60} = 0,60.$$

Teorema de Bayes e Teorema da Probabilidade Total

Usando o fato de que $P(A \cap B) = P(A)P(B|A)$, o teorema de Bayes em sua forma mais simples é obtido reescrevendo a equação (3.1) na forma

$$P(A|B) = P(A) \frac{P(B|A)}{P(B)}. \quad (3.2)$$

Dessa forma, temos um novo significado à probabilidade condicional $P(A|B)$. Para obtermos $P(A|B)$, basta multiplicar $P(A)$ por $P(B|A)/P(B)$. Uma versão generalizada do teorema de Bayes em (3.2) é dada pela equação (3.4), a qual, é especialmente aplicada quando o espaço amostral é particionado pelos eventos $A_1, A_2, \dots, A_k$ do seguinte modo:

- $A_i \cap A_j = \emptyset$, para todo $i \neq j$;
- $\bigcup_{i=1}^k A_i = \Omega$;
- $P(A_i) > 0$, para todo i .

Se B é um evento tal que $P(B) > 0$, então a probabilidade de B pode ser expressa na forma

$$P(B) = \sum_{i=1}^k P(B | A_i)P(A_i). \quad (3.3)$$

A expressão (3.3) é muitas vezes denominada de teorema da probabilidade total. Ela é utilizada na construção do teorema de Bayes que determina a probabilidade de um evento A_i dado que outro evento B ocorreu,

$$P(A_i | B) = \frac{P(B | A_i)P(A_i)}{\sum_{j=1}^k P(B | A_j)P(A_j)}, i = 1, \dots, k. \quad (3.4)$$

Independência de Eventos

Definição 3.6. Independência para dois eventos. Dois eventos $A, B \subseteq \Omega$ são independentes se, e somente se:

$$P(A \cap B) = P(A)P(B). \quad (3.5)$$

Esta condição é equivalente a $P(A|B) = P(A)$ e $P(B|A) = P(B)$.

Exemplo 3.2. Considere os eventos: A : "time visitante marca gol no primeiro tempo", com $P(A) = 0.3$ e B : "time visitante marca gol no segundo tempo", com $P(B) = 0.4$.

Se $P(A \cap B) = 0.12 = (0.3)(0.4)$, então A e B são independentes. Neste caso, saber que o time visitante marcou no primeiro tempo não altera a probabilidade de marcar no segundo tempo.

Definição 3.7. Independência para três eventos. Três eventos A, B e C são mutuamente independentes se, e somente se:

- (i) $P(A \cap B) = P(A)P(B)$,
- (ii) $P(A \cap C) = P(A)P(C)$,
- (iii) $P(B \cap C) = P(B)P(C)$,
- (iv) $P(A \cap B \cap C) = P(A)P(B)P(C)$.

Apenas satisfazer as condições par a par (i)-(iii) não implica independência mútua. A condição (iv) é essencial para independência completa.

Exemplo 3.3. Considere para um jogador em três partidas os eventos A, B e C tais que: A : "Faz gol na 1ª partida" ($P(A) = 0.5$), B : "Faz gol na 2ª partida" ($P(B) = 0.5$) e C : "Faz gol em ambas partidas 1 e 2 ou em nenhuma" ($P(C) = 0.5$)

Embora $P(A \cap B) = P(A)P(B)$ e similar para outros pares, temos $P(A \cap B \cap C) = 0.25 \neq 0.125 = P(A)P(B)P(C)$. Portanto, os eventos são independentes dois a dois mas não mutuamente independentes.

Definição 3.8. Independência para n eventos. Uma coleção de n eventos $A_1, A_2, \dots, A_n$ é mutuamente independente se para todo subconjunto de índices $1 \leq i_1 < i_2 < \dots < i_k \leq n$ com $2 \leq k \leq n$, temos:

$$P\left(\bigcap_{j=1}^k A_{i_j}\right) = \prod_{j=1}^k P(A_{i_j}). \quad (3.6)$$

Isso requer que todas as $2^n - n - 1$ possíveis interseções satisfaçam a condição de fatoração.

Exemplo 3.4. Considere 4 jogos consecutivos de um time. Sejam A_i : "vitória no i -ésimo jogo", com $P(A_i) = 0.6$ para $i = 1, 2, 3, 4$. Se os resultados são independentes, temos por exemplo,

$$P(A_1 \cap \bar{A}_2 \cap A_3 \cap \bar{A}_4) = (0.6)(0.4)(0.6)(0.4) = 0.0576$$

e

$$P(\text{ao menos 3 vitórias}) = \binom{4}{3}(0.6)^3(0.4)^1 + \binom{4}{4}(0.6)^4(0.4)^0 = 0.4752.$$

3.4 Variáveis aleatórias

Definição 3.9. Seja Ω um espaço amostral. Uma *função* X , que associa a cada elemento $\omega \in \Omega$ um número real $X(\omega)$, é denominada *variável aleatória*.

É comum na literatura encontrarmos outras denominações, como variável estocástica. Essas variáveis são geralmente representadas por letras maiúsculas, sendo as mais comuns X e Y . Dependendo do conjunto de valores que a variável pode assumir (isto é, sua imagem), ela pode ser classificada em:

- **Discretas:** têm como imagem um conjunto finito ou enumerável de valores;
- **Contínuas:** têm como imagem um intervalo (ou união de intervalos) de números reais.

"Como ressaltam autores consagrados, (MEYER, 1989, p. 56) observa que: "A terminologia acima é um tanto infeliz, mas é tão universalmente aceita, que não nos afastaremos dela. Tornamos tão claro quanto possível que X é uma função, e, contudo, a denominamos uma variável (aleatória)!"

No contexto do futebol, são exemplos de variáveis aleatórias:

- X_1 = o número de gols marcados pelo time da casa durante a partida;
- X_2 = o número de gols marcados pelo time visitante durante a partida;
- Y = a diferença entre número de gols marcados pelos times da casa e visitante;
- X_3 = o número de finalizações pelo time da casa até ocorrer o segundo gol;
- T_1 = o tempo até ocorrer o primeiro gol marcado pelo time da casa;
- T_2 = o tempo até ocorrer o primeiro gol marcado pelo time visitante;
- T = o tempo até ocorrer o primeiro gol marcado numa partida de futebol.

Variável Aleatória Discreta:

Definição 3.10. Uma variável aleatória X é discreta se o número de valores possíveis de X for finito ou infinito numerável.

Exemplo 3.5.

- o número de gols marcados por uma equipe em uma partida;
- o número de escanteios obtidos por um time durante um jogo;
- o número de impedimentos assinalados contra uma equipe em uma partida.

3.5 Função de Probabilidade de uma Variável Aleatória X

A função de probabilidade de uma variável aleatória discreta é uma correspondência que associa a cada valor assumido por X uma probabilidade específica. Em outras palavras, ela indica a chance de a variável assumir determinado valor no conjunto imagem. Essa função é expressa por:

$$p(x_i) = P(X = x_i), \quad (3.7)$$

Para todo x_i pertencente à imagem de X .

Exemplo 3.6. Seja X uma variável aleatória que representa o número de gols marcados pelo time visitante A em um jogo específico. Suponha que esta variável assume valores no conjunto $\{0, 1, 2, 3, 4\}$. A Tabela 3.1 mostra probabilidades atribuídas aos números de gols marcados. A notação $P(X = x) = p(x)$ indica a probabilidade de a variável aleatória X assumir o valor x , sendo x um elemento do conjunto dos valores possíveis da imagem da v.a. X . A FP de X também pode ser apresentada na forma gráfica, como mostra a Figura 3.1.

Tabela 3.1. Distribuição de probabilidade do número de gols marcados pelo time visitante A .

Número de Gols (x)	$P(X = x)$
0	$P(X = 0) = 0,30$
1	$P(X = 1) = 0,40$
2	$P(X = 2) = 0,20$
3	$P(X = 3) = 0,08$
4	$P(X = 4) = 0,02$

Fonte: Autor (2025).

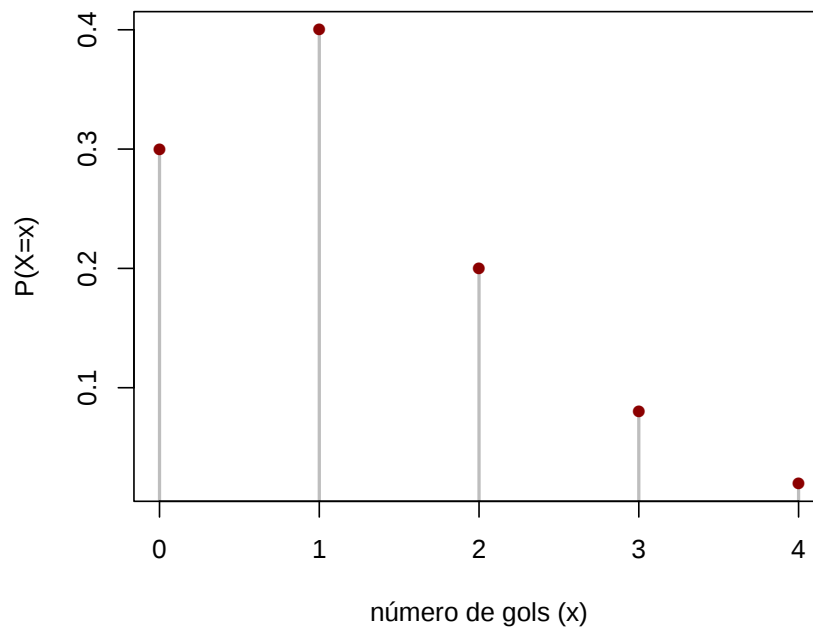


Figura 3.1. Função distribuição acumulada da v.a. X do Exemplo 3.6.

Fonte: 3.6

De forma resumida, a distribuição da Tabela 3.1 satisfaz os três axiomas de Kolmogorov como mostrado a seguir:

$$P(X = x) \geq 0, \quad \text{para todo } x \in \{0, 1, 2, 3, 4\};$$

$$\sum_x P(X = x) = 0,30 + 0,40 + 0,20 + 0,08 + 0,02 = 1;$$

$$P\left(\bigcup_x (X = x)\right) = \sum_x P(X = x) = 1.$$

3.6 Função de Distribuição Acumulada

A função de distribuição acumulada de uma variável aleatória discreta X , denotada por $F(x)$, é uma função que associa a cada número real x a probabilidade de que a variável aleatória X assuma um valor menor ou igual a x . Em outras palavras, representa a soma das probabilidades dos valores de X menores ou iguais a x . Essa função é expressa por:

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} P(X = x_i) \quad (3.8)$$

onde a soma é tomada sobre todos os valores x_i pertencentes à imagem de X tal que $x_i \leq x$.

A distribuição acumulada F correspondente a função de probabilidade da v.a. X do Exemplo 3.6, é dada por:

$$F(x) = \begin{cases} 0 & \text{se } x < 0, \\ 0,30 & \text{se } 0 \leq x < 1, \\ 0,70 & \text{se } 1 \leq x < 2, \\ 0,90 & \text{se } 2 \leq x < 3, \\ 0,98 & \text{se } 3 \leq x < 4, \\ 1 & \text{se } x \geq 4. \end{cases}$$

A função acumulada de X F também pode ser apresentada na forma gráfica, como mostra a Figura 3.2.

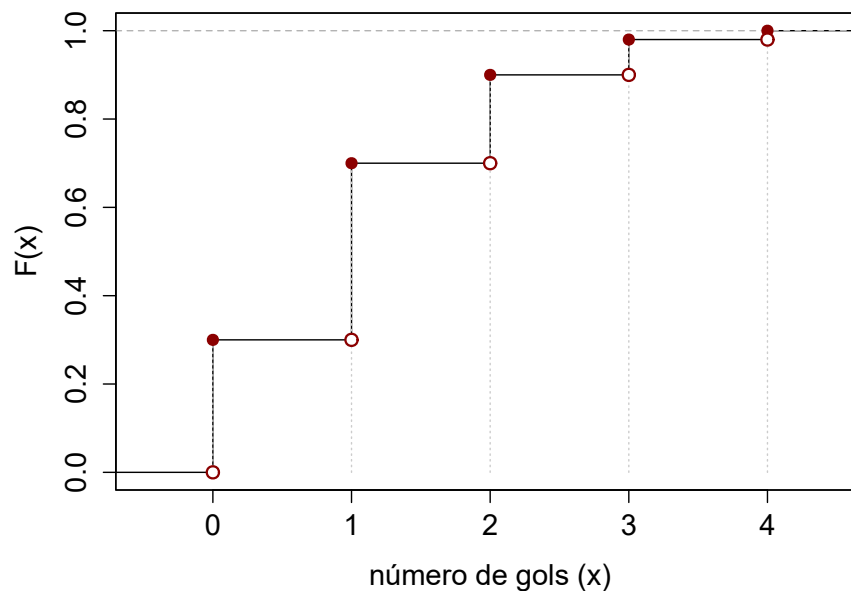


Figura 3.2. Função distribuição acumulada da v.a. X do Exemplo 3.6.

Fonte: Autor (2025)

3.7 Valor Esperado de uma Variável Aleatória

O valor esperado, também denominado de esperança matemática ou valor médio de uma variável aleatória discreta X , é dado por:

$$E(X) = \sum_x xp(x). \quad (3.9)$$

A notação $E(X) = \mu$ é usada para representar o valor esperado de uma variável aleatória X .

Temos a seguir algumas propriedades do valor esperado de uma v.a. A prova de cada uma pode ser vista em Meyer (1989). Seja X uma v.a. e c uma constante, então

- (i) $E(c) = c$,
- (ii) $E(cX) = cE(X)$,

- (iii) $E(X + Y) = E(X) + E(Y)$,
- (iv) $E(X + c) = E(X) + c$.

Se a v.a. X for uma uniforme discreta. Isto significa que X pode assumir n valores, todos com probabilidades iguais, $P(X = x_1) = P(X = x_2) = \dots = P(X = x_n) = \frac{1}{n}$. Então, o valor esperado, dado pela equação (3.9), se torna igual a média aritmética dos valores $(x_1, x_2, \dots, x_n)$:

$$E(X) = \frac{x_1 + x_2 + \dots + x_n}{n}.$$

Outra característica muito explorada de uma v.a. é a variância definida a seguir.

3.8 Variância de uma Variável Aleatória Discreta

A variância de uma variável aleatória, definida pela equação (3.10), é uma medida que expressa a dispersão ou o grau de espalhamento dos valores assumidos por X em torno da sua esperança matemática.

$$Var(X) = \sum_{i=1}^n (x_i - E(X))^2 P(X = x_i). \quad (3.10)$$

A notação $Var(X) = \sigma^2$ é usada para representar a variância de uma v.a. X . Uma forma de escrever a $Var(X)$ dada por

$$Var(X) = E(X^2) - [E(X)]^2. \quad (3.11)$$

Vamos agora deduzir a equação (3.11), expandindo o termo quadrático na equação (3.10), obtemos:

$$Var(X) = \sum_{i=1}^n (x_i^2 - 2x_i E(X) + [E(X)]^2) P(X = x_i).$$

Distribuindo a soma e as probabilidades, temos:

$$Var(X) = \sum_{i=1}^n x_i^2 P(X = x_i) - 2E(X) \sum_{i=1}^n x_i P(X = x_i) + [E(X)]^2 \sum_{i=1}^n P(X = x_i)$$

Observando a definição de valor esperado dada pela equação (3.9), temos que,

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

Além disso, a soma de todas as probabilidades é igual a 1, $\sum_{i=1}^n P(X = x_i) = 1$. Então, simplificando, podemos escrever a variância na forma:

$$Var(X) = E(X^2) - [E(X)]^2,$$

em que,

$$E(X^2) = \sum_{i=1}^n x_i^2 P(X = x_i).$$

Portanto, a equação (3.11) mostra que a variância pode ser calculada através da diferença entre o valor esperado dos quadrados da variável aleatória X e o quadrado do seu valor esperado $E(X)$.

Se X for uma v.a. e a uma constante, são válidas as seguintes propriedades.

$$\text{Var}(a + X) = \text{Var}(X) \quad (3.12)$$

e

$$\text{Var}(aX) = a^2 \text{Var}(X). \quad (3.13)$$

Se X e Y são variáveis aleatórias independentes, então:

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y). \quad (3.14)$$

3.9 Desvio Padrão de uma Variável Aleatória

O desvio padrão de uma variável aleatória X , $\sigma = dp(X)$, dado pela Equação (3.15) é à raiz quadrada da variância, $\sigma = \sqrt{\sigma^2}$.

$$dp(X) = \sqrt{\text{Var}(X)} = \sqrt{E[(X - E(X))^2]}. \quad (3.15)$$

Devemos observar que σ está na mesma unidade de medida da v.a. X .

3.10 Covariância entre duas Variáveis Aleatórias

A covariância é uma medida de dependência linear entre as variáveis. Se a covariância entre duas v.a. (X, Y) for zero não há dependência linear entre elas, entretanto pode haver outro tipo de relação e, portanto, não podemos afirmar que (X, Y) são independentes. Empregaremos a covariância para a distribuição de Poisson bivariada, em que são necessários três parâmetros, λ_1 , λ_2 e λ_3 , sendo que os dois primeiros representam as médias de gols dos dois times adversários, e λ_3 representa a covariância entre as variáveis aleatórias que modelam o número de gols de cada time. A equação (3.16) apresenta a definição de covariância para duas variáveis aleatórias X e Y .

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]. \quad (3.16)$$

3.11 Modelos Probabilísticos de Variáveis Aleatórias Discretas

Os modelos probabilísticos para variáveis aleatórias discretas permitem descrever a probabilidade da ocorrência de eventos associados a v.a. X . A seguir, descrevemos alguns modelos importantes na análise de dados esportivos. Em especial o modelo de Poisson, muito utilizado para prever resultados de futebol.

3.11.1 Distribuição de Bernoulli

A distribuição de Bernoulli é a base de diversos modelos probabilísticos, constituindo o caso mais simples de variável aleatória discreta. Considere um experimento com apenas dois resultados possíveis: $\{sucesso\}$ e $\{fracasso\}$. Para formalizar, definimos uma variável aleatória X tal que:

$$X = \begin{cases} 1, & \text{se ocorrer sucesso,} \\ 0, & \text{se ocorrer fracasso.} \end{cases}$$

Definição 3.11. Uma variável aleatória X que assume valores em $\{0, 1\}$, com probabilidades

$$P(X = x) = \begin{cases} p, & x = 1, \\ 1 - p, & x = 0, \end{cases} \quad (3.17)$$

é denominada variável aleatória de Bernoulli com parâmetro $p \in (0, 1)$. Utilizamos a notação

$$X \sim \text{Ber}(p).$$

Esta distribuição descreve, portanto, a probabilidade associada a um único ensaio de Bernoulli, conceito fundamental em para a distribuição binomial, apresentada na sequência, que é justamente a soma de n variáveis independentes de Bernoulli.

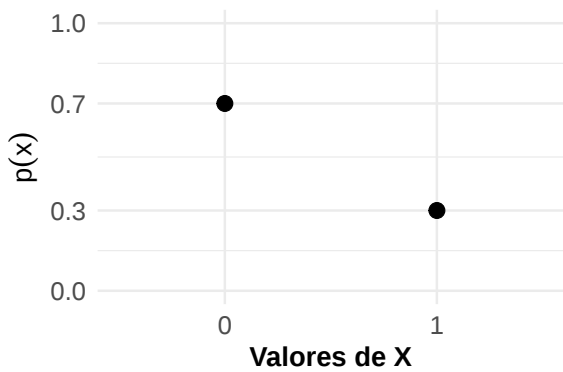
Proposição 3.1 (Esperança e variância da Bernoulli). *Se $X \sim \text{Ber}(p)$, então*

$$\mathbb{E}[X] = p \quad e \quad \text{Var}(X) = p(1 - p).$$

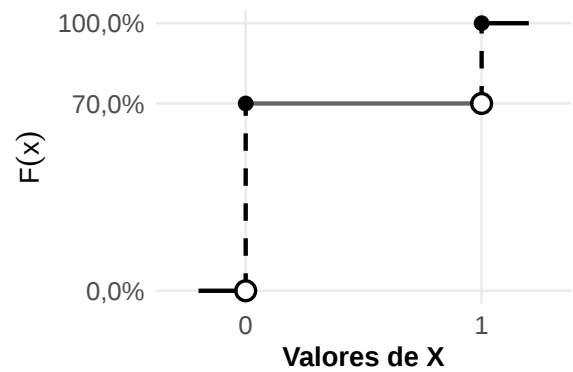
A função de distribuição acumulada $F(x) = P(X \leq x)$ de uma variável aleatória $X \sim \text{Ber}(p)$ é dada por:

$$F(x) = \begin{cases} 0, & x < 0, \\ 1 - p, & 0 \leq x < 1, \\ 1, & x \geq 1. \end{cases}$$

A Figura 3.3 apresenta, graficamente, a função de probabilidade e a função de distribuição acumulada para uma variável aleatória X com parâmetro $p = 0,3$.



(a) Função de probabilidade (fp) de $X \sim \text{Ber}(p)$.



(b) Função distribuição acumulada (fda) de $X \sim \text{Ber}(p)$.

Figura 3.3. Representações gráficas das funções fp e fda de uma v.a. $X \sim \text{Ber}(p = 0,3)$. Fonte: Autor (2025).

3.11.2 Distribuição Binomial

Considerando n ensaios de Bernoulli, uma lista como $(1, 0, 1, \dots, 0)$ representa uma amostra dos n ensaios. Sendo que, os valores 1 e 0 representam os eventos $\{\text{sucesso}\}$ e $\{\text{fracasso}\}$, respectivamente. A distribuição binomial está diretamente relacionada a quantidade de sucessos nos n ensaios de Bernoulli. Devemos observar que, se realizarmos 6 ensaios de Bernoulli e obtivermos a lista $(0, 1, 0, 0, 1, 0)$, então a probabilidade de $\{\text{sucesso}\}$ em que cada ensaio é independente dos demais. Temos,

$$P(\{0, 1, 0, 0, 1, 0\}) = (1-p)p(1-p)(1-p)p(1-p) = p^2(1-p)^4. \quad (3.18)$$

Qualquer sequência com 2 sucessos e 4 fracassos tem a mesma probabilidade $p^2(1-p)^4$, pois os eventos são independentes e mutuamente exclusivos. O número total de sequências possíveis com x sucessos em n ensaios é dado pelo coeficiente binomial $\binom{n}{x}$. Isto motiva a seguinte definição.

Definição 3.12. Uma variável aleatória X representando o número de sucessos em uma sequência de n ensaios de Bernoulli é chamada de variável aleatória Binomial com parâmetros n e p , cuja função probabilidade é dada por

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad \text{para } x = 0, 1, 2, \dots, n. \quad (3.19)$$

A notação $X \sim B(n, p)$ é utilizada para indicar uma v.a. X seguindo uma distribuição binomial com parâmetro n e p . Quando $n = 1$ a distribuição de probabilidade (3.19) é chamada de distribuição de Bernoulli.

Proposição 3.2. Esperança e variância da Binomial. Se $X \sim B(n, p)$, então

$$\mathbb{E}[X] = np \quad \text{e} \quad \text{Var}(X) = np(1-p).$$

3.11.3 Binomial Negativa

A distribuição binomial negativa modela o número de fracassos X antes que um número fixo r de sucessos seja alcançado em ensaios de Bernoulli independentes. Também é conhecida como distribuição de Pascal. A definição que melhor atende os propósitos deste trabalho é a seguinte.

Definição 3.13. Uma variável aleatória X segue uma distribuição binomial negativa com parâmetros r (número de sucessos desejados) e p (probabilidade de sucesso), denotada por $X \sim NB(r, p)$, se sua função de probabilidade for:

$$P(X = x) = \binom{x+r-1}{x} p^r (1-p)^x, \quad \text{para } x = 0, 1, 2, \dots \quad (3.20)$$

A obtenção de (3.20) segue diretamente a partir da distribuição Binomial. O evento $\{X = x\}$ ocorre somente se houver nos primeiros $x+r-1$ ensaios x fracassos, e um sucesso no ensaio $x+r$. Agora basta observar que, a probabilidade de x fracassos nos $x+r-1$ ensaios é igual a $\binom{x+r-1}{x} p^{r-1} (1-p)^x$, e que a probabilidade de um sucesso no ensaio $x+r$ é p . E o resultado (3.20) segue ao se multiplicar estas probabilidades.

Proposição 3.3. Esperança e variância da Binomial Negativa. Se $X \sim \text{NB}(r, p)$, então

$$\mathbb{E}[X] = \frac{r(1-p)}{p} \quad \text{e} \quad \text{Var}(X) = \frac{r(1-p)}{p^2}.$$

Um caso especial da distribuição Binomial Negativa é a distribuição Geométrica obtida ao se fazer na expressão (3.20) $r = 1$. A v.a. X , neste caso, é interpretada como o ensaio no qual ocorre o primeiro sucesso.

3.11.4 Distribuição Hipergeométrica

A distribuição Hipergeométrica é utilizada em situações nas quais se retiram elementos de uma população finita e sem reposição, sendo essa população composta por dois tipos distintos de elementos. Os que contém a propriedade A e os que não contém.

Cada retirada de um elemento da população altera a probabilidade das retiradas subsequentes, diferentemente da v.a. Binomial em que os ensaios são independentes.

Definição 3.14. Uma variável aleatória X segue uma distribuição Hipergeométrica, denotada por $X \sim \text{Hiper}(N, k, n)$ em que os parâmetros N , k e n , representam, respectivamente o tamanho da população, o número de elementos com a característica de interesse e o tamanho da amostra, se a sua função de probabilidade se sua função de probabilidade for dada por

$$P(X = x) = \frac{\binom{k}{x} \binom{N-k}{n-x}}{\binom{N}{n}}, \quad \text{para } x = \max(0, n - (N - k)) \leq x \leq \min(n, k). \quad (3.21)$$

Proposição 3.4. Esperança e variância da Hipergeométrica. Se $X \sim \text{Hiper}(N, k, n)$, então,

$$\mathbb{E}[X] = n \frac{k}{N} \quad \text{e} \quad \text{Var}(X) = n \frac{k}{N} \left(1 - \frac{k}{N}\right) \frac{N-n}{N-1}.$$

No contexto esportivo a distribuição Hipergeométrica pode ser empregada conforme o exemplo seguinte.

Exemplo 3.7. Selecionando aleatoriamente e sem reposição n de um total de N partidas de um campeonato de futebol, deseja-se estimar a probabilidade de ocorrer exatamente x vitórias de um determinado time, sabendo que ao todo ele possui k vitórias no campeonato.

Suponha que em um campeonato com $N = 38$ jogos, um clube venceu $k = 20$ partidas. Se escolhermos aleatoriamente $n = 5$ jogos desse clube para analisar, podemos empregar a distribuição Hipergeométrica para determinar a probabilidade de que exatamente x dessas partidas sejam vitórias.

3.11.5 Distribuição de Poisson

Definição 3.15. Uma variável aleatória X com valores possíveis no conjunto $\{0, 1, 2, \dots\}$ tem distribuição de Poisson com parâmetro $\lambda > 0$ se, sua função de probabilidade é dada por

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, 2, \dots \quad (3.22)$$

É imediato observar que a expressão (3.22) define uma função de probabilidade, pois

$$\sum_{i=0}^{\infty} P(X = x) = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda} e^{\lambda} = 1.$$

Proposição 3.5. Esperança e variância da distribuição de Poisson. Se $X \sim \text{Poisson}(\lambda)$, então

$$\mathbb{E}[X] = \lambda \quad \text{e} \quad \text{Var}(X) = \lambda.$$

Distribuição de Poisson como Limite da Binomial

Sob certas condições, a v.a. de Poisson pode ser usada como uma aproximação para a v.a. Binomial com parâmetros (n, p) no caso particular de n grande e p suficientemente pequeno, de modo que $np = \lambda$ tenha tamanho moderado. Considere a função de probabilidade Binomial:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}.$$

Podemos reescrever como:

$$P(X = x) = \frac{n(n-1)(n-2) \cdots (n-x+1)}{x!} p^x (1-p)^{n-x}.$$

Considerando a reparametrização $np = \lambda$. Segue-se que,

$$p = \frac{\lambda}{n}, \quad 1-p = 1 - \frac{\lambda}{n} = \frac{n-\lambda}{n}.$$

Assim,

$$P(X = x) = \frac{n(n-1) \cdots (n-x+1)}{x!} \left(\frac{\lambda}{n}\right)^x \left(\frac{n-\lambda}{n}\right)^{n-x}.$$

De forma mais conveniente,

$$P(X = x) = \frac{\lambda^x}{x!} \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{x-1}{n}\right) \left(1 - \frac{\lambda}{n}\right)^{n-x}.$$

Agora, tomemos o limite quando $n \rightarrow \infty$, com $np = \lambda$ constante. Note que:

$$\lim_{n \rightarrow \infty} \left(1 - \frac{j}{n}\right) = 1, \quad j = 1, 2, \dots, x-1,$$

e, além disso,

$$\lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda}.$$

Portanto,

$$\lim_{n \rightarrow \infty} P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda} = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, 2, \dots, \quad (3.23)$$

que é a função de probabilidade da distribuição de Poisson.

Outras Observações Sobre a Distribuição de Poisson

A distribuição que hoje leva o nome de Poisson remonta a um tratado de 1837 de Siméon Denis Poisson, no qual o autor aplicou a teoria das probabilidades à avaliação de decisões judiciais e de erros de veredictos. O trabalho em questão, intitulado *Recherches sur la probabilité des jugements en matière criminelle et en matière civile*, discutia a modelagem de contagens de eventos e forneceu a base conceitual para o que viria a ser conhecido como *distribuição de Poisson*. Ao longo do século XIX e início do XX, essa distribuição consolidou-se como o limite da binomial quando $n \rightarrow \infty$ e $p \rightarrow 0$ com $np = \lambda$ fixo, tornando-se o modelo de referência para contagens de eventos “raros” em intervalos de tempo ou de espaço (por exemplo, decadências radioativas, defeitos em processos industriais e chegadas em sistemas de filas) e servindo de fundamento ao processo de Poisson.

Na seção anterior, utilizamos a distribuição de Poisson como uma aproximação para a binomial em regimes de eventos raros (n grande, p pequeno e $\lambda = np$ fixo). É importante destacar, contudo, que a distribuição de Poisson não é usada somente como uma aproximação; ela é um modelo probabilístico fundamental por si só, ideal para descrever a contagem de eventos independentes em intervalos de tempo ou de espaço. Por essa razão, surge com frequência como representação simples e eficaz de uma ampla gama de fenômenos observáveis em contextos como confiabilidade, filas, telecomunicações e processos industriais. O leitor interessado no processo de poisson, poderá consultar Ross (2010).

A distribuição de Poisson determina a probabilidade de eventos raros que ocorrem em um espaço de tempo fixo. No contexto esportivo, o parâmetro λ pode ser usado para representar a média de gols marcados por um time, calculado conforme seu desempenho em partidas anteriores.

3.11.6 Distribuição Conjunta de Duas Variáveis Aleatórias

A distribuição conjunta das variáveis aleatórias X e Y descreve como se comportam probabilisticamente essas duas variáveis simultaneamente; também pode ser generalizada para mais de duas variáveis.

A função de probabilidade conjunta (ou distribuição de probabilidade conjunta) entre duas variáveis aleatórias discretas X e Y é dada por

$$P(X = x, Y = y) = f(x, y), \quad (3.24)$$

sujeita às condições:

- $f(x, y) \geq 0$;
- $\sum_{x \in X} \sum_{y \in Y} f(x, y) = 1$.

Suponhamos que X assuma os valores $x_1, x_2, \dots, x_m$ e Y assuma os valores $y_1, y_2, \dots, y_n$. Uma tabela de probabilidade conjunta pode ser apresentada conforme a Tabela 3.2. Observe que, ao fixar $X = x_j$, segue-se:

$$P(X = x_j) = f_1(x_j) = \sum_{k=1}^n f(x_j, y_k), \quad (3.25)$$

e, de forma análoga, ao fixar $Y = y_k$:

$$P(Y = y_k) = f_2(y_k) = \sum_{j=1}^m f(x_j, y_k). \quad (3.26)$$

Tabela 3.2. Distribuição conjunta de X e Y

$Y \backslash X$	y_1	y_2	$\dots$	y_{n-1}	y_n	Total
x_1	$f(x_1, y_1)$	$f(x_1, y_2)$	$\dots$	$f(x_1, y_{n-1})$	$f(x_1, y_n)$	$f_1(x_1)$
x_2	$f(x_2, y_1)$	$f(x_2, y_2)$	$\dots$	$f(x_2, y_{n-1})$	$f(x_2, y_n)$	$f_1(x_2)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
x_m	$f(x_m, y_1)$	$f(x_m, y_2)$	$\dots$	$f(x_m, y_{n-1})$	$f(x_m, y_n)$	$f_1(x_m)$
Total	$f_2(y_1)$	$f_2(y_2)$	$\dots$	$f_2(y_{n-1})$	$f_2(y_n)$	1

Fonte: Autor.

Como $f_1(x)$ e $f_2(y)$ são representadas nas margens de tabelas conjuntas, conforme a Tabela (3.2), elas são comumente chamadas de distribuições de probabilidades marginais de X e de Y , respectivamente. Dessa forma, definem-se:

$$f_1(x) = \sum_{y=1}^{n_y} f(x, y), \quad f_2(y) = \sum_{x=1}^{n_x} f(x, y).$$

Observe ainda que

$$\sum_{x=1}^{n_x} \sum_{y=1}^{n_y} f(x, y) = 1,$$

o que implica

$$\sum_{x=1}^{n_x} f_1(x) = 1 \quad \text{e} \quad \sum_{y=1}^{n_y} f_2(y) = 1.$$

A função de distribuição acumulada conjunta de (X, Y) é então dada por

$$F(x, y) = P(X \leq x, Y \leq y) = \sum_{\substack{x_i \leq x \\ i=1, \dots, n_x}} \sum_{\substack{y_j \leq y \\ j=1, \dots, n_y}} f(x_i, y_j).$$

Neste trabalho, aplicamos a distribuição de Poisson para representar o número de gols em partidas de futebol, considerando dados históricos dos jogos anteriores. Para modelar a probabilidade dos placares, utilizaremos o produto de duas distribuições de Poisson independentes, sendo $P(N_A = n_A)$ a probabilidade de o time A marcar exatamente n_A gols em uma partida, e $P(N_B = n_B)$ a probabilidade de o time B marcar exatamente n_B gols. Assim, a probabilidade conjunta será dada por:

$$P(N_A = n_A, N_B = n_B) = P(N_A = n_A)P(N_B = n_B), \quad (3.27)$$

em que cada termo é obtido pela função massa de probabilidade da distribuição de Poisson com parâmetros λ_A (para o time A) e λ_B (para o time B).

Esses parâmetros (λ_A e λ_B) podem ser modelados das seguintes formas:

- λ = força de ataque do time $\times$ força defensiva do adversário;
- λ = força de ataque do time $\times$ força defensiva do adversário $\times$ fator casa, caso o time jogue como mandante.

Adicionalmente, o modelo pode ser ajustado segundo a proposta de Dixon e Coles (1997), permitindo capturar adequadamente probabilidades relacionadas a placares baixos, tais como (0–0), (1–0), (0–1) e (1–1). Com esse ajuste, a fórmula para a probabilidade conjunta se torna:

$$P(N_A = k, N_B = j) = \tau(k, j)P(N_A = k)P(N_B = j), \quad (3.28)$$

onde $\tau(k, j)$ é o fator de correção introduzido por Dixon e Coles.

3.11.7 Distribuição de Poisson Bivariada

Ao utilizarmos duas distribuições de Poisson separadas para modelar a quantidade de gols de cada time, assumimos implicitamente que as equipes atuam de forma independente. No entanto, essa suposição ignora o fato de que um gol marcado por uma equipe pode influenciar o desempenho da outra. Como essa hipótese de independência nem sempre se verifica, opta-se pela utilização da distribuição de Poisson bivariada como alternativa mais realista.

A distribuição de Poisson bivariada permite modelar eventos com uma possível dependência entre os gols marcados por duas equipes adversárias, ao mesmo tempo em que não subestima os casos em que há independência, tal como ocorre na modelagem com duas Poissons independentes.

Sejam Z_1 , Z_2 e Z_3 variáveis aleatórias discretas, independentes entre si, que seguem distribuições de Poisson com médias λ_1 , λ_2 e λ_3 , respectivamente. Define-se então as variáveis:

$$\begin{aligned} X &= Z_1 + Z_3, \\ Y &= Z_2 + Z_3, \end{aligned}$$

Onde X e Y representam, respectivamente, o número de gols das duas equipes. Os pares (X, Y) também são variáveis aleatórias discretas.

A função de probabilidade conjunta da distribuição de Poisson bivariada é dada por:

$$P(X = x, Y = y) = e^{-(\lambda_1 + \lambda_2 + \lambda_3)} \sum_{k=0}^{\min(x, y)} \frac{\lambda_1^{x-k} \lambda_2^{y-k} \lambda_3^k}{(x-k)!(y-k)!k!},$$

em que:

- λ_1 e λ_2 são os parâmetros exclusivos de cada equipe;
- λ_3 é o parâmetro compartilhado, que introduz a correlação entre X e Y .

A utilização prática tanto da distribuição de Poisson univariada quanto da bivariada depende da estimação adequada de seus parâmetros, como λ , λ_1 , λ_2 e λ_3 . Esses valores são fundamentais para que o modelo reflita o comportamento real dos dados, permitindo previsões mais consistentes e interpretações estatísticas confiáveis. O método da máxima verossimilhança constitui uma alternativa robusta e frequentemente utilizada para a estimação de parâmetros, sendo discutido em detalhe no Apêndice C.

PROBABILIDADE NAS APOSTAS ESPORTIVAS

Neste capítulo, abordaremos inicialmente o conceito de *odds*, que representa a probabilidade de um time vencer, empatar ou ser derrotado, assim como o retorno que o apostador pode obter ao realizar uma aposta. Em um segundo momento, empregaremos o modelo de Poisson para previsão de placares, utilizando o produto de duas distribuições de Poisson independentes, cujos parâmetros são modelados para captar a força de ataque e defesa dos times. Esse modelo foi um dos pioneiros na predição de placares e será abordado como uma proposta didática para a abordagem de probabilidades de placares em sala de aula.

4.1 Odds

No universo das apostas esportivas, compreender como as probabilidades (ou odds) são apresentadas é o primeiro passo para tomar decisões informadas. As *odds* não apenas indicam a probabilidade de um resultado ocorrer, mas também determinam o valor do seu potencial retorno financeiro. Existem diferentes formatos para representar essas probabilidades, cada um com sua própria lógica de cálculo e popularidade de acordo com o tipo de esporte e região específica.

A seguir, vamos detalhar os três principais tipos de *odds* encontrados nas plataformas de apostas: as Decimais, as mais comuns no Brasil; as Fracionárias, populares no Reino Unido; e as Americanas, predominantes nos esportes dos Estados Unidos.

4.1.1 Odds Decimais

São as mais utilizadas nos sites de apostas disponíveis para os brasileiros. Para calcular o retorno esperado em caso de uma aposta vencedora, basta multiplicar o valor da *odd* pelo valor apostado. Por exemplo, ao apostar R\$10,00 em uma *odd* de 1,15, o retorno esperado é:

$$R\$10,00 \times 1,15 = R\$11,50.$$

4.1.2 Odds Fracionárias

Muito comuns na Europa, a interpretação desse formato de odds é a seguinte: a odd $\frac{a}{b}$ indica que, ao apostar um valor b , o lucro esperado será a . Por exemplo, ao apostar

R\$10,00 em uma aposta esportiva cuja odd fracionária é $\frac{3}{5}$, o lucro será:

$$\frac{R\$10,00}{5} \times 3 = R\$6,00.$$

O retorno esperado, neste caso, será a soma do valor apostado com o lucro, resultando em R\$16,00.

4.1.3 Odds Americana

Formato de *odds* amplamente utilizado nas ligas americanas, como NBA e NFL, por exemplo. As odds americanas sempre vêm acompanhadas de um sinal “+” ou “-”. As odds precedidas pelo sinal “+” indicam que o time é o “azarão” e mostram que, apostando R\$100,00, o lucro será equivalente ao valor da odd. Já as odds precedidas pelo sinal “-” indicam que o time é o “favorito” e, ao apostar o valor da odd, o retorno será o valor apostado mais R\$100,00.

Por exemplo, suponhamos que um time A tenha uma odd americana representada por “+300”. Isso significa que, ao apostar R\$100,00, o lucro será de R\$300,00, totalizando um retorno de R\$400,00. Já um time B, com uma odd americana de “-150”, indica que é necessário apostar R\$150,00 para ganhar R\$100,00, totalizando um retorno de R\$250,00.

4.2 Lucro das Casas de Apostas

Uma pergunta que pode surgir é: se as apostas esportivas envolvem eventos que não podem ser prognosticados com certeza absoluta, como as casas de apostas lucram? A resposta está no chamado *overround*, que é a forma como as casas de apostas inserem comissões nas *odds*. Uma *odds* decimal é dada pela seguinte fórmula:

$$\text{odds} = \frac{1}{\text{probabilidade}}$$

Logo, se um time A tem 50% de chance de vitória, sua *odds* deveria ser:

$$\frac{1}{50\%} = 2$$

Ou seja, o retorno da aposta, em caso de vitória, é o valor apostado multiplicado por 2. Entretanto, se um jogador apostar R\$ 50,00 na vitória e R\$ 50,00 na derrota do time A, ele não irá ganhar nem perder (excluindo-se a hipótese de empate), já que, caso o time perca, como foi apostado R\$ 50,00 na derrota, o retorno esperado é:

$$\text{Retorno} = R\$50,00 \times 0 + R\$50,00 \times 2 = R\$100,00$$

Suponhamos que o time A e o time B disputem uma partida e ambos tenham igual probabilidade de vencer. Logo, as *odds* oferecidas pelas casas de apostas deveriam ser 2. Entretanto, geralmente esse valor é menor, ficando em torno de 1,90. Assim, se um apostador apostar R\$ 50,00 na vitória do time A e R\$ 50,00 na vitória do time B, o retorno esperado (novamente excluindo a possibilidade de empate) será:

$$\text{Retorno} = R\$50,00 \times 0 + R\$50,00 \times 1,90 = R\$95,00$$

Portanto, a casa de apostas lucra R\$ 5,00. Uma fórmula para o cálculo do *overround* é dada por:

$$\text{overround} = \frac{1}{\text{odd}_1} + \frac{1}{\text{odd}_2}$$

No nosso exemplo,

$$\text{overround} = \frac{1}{1,90} + \frac{1}{1,90} \approx 1,052$$

Note que a soma das probabilidades ultrapassa 100%, garantindo o lucro da casa de apostas a longo prazo.

4.3 Cálculo de Odds Usando a Distribuição de Poisson

Nesta seção, empregamos a distribuição de Poisson, expressa na Equação 3.22, para calcular as probabilidades de diversos placares entre duas equipes. O método consiste em obter, para cada time, a probabilidade de marcar i gols (time A) e j gols (time B) e, em seguida, tomar o produto dessas probabilidades. Em outras palavras, supomos que as variáveis aleatórias G_A e G_B , números totais de gols dos times A e B, respectivamente, sejam independentes; logo, um gol de uma equipe não interfere no desempenho da outra.

Na literatura, encontram-se diferentes maneiras de modelar o parâmetro λ . Um dos pioneiros a decompor λ “por time” para prever placares utilizando a distribuição de Poisson foi Maher (1982). Segundo o autor, a separação em parâmetros de ataque e defesa, distintos para jogos em casa e fora, mostrou-se adequada para explicar os resultados da First Division da English Football League (temporada 1971–1972). Maher aplicou, de forma empírica, o modelo de Poisson independente com o parâmetro estruturado da seguinte forma:

$$\begin{aligned} X_{ij} &\sim \text{Poisson}(\lambda_{ij}), & \lambda_{ij} &= \alpha_i \beta_j, \\ Y_{ij} &\sim \text{Poisson}(\mu_{ij}), & \mu_{ij} &= \gamma_j \delta_i, \end{aligned}$$

sujeito às restrições de identificabilidade

$$\sum_{i=1}^n \alpha_i = \sum_{i=1}^n \beta_i \quad \text{e} \quad \sum_{i=1}^n \gamma_i = \sum_{i=1}^n \delta_i.$$

- X_{ij} – número de gols marcados pelo *mandante* i contra j .
- Y_{ij} – número de gols marcados pelo *visitante* j contra i .
- α_i – **força de ataque em casa** do time i .
- β_j – **fraqueza defensiva fora** do time j .
- γ_j – **fraqueza defensiva em casa** do time j .
- δ_i – **força de ataque fora** do time i .
- λ_{ij} – média de gols esperados do mandante i contra j .
- μ_{ij} – média de gols esperados do visitante j contra i .

Segundo Maher (1982), há três razões fundamentais que tornam a distribuição de Poisson adequada para modelar o número de gols em partidas de futebol:

- a) **Eventos raros e independentes:** cada chegada à área adversária com potencial de gol é relativamente rara em relação ao total de posses de bola, e pode-se supor razoavelmente que o sucesso (gol) em cada tentativa seja independente das demais.
- b) **Grande número de oportunidades:** durante uma partida, as equipes realizam muitas ações ofensivas, o que configura um número elevado de “tentativas” de marcar, aproximando o processo binomial à Poisson quando a probabilidade de gol em cada tentativa é pequena.
- c) **Homogeneidade temporal:** supõe-se que a taxa de ocorrência de gols se mantém aproximadamente constante ao longo dos 90 minutos, permitindo modelar o total de gols como um único disparo de Poisson com parâmetro igual à média de gols esperados.

Em consequência, sejam X_{ij} e Y_{ij} o número de gols marcados pelo mandante i e pelo visitante j , respectivamente, na partida (i, j) . Se a taxa de gols esperada for λ_{ij} e μ_{ij} , então:

$$X_{ij} \sim \text{Poisson}(\lambda_{ij}), \quad Y_{ij} \sim \text{Poisson}(\mu_{ij}),$$

o que fornece uma base teórica consistente para a decomposição de λ e μ em parâmetros de ataque e defesa conforme apresentado no modelo de Maher.

4.3.1 Campeonato Paulista 2025

Vamos determinar, utilizando a distribuição de Poisson, as probabilidades de placares entre *Corinthians* e *Palmeiras* no jogo de volta do **Campeonato Paulista**. No formato atual, o **Campeonato Paulista** é disputado por 16 times, divididos em quatro grupos de quatro equipes cada. Na primeira fase, os clubes de um grupo enfrentam apenas os adversários dos outros grupos, totalizando uma rodada de 12 jogos. Os dois melhores colocados de cada grupo avançam às *quartas-de-final*, disputadas em partida única, com mando da equipe de melhor campanha. A semifinal segue a mesma lógica (jogo único, mando do melhor colocado). A decisão é realizada em dois confrontos (ida e volta), e a pontuação da fase de classificação define quem decide em casa.

A Tabela 4.1 apresenta a campanha realizada pelo Corinthians no Campeonato Paulista de 2025; juntamente com o Palmeiras, o clube chegou à final da competição.

Tabela 4.1. Jogos do Corinthians no Campeonato Paulista de 2025

Data	Adversário	Local	Resultado	Fase
16/01/2025	Red Bull Bragantino	Fora	1–2	Fase de Grupos
19/01/2025	Velo Clube	Casa	2–1	Fase de Grupos
22/01/2025	Água Santa	Casa	2–1	Fase de Grupos
26/01/2025	São Paulo	Fora	3–1	Fase de Grupos
29/01/2025	Ponte Preta	Fora	0–1	Fase de Grupos
01/02/2025	Noroeste	Casa	2–1	Fase de Grupos
03/02/2025	Novorizontino	Fora	0–1	Fase de Grupos
06/02/2025	Palmeiras	Fora	1–1	Fase de Grupos
09/02/2025	São Bernardo	Casa	2–0	Fase de Grupos
12/02/2025	Santos	Casa	2–1	Fase de Grupos
15/02/2025	Portuguesa	Fora	2–2	Fase de Grupos
23/02/2025	Guarani	Casa	2–2	Fase de Grupos
02/03/2025	Mirassol	Casa	2–0	Quartas de Final
09/03/2025	Santos	Casa	2–1	Semifinal
16/03/2025	Palmeiras	Fora	0–1	Final (Jogo de Ida)

Fonte: Globo Esporte (2025)

A Tabela 4.2 representa, por sua vez, a trajetória do Palmeiras na mesma competição.

Tabela 4.2. Jogos do Palmeiras no Campeonato Paulista de 2025

Data	Adversário	Local	Resultado	Fase
15/01/2025	Portuguesa	Casa	2–0	Fase de Grupos
18/01/2025	Noroeste	Fora	1–1	Fase de Grupos
22/01/2025	Santos	Fora	1–2	Fase de Grupos
25/01/2025	Novorizontino	Casa	1–2	Fase de Grupos
28/01/2025	Red Bull Bragantino	Casa	0–0	Fase de Grupos
02/02/2025	Guarani	Fora	1–4	Fase de Grupos
06/02/2025	Corinthians	Casa	1–1	Fase de Grupos
13/02/2025	Inter de Limeira	Fora	0–3	Fase de Grupos
16/02/2025	São Paulo	Casa	0–0	Fase de Grupos
20/02/2025	Botafogo-SP	Casa	3–1	Fase de Grupos
23/02/2025	Mirassol	Fora	2–3	Fase de Grupos
27/02/2025	Água Santa	Fora	1–1	Fase de Grupos
01/03/2025	São Bernardo	Fora	0–3	Quartas de Final
10/03/2025	São Paulo	Casa	1–0	Semifinal
16/03/2025	Corinthians	Casa	0–1	Final (Jogo de Ida)

Fonte: Globo Esporte (2025)

A coluna **Local** indica a condição do time da tabela: **Casa** (mandante) e **Fora** (visitante). A coluna **Resultado** usa o formato *mandante–visitante*; logo, quando o Corinthians estiver como *Fora*, o primeiro número é o do adversário e o segundo é o do Corinthians (e vice-versa quando *Casa*). Essa convenção ajuda a evidenciar o mando de campo em análises, como na estimação da vantagem de casa em modelos de Poisson.

Vamos utilizar o programa R, amplamente adotado por estatísticos. O R emprega uma linguagem semelhante à do Python, permite a escrita de diversos *scripts* voltados à Estatística e à Probabilidade, além da manipulação de matrizes. Neste contexto, cada entrada p_{ij} de uma matriz representa:

$$p_{ij} = P\{G_1 = i, G_2 = j\},$$

Isto é, a probabilidade de a equipe mandante marcar i gols e a equipe visitante marcar j gols. O script em R encontra-se no apêndice

A média de gols esperados para cada equipe na partida — representada por λ — foi estimada como o produto da média de gols marcados por jogo da própria equipe (força ofensiva) pela média de gols sofridos por jogo da equipe adversária (fragilidade defensiva). Assim, temos:

$$\lambda_X = A_{\text{mandante}} \times D_{\text{visitante}},$$

$$\lambda_Y = A_{\text{visitante}} \times D_{\text{mandante}},$$

em que A representa a taxa média de ataque (gols marcados por jogo) e D , a taxa média de defesa (gols sofridos por jogo). Esses parâmetros foram calculados a partir dos dados reais do Campeonato Paulista de 2025.

Assumindo independência entre os gols marcados pelas duas equipes — hipótese simplificadora comum nesse tipo de modelagem —, a probabilidade de um placar específico (i, j) é dada pelo produto de duas distribuições de Poisson:

$$P(G_1 = i, G_2 = j) = \frac{e^{-\lambda_X} \lambda_X^i}{i!} \cdot \frac{e^{-\lambda_Y} \lambda_Y^j}{j!}.$$

A partir dessas distribuições, é possível construir uma matriz de probabilidades dos placares possíveis para a partida. O Apêndice A.1 detalha o código, em linguagem R, empregado na estimativa dessas probabilidades para o jogo de volta da final entre Corinthians e Palmeiras.

A matriz de probabilidades gerada pelo código apresentado no apêndice A.1, com os valores arredondados e restrita aos placares de 0 a 7 gols para cada equipe, está representada na Tabela 4.3.

Tabela 4.3. Matriz de probabilidades (%)

Gols	0	1	2	3	4	5	6	7
0	5,9	9,4	7,5	4,0	1,6	0,5	0,1	0,0
1	7,2	11,6	9,3	4,9	2,0	0,6	0,2	0,0
2	4,5	7,2	5,7	3,1	1,2	0,4	0,1	0,0
3	1,9	3,0	2,4	1,3	0,5	0,2	0,0	0,0
4	0,6	0,9	0,7	0,4	0,2	0,1	0,0	0,0
5	0,1	0,2	0,2	0,1	0,0	0,0	0,0	0,0
6	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
7	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0

Fonte: Elaborado pelo autor (2025).

Nesta matriz, as linhas indicam o número de gols marcados pelo **Corinthians** (mandante) e as colunas o número de gols do **Palmeiras** (visitante). Cada célula apresenta a probabilidade percentual estimada para o respectivo placar.

Por exemplo, a célula na **linha 3** e **coluna 1** corresponde ao placar **2 x 0** a favor do Corinthians, com uma probabilidade de **4,5%**, conforme calculado pelo modelo. uma observação interessante a se fazer, é a seguinte:

- Vitória do Corinthians = soma das células abaixo da diagonal principal.
- Empate = soma da diagonal principal.
- Vitória do Palmeiras = soma das células acima da diagonal principal.

Abaixo, apresentamos a cotação de algumas casas de aposta para a final do Campeonato Paulista, realizada em 27 de março, O placar saiu 0 – 0, com o Corinthians jogando em casa. A Figura 4.1 exibe valores bastante próximos aos obtidos pelo código no Apêndice A.1 e apresentados na Tabela 4.4.



Figura 4.1. Odds oferecidas por diferentes casas de aposta para a partida final do Paulistão.

Fonte: Oddspedia (2025) .

A probabilidade e *odd* que o script do apêndice A.1 nos retorna está na Tabela 4.4. Observe que os valores são bem próximos daqueles que encontramos na Figura 4.1

Tabela 4.4. Percentual de resultados e odds por categoria

Categoria	Percentual (%)	Odds
Corinthians	30,0	3,33
Empate	25,0	4,00
Palmeiras	45,1	2,21

Fonte: Elaborado pelo autor (2025).

4.3.2 Brasileiro Série A

Para determinar os placares da partida entre Palmeiras e Flamengo-RJ, aplicamos o produto de duas distribuições de Poisson. Os dados históricos dos jogos foram obtidos no portal Football Data (2025) , que oferece planilhas em formato CSV e Excel com informações detalhadas de partidas — incluindo datas, resultados, odds, probabilidades de totais de gols, handicap (método de equilibrar as chances entre equipes de níveis diferentes, atribuindo uma “vantagem” ou “desvantagem” inicial em gols) e outros indicadores — das principais competições de futebol desde de o ano de 2012.

No atual formato da Série A do Campeonato Brasileiro, o título é concedido ao clube que somar mais pontos ao término das 38 rodadas. Os 20 clubes disputam o torneio em sistema de pontos corridos, enfrentando-se em turno e retorno. Cada equipe realiza **38** partidas (19 como mandante e 19 como visitante), totalizando **380** jogos na competição. A pontuação é atribuída da seguinte forma:

- i) 3 pontos por vitória;
- ii) 1 ponto por empate;
- iii) 0 ponto por derrota.

Utiliza-se o código apresentado no Apêndice B.1 para a análise das probabilidades e das *odds*, bem como para a construção da matriz de probabilidade dos placares entre Palmeiras e Flamengo – RJ. Para isso, foram coletadas partidas realizadas entre 06 de agosto de 2023 e 05 de agosto de 2025. O cálculo da média esperada (λ) foi realizado da seguinte forma:

- **Para o Flamengo (mandante):** calcula-se a média de gols que se espera que o Flamengo marque. A fórmula utilizada é:

$$\lambda_{\text{Flamengo}} = \text{Vantagem de Casa} \times \text{Ataque do Flamengo} \times \text{Defesa do Palmeiras}$$

- **Para o Palmeiras (visitante):** calcula-se a média de gols que se espera que o Palmeiras marque. A fórmula utilizada é:

$$\lambda_{\text{Palmeiras}} = \text{Ataque do Palmeiras} \times \text{Defesa do Flamengo}$$

O código no Apêndice B.1 emprega o mesmo procedimento empregado por Maher (1982), a seguir tem-se a Tabela 4.5 da matriz de probaildades:

Tabela 4.5. Matriz de probabilidades arredondada (%) – Flamengo × Palmeiras

Gols	0	1	2	3	4	5	6	7
0	10,2	10,1	5,0	1,7	0,4	0,1	0,0	0,0
1	13,1	13,1	6,5	2,2	0,5	0,1	0,0	0,0
2	8,5	8,4	4,2	1,4	0,3	0,1	0,0	0,0
3	3,6	3,6	1,8	0,6	0,1	0,0	0,0	0,0
4	1,2	1,2	0,6	0,2	0,0	0,0	0,0	0,0
5	0,3	0,3	0,1	0,1	0,0	0,0	0,0	0,0
6	0,1	0,1	0,0	0,0	0,0	0,0	0,0	0,0
7	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0

Fonte: Elaborado pelo autor (2025).

Nesta matriz, as **linhas** indicam o número de gols do **Flamengo** (mandante) e as **colunas** o número de gols do **Palmeiras** (visitante). Cada célula apresenta a **probabilidade percentual** estimada para o respectivo placar, arredondada a uma casa decimal.

A Tabela 4.6 trás as *odds* e probabilidades do confronto entre Palmeiras e flamengo-RJ. um confronto entre esses times, pelo calendário do cameponato brasileiro Série A está previsto para o dia 5 de novembro de 2025 na 29^o rodadda. Eles já se enfrentaram na 10^a rodada, em 25 de maio de 2025, no Allianz Parque, com vitória do Flamengo por 2–0.

Tabela 4.6. Percentual de resultados e odds por categoria

Categoria	Percentual (%)	Odds
Flamengo	42,5	2,35
Empate	28,3	3,53
Palmeiras	29,1	3,43

Fonte: Elaborado pelo autor (2025).

ENSINO DE PROBABILIDADE COM FOCO EM APOSTAS ESPORTIVAS

Com o crescimento acelerado do mercado de apostas no Brasil, surgem desafios que o ambiente escolar não pode ignorar. A Base Nacional Comum Curricular (BNCC) determina que o currículo dialogue com as vivências cotidianas dos estudantes, levando em conta seus contextos locais e as especificidades do país, princípio igualmente reforçado pela Lei de Diretrizes e Bases (LDB). Em seu art. 26, essa lei estabelece:

Os currículos da educação infantil, do ensino fundamental e do ensino médio devem ter base nacional comum, a ser complementada, em cada sistema de ensino e em cada estabelecimento escolar, por uma parte diversificada, exigida pelas características regionais e locais da sociedade, da cultura, da economia e dos educandos. (BRASIL, 1996, art. 26)

Juntos, BNCC e LDB asseguram uma formação que integra dimensões econômicas, culturais e regionais, promovendo uma aprendizagem crítica e significativa.

Nesse cenário, as apostas esportivas — cada vez mais presentes no dia a dia dos jovens, impulsionadas por intensa publicidade — tornam-se tema relevante para a Matemática. A disciplina pode problematizar, de forma crítica, o conceito de *odds*, explicar como eles são calculados, esclarecer a lógica dos lucros das casas de aposta e, sobretudo, discutir os riscos envolvidos. Assim, o estudo ultrapassa a teoria, aproximando-se da realidade dos alunos e fomentando uma postura consciente diante desse fenômeno.

Dessa forma, o objetivo deste capítulo é elaborar uma sequência didática para estimar os placares de uma partida de futebol a partir dos resultados de jogos anteriores, utilizando a distribuição de Poisson. Segundo Zabala (1998, p. 18), “uma sequência didática é um conjunto ordenado, articulado e estruturado de atividades de sala de aula, com começo, meio e fim, que visa à realização de determinados objetivos de aprendizagem”. Desse modo, o roteiro proposto organiza conteúdos, atividades e avaliações de forma encadeada, garantindo coerência pedagógica ao longo do trabalho com modelos probabilísticos aplicados ao futebol. A seguir, apresentamos um breve roteiro para utilização do GeoGebra Calculator Suite.

5.1 GeoGebra: Cálculo da Distribuição de Poisson

Este guia rápido detalha como utilizar a ferramenta Calculadora de Probabilidade, disponível no *GeoGebra Calculator Suite* (INTERNATIONAL GEOGEBRA INSTITUTE, 2025), para analisar a Distribuição de Poisson.

5.1.1 Passo 1: Acesso à Calculadora de Probabilidade

1. Abra o GeoGebra (versão Clássica ou *Calculator Suite*).
2. No menu superior, clique em **Exibir**.
3. Selecione a opção **Calculadora de Probabilidade**. Uma nova janela será aberta, apresentando uma distribuição (geralmente a Normal como padrão) e seus respectivos controles.

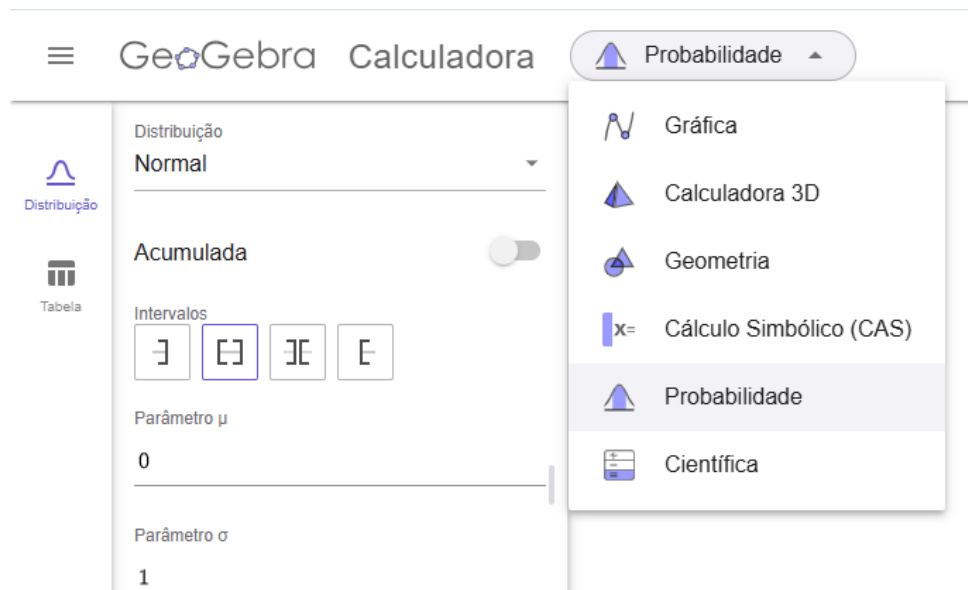


Figura 5.1. Acesso à Calculadora de Probabilidade no menu Exibir do GeoGebra.

5.1.2 Passo 2: Seleção da Distribuição de Poisson

1. Na janela da Calculadora de Probabilidade, localize o menu suspenso (dropdown) que indica o tipo de distribuição.
2. Clique neste menu e, na lista de distribuições discretas, selecione a opção **Poisson**.

A interface será atualizada para refletir os parâmetros e o gráfico da Distribuição de Poisson.

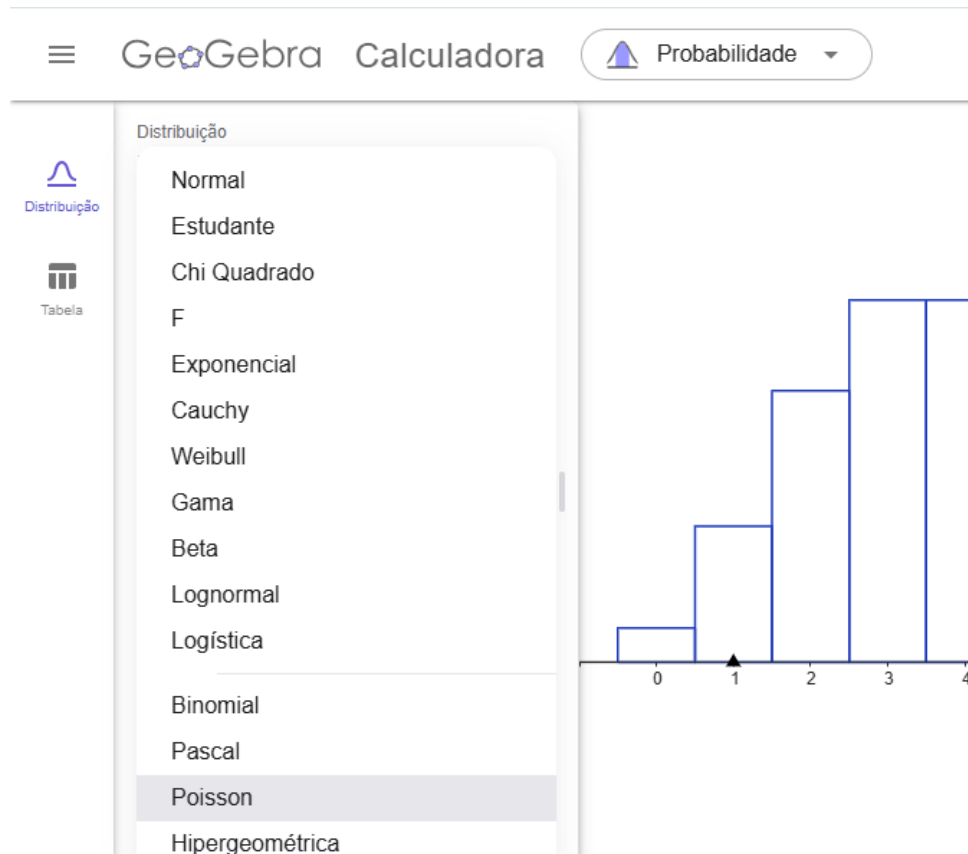


Figura 5.2. Seleção da distribuição de Poisson no menu suspenso.

5.1.3 Passo 3: Inserção de Parâmetros e Cálculo de Probabilidades

A interface para a Distribuição de Poisson solicitará o seguinte parâmetro:

- **Média (μ):** Neste campo, insira o valor de λ (a média de gols, no contexto do futebol).

Abaixo do campo da média, você encontrará os controles para calcular as probabilidades:

- **Intervalo de Probabilidade:** A calculadora permite o cálculo de probabilidades para um valor exato ou um intervalo.
 - **Probabilidade Pontual ($P(X = k)$):** Para calcular a probabilidade de um time marcar exatamente k gols, insira o valor de k nos dois campos do intervalo. Por exemplo, para $P(X = 2)$, defina o intervalo como $[2 , 2]$.
 - **Probabilidade Acumulada à Esquerda ($P(X \leq k)$):** Para calcular a probabilidade de marcar *até* k gols, utilize o botão de intervalo fechado à esquerda $] e insira k no campo correspondente. O GeoGebra calculará a soma das probabilidades de $0, 1, \dots, k$.$
 - **Probabilidade Acumulada à Direita ($P(X \geq k)$):** Para calcular a probabilidade de marcar *pelo menos* k gols, utilize o botão de intervalo fechado à direita $[e insira k no campo.$

- **Probabilidade em um Intervalo** ($P(k_1 \leq X \leq k_2)$): Para calcular a probabilidade de o número de gols estar entre k_1 e k_2 (inclusive), basta preencher os campos do intervalo.

O resultado numérico da probabilidade é exibido instantaneamente, e o gráfico de barras é atualizado, destacando os valores de k selecionados.

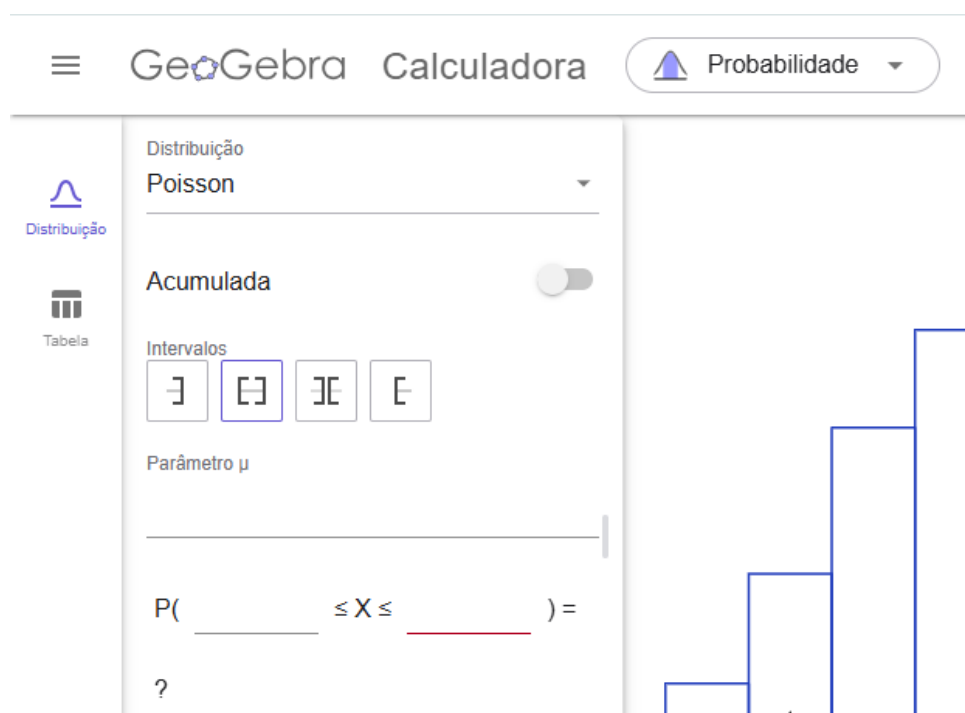


Figura 5.3. Interface para a Distribuição de Poisson.

5.2 Público-alvo

A sequência didática desenvolvida neste trabalho é voltada principalmente para estudantes do 2º e 3º anos do Ensino Médio, com idade entre 16 e 18 anos, tanto da rede pública quanto privada. Espera-se que esses alunos já tenham familiaridade com operações algébricas básicas, porcentagem, leitura de tabelas e gráficos, bem como noções iniciais de probabilidade.

O material também se alinha aos objetivos do Novo Ensino Médio, em especial aos itinerários formativos da área de Matemática e suas Tecnologias, nos quais a Base Nacional Comum Curricular (BNCC) incentiva o aprofundamento em temas como Modelagem Matemática, Estatística e Probabilidade. Além disso, é pensado para jovens inseridos em um contexto social marcado pela presença constante do futebol no cotidiano e pela crescente visibilidade das apostas esportivas. Nesse cenário, trabalhar conceitos como *odds*, risco e retorno pode contribuir para a construção de uma postura crítica e consciente diante das promessas de lucro fácil amplamente divulgadas.

Por fim, esta proposta pode ser útil a professores de Matemática que buscam alternativas para abordar conteúdos estatísticos de forma contextualizada, conforme orienta Brasil (2018) e também a Brasil (1996), ao preverem a articulação entre os conhecimentos escolares e a realidade vivida pelos estudantes.

5.3 Objetivo

O objetivo é estimar placares de partidas de futebol a partir de resultados históricos, empregando a distribuição de Poisson – abordagem iniciada por Moroney (1951), formalizada por Maher (1982) e posteriormente refinada por Dixon e Coles (1997), e ainda hoje empregada na previsão de placares. A sequência organizará conteúdos, atividades e avaliações de modo a garantir coerência pedagógica ao estudo de modelos probabilísticos aplicados ao futebol.

5.4 Habilidades e Competências

A Tabela 5.1 sintetiza como as competências e habilidades da Base Nacional Comum Curricular (BNCC, (2018)), dos Parâmetros Curriculares Nacionais do Ensino Médio (PCNEM, (1999)) e das Competências Gerais da BNCC são aplicadas na sequência didática proposta.

Tabela 5.1. Enquadramento curricular: documentos, competências/habilidades e aplicação na sequência didática

Documento	Competências / Habilidades	Aplicação na sequência
BNCC – Ensino Médio	EM13MAT311 – Identificar e descrever o espaço amostral de eventos aleatórios. EM13MAT312 – Formular e resolver problemas de probabilidade em experimentos sucessivos. EM13MAT102 – Analisar e interpretar representações estatísticas de forma crítica.	Definição do espaço amostral de placares; cálculo de λ pela distribuição de Poisson; interpretação de gráficos de frequência.
PCNEM (1999)	Abordar probabilidade e estatística em contextos de risco e tomada de decisão.	Discussão sobre apostas esportivas responsáveis; análise crítica do uso dos modelos matemáticos.
Competências Gerais da BNCC	1 – Linguagens 2 – Pensamento científico, crítico e criativo 5 – Cultura digital 7 – Argumentação	Uso de calculadora e planilhas; apresentação oral e escrita dos resultados; debates sobre limitações do modelo.

Fonte: Autor (2025).

5.5 Roteiro para a Sequência Didática

O roteiro proposto para a sequência didática é:

- I. Introdução à probabilidade associada às apostas esportivas, discussão inicial e levantamento de conhecimentos prévios.
- II. Revisão de conceitos de espaço amostral e probabilidade básica, inserindo a ideia de *odds* e *overround* como formas de garantir o lucro das casas de apostas.
- III. Introdução à distribuição de Poisson de forma intuitiva, preferencialmente como uma construção a partir da distribuição binomial.
- IV. Coleta de dados de partidas anteriores em sites como Football Data (2025), que disponibiliza uma tabela em formato XLS com todas as partidas do Campeonato Brasileiro Série A desde 2012, ou em Confederação Brasileira de Futebol (2025).
- V. Cálculo das médias a partir das frequências para estimar o parâmetro λ utilizado na distribuição de Poisson, dada pela fórmula 3.22.
- VI. Análise dos resultados obtidos e discussão crítica do modelo, enfatizando a conscientização de que probabilidades não representam certezas.

5.6 Planejamento das Aulas

Este roteiro pode ser dividido em cinco aulas, conforme descritas a seguir:

5.6.1 Aula 1 – Probabilidade no Universo das Apostas Esportivas

Duração: 1 aula de 45 minutos

Objetivo Geral

Contextualizar o tema das apostas esportivas e despertar conhecimentos prévios de probabilidade, preparando o terreno para as aulas subsequentes.

Objetivos Específicos

1. Reconhecer a presença das apostas esportivas no cotidiano da turma e problematizar a influência da publicidade;
2. Definir, em linguagem acessível, o conceito preliminar de *odds* e sua relação com risco e retorno;
3. Mapear rapidamente — via conversa guiada — os conhecimentos prévios sobre espaço amostral e eventos aleatórios.

Habilidades BNCC Mobilizadas

EM13MAT311, EM13MAT312, EM13MAT102 BRASIL (2018a)

Conteúdo Abordado

- Definição de *odds*;
- Princípios básicos de probabilidade;
- Estimativas rápidas de porcentagem.

Recursos Didáticos

- Slides ou quadro digital com manchetes reais sobre *bets*;
- Quadro branco e acessórios (apagador, pincel);
- Calculadora simples (ou celular) para estimativas rápidas de porcentagem.

Metodologia

1. **Pergunta disparadora (5 min)** — “Se você apostar R\$10 no seu time favorito a *odds* 1,80, quanto recebe se ele vencer?” Alunos respondem oralmente; professor registra variações.
2. **Exposição dialogada (10 min)** — Apresentar dados do crescimento do setor de apostas no Brasil e projeções (CNN BRASIL, 2024). Conectar ao cotidiano dos estudantes. Explorar o aumento do número de casas de apostas e discutir potenciais benefícios e prejuízos. Utilizar o gráfico da Figura 5.4.

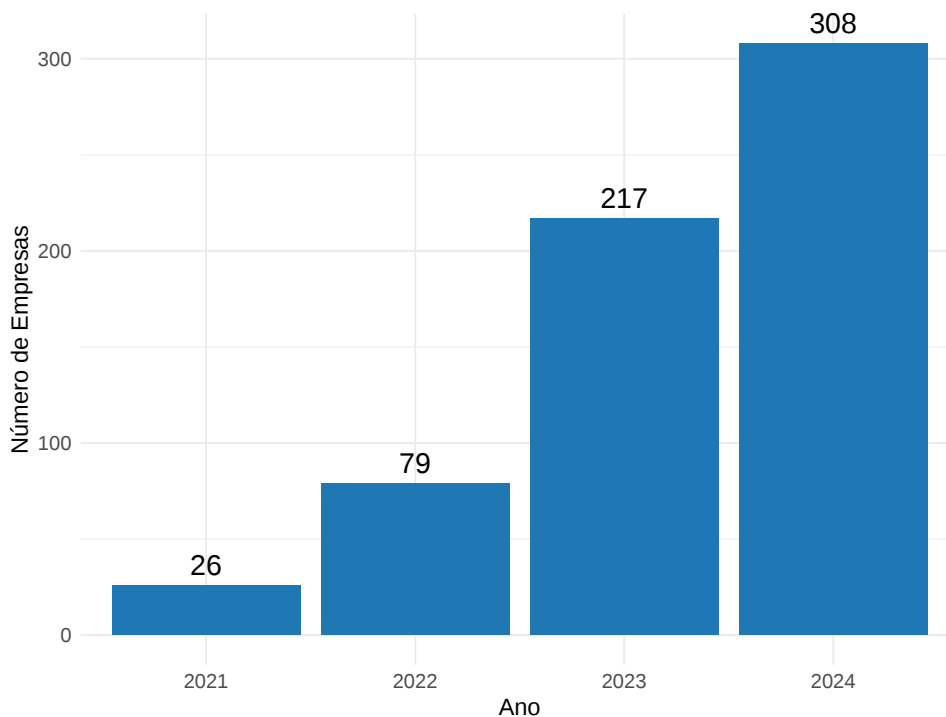


Figura 5.4. Empresas de apostas registradas no Brasil ao longo do tempo.

Fonte: CNN Brasil (2024).

3. **Questões a serem resolvidas em sala (25 min)** — Resolução em duplas/trios dos itens 1–9 do plano (com socialização rápida no quadro).
4. **Tarefa de casa (5 min)** — Observar ao menos três propagandas de apostas durante a semana, anotando: jogo citado, *odd* anunciada e promessa de ganho. Trazer para a próxima aula.

Lista de Questões (em sala)

1. Em um certo clássico entre o Time A e o Time B, foram disputadas 10 partidas nos últimos dois anos. O Time A venceu 4, empatou 3 e perdeu 3.
 - a) Qual a probabilidade de o Time A vencer uma partida escolhida ao acaso?
 - b) Qual a probabilidade de empate?
 - c) Qual a probabilidade de derrota do Time A?
2. Num sorteio aleatório de 3 partidas do mesmo histórico (acima), qual é a probabilidade de que nenhuma termine em empate?
3. Em um campeonato, sabe-se que a taxa de conversão de pênaltis é, em média, 75%.
 - a) Qual a probabilidade de um jogador converter exatamente 2 pênaltis em 3 cobranças?
 - b) Qual a probabilidade de converter pelo menos 2 em 3?
4. Durante a temporada, o atacante X finalizou 12 vezes ao gol e acertou 8 delas.
 - a) Qual a probabilidade empírica de X acertar uma finalização?
 - b) Em 5 novas finalizações, qual o número esperado de acertos?
5. Em um campeonato com 20 equipes, apenas 4 se classificam para a fase final.
 - a) Se todas as equipes tivessem desempenho igual, qual seria a probabilidade de uma equipe qualquer ficar entre as quatro primeiras?
6. Num sorteio para escolher o capitão de campo, participa todo o elenco de 11 jogadores.
 - a) Qual a probabilidade de o goleiro ser escolhido?
 - b) Qual a probabilidade de um dos três atacantes ser escolhido?
7. Sabendo que, em média, um time visitante marca 1 gol por partida, defina, de modo qualitativo, o que significaria para esse time:
 - a) “Marcar pelo menos 2 gols”;
 - b) “Não marcar nenhum gol”.
8. Em uma rodada, seleciona-se aleatoriamente uma das 380 partidas do Brasileirão Série A. Sabe-se que, historicamente, cerca de 30% dos jogos terminam em empate.
 - a) Qual a probabilidade de a partida sorteada ter vencedor definido (não empate)?

- b) Em 5 sorteios independentes, qual a probabilidade de ocorrer exatamente 2 empates?
9. Um exame antidoping tem eficiência de 95% para detectar substâncias proibidas quando um atleta está de fato dopado. Entretanto, o teste apresenta resultado falso-positivo para 1% dos atletas testados. Se apenas 0,5% dos atletas utilizam substâncias proibidas, qual é a probabilidade de um atleta estar dopado, dado que o resultado do seu exame foi positivo? (Questão adaptada de César e Morgado (2022).)

Avaliação Formativa

- Participação nas discussões e coerência das probabilidades atribuídas;
- Registro individual (breve parágrafo) resumindo o que mudou na percepção de “probabilidade” após a aula.

Nota ao professor: Esta aula cumpre a etapa I do roteiro apresentado no item 5.5 e inaugura o *planejamento de cinco aulas* descrito em 5.6.

5.6.2 Aula 2 – Como as Casas de Aposta Transformam *Odds* em Lucro?

Duração: 1 aula de 45 minutos

Objetivo Geral

Conceituar *odds* e ensinar aos alunos como converter *odds* em probabilidades, evidenciando os *overrounds* como vantagem da casa de apostas para aumentar sua margem de lucro.

Objetivos Específicos

- a) Transformar *odds* em porcentagem usando a regra $p = \frac{100}{\text{odd}}$;
- b) Verificar se a soma das probabilidades ultrapassa 100%;
- c) Calcular o valor que excede esse limite de 100%.

Materiais

- Tabela impressa com *odds* de três (ou mais) jogos do campeonato analisado;
- Calculadora ou celular;
- Data-show para socializar os resultados.

Metodologia

1. **Acolhida e retomada (3 min)** — Relembrar a noção de *odds* e a tarefa de casa sobre propagandas. Registrar 1–2 exemplos trazidos pela turma.
2. **Apresentação do problema (Figura) (6 min)** — Observar a Figura 5.5 com *odds*. Explicar que será feita a “tradução” das *odds* em probabilidades e a análise do lucro implícito da casa.



Al Ahly 3 – 1 Al-Ittihad		Resultado final 3:1 (1:0, 2:1)		
Sexta, 15 Dec 2023, 15:00				
1X2 Acima/Abaixo Handicap Asiático Ambos marcam Para classificar				
Tempo Total 1º Tempo 2º Tempo				
Casas de apostas	1 ↑	X ↑	2 ↑	
bet365 bet365	▲ 5.25	▼ 3.75	▼ 1.67	
Betano Betano.br	▲ 4.45	3.55	▼ 1.78	
betfair Betfair	▲ 5.00	▲ 3.60	▼ 1.70	

Figura 5.5. Odds do jogo entre Al Ahly (Egito) e Al-Ittihad (Arábia Saudita).

Fonte: Odds Agora (2023).

3. **Atividade guiada: conversão de odds (12 min)** — Em duplas, aplicar $p = \frac{100}{\text{odd}}$ para os desfechos do jogo e registrar as probabilidades.
4. **Discussão do somatório e overround (8 min)** — Socializar resultados e discutir por que a soma ultrapassa 100%. Introduzir o termo *overround*.
5. **Análise do resultado real (6 min)** — Checar se o placar final era ou não o mais provável segundo as *odds*. Discutir risco e “zebra”.
6. **Consolidação conceitual (6 min)** — Sistematizar no quadro: (i) fórmula de conversão; (ii) soma das probabilidades; (iii) $\text{overround} = (\text{soma}) - 100\%$; (iv) interpretação econômica para a banca.
7. **Registro individual e fechamento (4 min)** — Em 3–5 linhas: “O que é *overround* e como ele gera lucro para a casa?”

Atividade (Questões em Sala)

Com base na Figura 5.5, proponha:

1. Se você apostar R\$10,00, qual será o retorno financeiro em caso de vitória do mandante, empate ou vitória do visitante?
2. Converta as *odds* em probabilidades usando $p = \frac{100}{\text{odd}}$.
3. Some as probabilidades obtidas. Elas totalizam 100%? Se não, por que isso ocorre? Qual a implicação do “excesso” para quem aposta?

4. O resultado do jogo foi 3–1 para o Al Ahly. Explique se esse resultado era o mais provável segundo as *odds* (lembre-se: *odds* são inverso da probabilidade).

Atividade para Casa

Coletar dados de pelo menos 10 partidas de dois times em um campeonato e organizar em tabelas como a Tabela 5.2.

Tabela 5.2. Exemplo de construção de tabela de placares

Data	Adversário	Local	Resultado	Fase
dd/mm/aaaa	Time A	Mandante ou Visitante	x–y	Fase de Grupos
dd/mm/aaaa	Time B	Mandante ou Visitante	x–y	Fase de Grupos
⋮	⋮	⋮	⋮	⋮
dd/mm/aaaa	Time N	Mandante ou Visitante	x–y	Quartas de Final
dd/mm/aaaa	Time O	Mandante ou Visitante	x–y	Semifinal
dd/mm/aaaa	Time P	Mandante ou Visitante	x–y	Final

Fonte: Dados fictícios para uso didático.

Avaliação formativa

- **Conversão de *odds*:** aplicação correta de $p = \frac{100}{\text{odd}}$ e interpretação percentual;
- **Análise do somatório:** reconhecimento de que a soma excede 100% e explicação do *overround*;
- **Participação oral:** clareza e fundamentação das intervenções;
- **Registro individual:** síntese sobre o papel do *overround* no lucro da casa;
- **Retomada da tarefa anterior:** entrega e qualidade dos exemplos de propagandas (quando aplicável).

5.6.3 Aula 3 – O Modelo Probabilístico de Poisson a Partir do Modelo Binomial

Duração: 45 min

Objetivo Geral

Construir, de forma intuitiva, a distribuição de Poisson a partir da distribuição binomial, contextualizando-a na previsão de gols em partidas de futebol.

Objetivos Específicos

- Revisar os elementos da distribuição binomial (n , p , variável discreta);
- Identificar limitações da binomial quando há muitos ensaios e baixa probabilidade de sucesso;

- c) Compreender a Poisson como modelo para eventos raros em intervalos fixos, com parâmetro igual à média observada;
- d) Aplicar a fórmula da Poisson para calcular $P(X = k)$.

Materiais

- Projetor;
- Quadro branco e acessórios;
- Calculadora simples ou celulares.

Metodologia

1. **Ativação do conhecimento prévio (5 min)** — Revisão do conceito de probabilidade e do contexto de apostas.
2. **Distribuição binomial: situação-problema (15 min)** — Apresentar a definição a seguir e resolver o exemplo sobre pênaltis, com variações “todas as 5” e “pelo menos 4”.

Definição 5.1 (Distribuição Binomial). Considere n repetições independentes de uma mesma ação, cada uma com sucesso (probabilidade $p \in [0, 1]$) ou fracasso (probabilidade $1 - p$). A probabilidade de obter exatamente k sucessos é

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k = 0, 1, \dots, n.$$

Exemplo 5.1. Um jogador converte pênaltis com 75% de acerto. Em 5 cobranças, qual a probabilidade de converter exatamente 4?

3. **Limites da binomial e ponte para Poisson (10 min)** — Discutir a dificuldade quando n é grande e p é pequeno. Enunciar a aproximação:

Proposição 5.1 (Aproximação da Binomial pela Poisson). *Seja $X \sim \text{Bin}(n, p_n)$ com $\lambda = np_n$ constante. Então, para $k = 0, 1, 2, \dots$,*

$$\lim_{n \rightarrow \infty} \binom{n}{k} p_n^k (1 - p_n)^{n-k} = \frac{e^{-\lambda} \lambda^k}{k!} = P_{\text{Pois}}(X = k).$$

Esboço. Usando $p_n = \lambda/n$ e $\lim_{n \rightarrow \infty} (1 - \lambda/n)^n = e^{-\lambda}$, obtém-se o resultado. $\square$

$$P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

4. **Aplicação com média de gols (10 min)** — Exemplo: $\lambda = 1,2$. Calcular $P(X = 2)$ e interpretar (previsão de placar).
5. **Fechamento (5 min)** — Síntese comparativa: binomial (ensaios fixos, p constante) vs. Poisson (contagem de eventos por intervalo). Registro individual: “Quando usar Poisson em vez de binomial?”

Avaliação Formativa

- **Participação e resolução coletiva:** acompanhar o raciocínio no problema dos pênaltis;
- **Cálculo orientado:** aplicar a Poisson no exemplo com $\lambda = 1,2$;
- **Síntese individual curta:** responder, em até 4 linhas, à pergunta sobre quando usar cada modelo.

5.6.4 Aula 4 – Uso da Calculadora de Probabilidades do GeoGebra (Distribuição de Poisson)

Duração: 45 min

Objetivo Geral

Familiarizar os alunos com a *Calculadora de Probabilidades* do GeoGebra, utilizando a distribuição de Poisson para explorar, de forma interativa, a previsão de gols.

Objetivos Específicos

- a) Acessar a Calculadora de Probabilidades (web/app) e selecionar “Poisson”;
- b) Definir λ (média de gols) e gerar gráfico/tabela;
- c) Calcular $P(X = k)$ e $P(X \leq k)$ no GeoGebra;
- d) Registrar, por captura de tela, um exemplo de aplicação.

Materiais

- Dispositivos com GeoGebra (web ou app);
- Projetor/data-show;
- Anotações das aulas 2 e 3.

Metodologia

1. **Revisão rápida (3 min)** — Relembrar o significado de λ .
2. **Demonstração guiada (7 min)** — Professor projeta a Calculadora de Probabilidades $\rightarrow$ “Poisson” $\rightarrow$ define $\lambda = 1,7$ e mostra o uso.

Problema (na demonstração)

Suponha média de gols do Time A em casa $\lambda = 1,7$. Calcule no GeoGebra:

- a) $P(X = 0)$;
- b) $P(X = 2)$;
- c) $P(X \leq 1)$.

Solução no GeoGebra

- a) $P(X = 0)$: $\mu = 1,7$, intervalo $[0, 0] \Rightarrow \approx 0,1827$ (18,27%).
 b) $P(X = 2)$: $\mu = 1,7$, intervalo $[2, 2] \Rightarrow \approx 0,2640$ (26,40%).
 c) $P(X \leq 1)$: $\mu = 1,7$, somar $P(0) + P(1) \Rightarrow \approx 0,4932$ (49,32%).

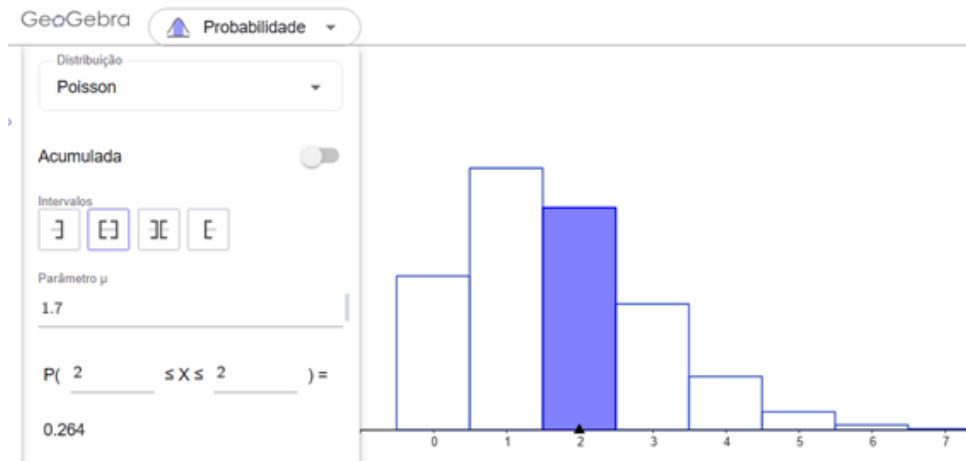


Figura 5.6. Exemplo de $P(X = 2)$ na Poisson com $\lambda = 1,7$ no GeoGebra.

- Prática guiada (10 min)** — Em duplas, repetir com $\lambda = 0,8$ e $\lambda = 1,5$: $P(0), P(1), P(2), P(\leq 2)$.
- Investigação exploratória (12 min)** — Usar *slider* para $\lambda \in [0,5, 3,0]$ (passo 0,5) e registrar $P(0), P(1), P(2)$. Observar deslocamento do pico.
- Compartilhamento (8 min)** — Apresentações rápidas: “Como o aumento de λ afeta $P(X = 2)$?”
- Síntese (5 min)** — Vantagens do GeoGebra: verificação rápida, visualização do efeito de λ , conexão com *odds/overround*.

Avaliação formativa

- **Observação em tempo real:** inserção correta de λ e leitura de $P(X = k)$;
- **Captura de tela:** cada dupla entrega print com gráfico e $P(X = 2)$;
- **Ticket de saída:** em até 3 linhas, “Qual a principal vantagem de usar o GeoGebra e como λ influenciou as probabilidades?”.

5.6.5 Aula 5 – Aplicação, Análise Crítica e Reflexões

Duração: 45 min

Objetivo Geral

Aplicar o modelo de Poisson aos dados coletados, refletindo sobre sua validade e discutindo limitações na previsão de placares, especialmente no contexto das apostas.

Objetivos Específicos

- a) Estimar λ (média de gols) a partir dos dados coletados;
- b) Calcular probabilidades de placares específicos via Poisson;
- c) Comparar com as *odds* oferecidas por casas de apostas;
- d) Refletir criticamente sobre limites do modelo e decisões conscientes.

Habilidades BNCC Mobilizadas

EM13MAT311, EM13MAT312, EM13MAT102; Competência Geral 2 (pensamento científico); Competência Geral 7 (argumentação) (BRASIL, 2018a).

Materiais

- Dados tabulados dos jogos (trazidos pelos alunos);
- Calculadora (ou app do celular);
- GeoGebra (web/app);
- Projetor e quadro branco.

Metodologia

1. **Ativação inicial (4 min)** — Retomar conjuntos de dados e escolher um para análise coletiva.
2. **Estimativa de λ (8 min)** — Calcular a média de gols do conjunto; definir contexto (mandante/visitante).
3. **Aplicação da Poisson (10 min)** — Estimar $P(X = k)$ e $P(Y = j)$ para $k, j = 0, 1, 2, 3, 4, 5$ utilizando a fórmula da distribuição de Poisson:

$$P(K = n) = \frac{e^{-\lambda} \lambda^n}{n!},$$

onde λ representa a média de gols esperados. Cada grupo deverá calcular dois valores distintos de λ :

- λ_A = média de gols marcados do time A $\times$ média de gols sofridos do time B;
- λ_B = média de gols marcados do time B $\times$ média de gols sofridos do time A.

Esses valores representam, respectivamente, os gols esperados do time A contra o B, e do time B contra o A. Com esses parâmetros, os alunos devem calcular $P(X = k)$ e $P(Y = j)$ para $k, j \in \{0, 1, 2, 3, 4, 5\}$, podendo usar a fórmula ou a Calculadora de Probabilidades do GeoGebra. Os resultados serão usados na próxima etapa para montar a matriz de probabilidades do placar.

Após os cálculos, os grupos comparam seus resultados e discutem as diferenças, refletindo sobre como a escolha dos dados afeta as previsões.

4. **Construção da matriz de probabilidade (8 min)** — Montar a Tabela 5.3 para dois times, usando $p_{i,j} = P(X = i) P(Y = j) \times 100\%$.

Tabela 5.3. Matriz de probabilidades de gols (%) — modelo genérico (0–5 gols)

Gols	0	1	2	3	4	5
0	$p_{0,0}$	$p_{0,1}$	$p_{0,2}$	$p_{0,3}$	$p_{0,4}$	$p_{0,5}$
1	$p_{1,0}$	$p_{1,1}$	$p_{1,2}$	$p_{1,3}$	$p_{1,4}$	$p_{1,5}$
2	$p_{2,0}$	$p_{2,1}$	$p_{2,2}$	$p_{2,3}$	$p_{2,4}$	$p_{2,5}$
3	$p_{3,0}$	$p_{3,1}$	$p_{3,2}$	$p_{3,3}$	$p_{3,4}$	$p_{3,5}$
4	$p_{4,0}$	$p_{4,1}$	$p_{4,2}$	$p_{4,3}$	$p_{4,4}$	$p_{4,5}$
5	$p_{5,0}$	$p_{5,1}$	$p_{5,2}$	$p_{5,3}$	$p_{5,4}$	$p_{5,5}$

Fonte: Elaborado pelo autor (2025).

Em cada elemento, $p_{i,j} = P(X = i) P(Y = j) \times 100\%$, onde $P(X = i)$ e $P(Y = j)$ são probabilidades Poisson (times A e B, respectivamente).

5. **Comparação com odds reais (8 min)** — Exibir *odds* do confronto, converter em probabilidades implícitas, verificar o *overround* e contrastar com o modelo, ressaltando que a soma das probabilidades na diagonal indica empate, as probabilidades acima da diagonal correspondem à vitória do time B e as abaixo da diagonal, à vitória do time A.
6. **Discussão crítica (4 min)** — O que a Poisson não capta? (lesões, contexto tático, dependência entre gols, etc.). Matemática como ferramenta crítica.
7. **Síntese (3 min)** — Registrar no quadro as principais limitações identificadas e encaminhar a atividade final.

Avaliação formativa

- **Cálculo de λ e aplicação da Poisson;**
- **Comparação crítica com odds reais e identificação de *overround*;**
- **Participação e argumentação baseada em dados;**
- **Registro final:** parágrafo respondendo “Quais os limites do modelo de Poisson para prever placares de futebol e como essa experiência alterou sua visão sobre as apostas esportivas?”

CONSIDERAÇÕES FINAIS

A modelagem probabilística aplicada à educação pode contribuir para mitigar efeitos negativos associados às apostas esportivas, como problemas de saúde pública, endividamento e conflitos familiares. Ao mesmo tempo, possibilita uma aprendizagem significativa e alinhada a regulamentação da educação básica. Naturalmente, isso deve estar vinculado a uma legislação que tente assegurar a prática responsável, prevenindo fraudes, lavagem de dinheiro e desonestidade aos clientes das casas de apostas.

Na análise das *odds* do Campeonato Paulista de Futebol de 2025, verificou-se que os valores calculados estão coerentes com os divulgados pelas casas de apostas. Isto evidencia que, mesmo com a utilização de modelos probabilísticos simples, é possível alcançar um nível satisfatório nas estimativas das odds. Entretanto, deve-se destacar que as casas de apostas utilizam modelos mais sofisticados, baseados em *machine learning* e *deep learning*. Além disso, recorrem a ajustes dinâmicos de risco e a estratégias de sobrecarga (*overrounds*), que asseguram sua lucratividade no médio e longo prazo.

A sequência didática apresentada buscou, em sala de aula, estimular a criticidade, permitindo que o estudante, às vésperas de concluir o Ensino Médio, desenvolvesse uma compreensão mais reflexiva sobre as apostas esportivas. É igualmente importante que os alunos se conscientizem dos impactos sociais e econômicos negativos que as apostas podem acarretar, como problemas de saúde coletiva e desequilíbrios financeiros, conforme indicam estudos sobre a compulsão por apostar.

A maior parte da literatura trata da modelagem do número de gols marcados por cada equipe, em que as distribuições Poisson bivariada e Poisson univariada têm sido muito utilizadas. Uma proposta de trabalho futuro é modelar a diferença do número de gols. Isto apresenta algumas vantagens como a eliminação da correlação imposta pelo fato de os dois times adversários competirem entre si. Neste caso a distribuição Skellam é a que vem sendo proposta por alguns autores como Karlis e Ntzoufras (2009)

Referências

BRASIL. **Base Nacional Comum Curricular**. 2018. Disponível em:
<<http://basenacionalcomum.mec.gov.br>>. Acesso em: 14 jun. 2025.

_____. **Constituição da República Federativa do Brasil de 1988**. Brasília: Senado Federal, 1988. Disponível em:
<https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm>. Acesso em: 20 nov. 2024.

_____. **Lei nº 13.756, de 12 de dezembro de 2018**. [S.l.: s.n.], 2018. Disponível em: <https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13756.htm>. Acesso em: 20 nov. 2024.

_____. **Lei nº 14.790, de 29 de dezembro de 2023**. [S.l.: s.n.], 2023. Dispõe sobre a tributação e regulamentação das apostas de quota fixa, e altera as Leis nº 5.768/71 e nº 13.756/2018. Disponível em: <https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2023/lei/L14790.htm>. Acesso em: 20 nov. 2024.

_____. **Parâmetros Curriculares Nacionais: Ensino Médio. Matemática**. 1999. Disponível em: <<http://portal.mec.gov.br/seb/arquivos/pdf/mat.pdf>>. Acesso em: 1 jun. 2025.

BRASIL. **Base Nacional Comum Curricular**. Disponível em: <https://basenacionalcomum.mec.gov.br/>. Acesso em: 20 jul. 2025. Ministério da Educação. 2018. Disponível em: <<https://basenacionalcomum.mec.gov.br/>>. Acesso em: 20 jul. 2025.

_____. **Lei de Diretrizes e Bases da Educação Nacional, Lei nº 9.394, de 20 de dezembro de 1996**. Brasília: Presidência da República, 1996. Disponível em: <https://www.planalto.gov.br/ccivil_03/leis/l9394.htm>. Acesso em: 20 nov. 2024.

CÉSAR, Augusto; MORGADO, Paulo Cezar. **Matemática Discreta**. 3. ed. Rio de Janeiro: Sociedade Brasileira de Matemática, 2022. Material do Mestrado Profissional em Matemática em Rede Nacional (PROFMAT).

CNN BRASIL. **Setor de apostas online cresceu 734% desde 2021, aponta pesquisa**. 2024. Disponível em: <<https://www.cnnbrasil.com.br/economia/negocios/setor-de-apostas-online-cresceu-734-desde-2021-aponta-pesquisa/>>. Acesso em: 13 jan. 2025.

CONFEDERAÇÃO BRASILEIRA DE FUTEBOL. **Classificação oficial – Brasileirão Série B 2024**. 2024. Disponível em: <<https://www.cbf.com.br/futebol-brasileiro/tabelas/campeonato-brasileiro/serie-b/2024>>. Acesso em: 7 fev. 2025.

_____. **Confederação Brasileira de Futebol**. 2025. Disponível em: <<https://www.cbf.com.br/>>. Acesso em: 2 jun. 2025.

DIXON, Mark J.; COLES, Stuart G. Modelling Association Football Scores and Inefficiencies in the Football Betting Market. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, v. 46, n. 2, p. 265–280, 1997. DOI: 10.1111/1467-9876.00065.

FAZENDA, MINISTÉRIO DA. **SIGAP - Sistema de Gestão de Acessos e Perfis**. 2024. Disponível em: <<https://sigap.fazenda.gov.br/login>>. Acesso em: 20 nov. 2024.

FÉDÉRATION INTERNATIONALE DE FOOTBALL ASSOCIATION (FIFA). **Resultados da Copa do Mundo FIFA 2022**. Acesso em: 14 jun. 2025. FIFA. 2022. Disponível em: <<https://www.fifa.com/pt/tournaments/mens/worldcup/qatar2022/scores-fixtures>>. Acesso em: 14 jun. 2025.

FOOTBALL DATA. **Resultados de partidas – Football-Data.co.uk**. [S.l.: s.n.], 2025. Football Data. Disponível em: <<http://www.football-data.co.uk/matches.php>>. Acesso em: 15 maio 2025.

GLOBO ESPORTE. **Tabela, jogos e horários do Campeonato Paulista 2025**. [S.l.: s.n.], 2025. Globo. Disponível em: <<https://ge.globo.com/sp/futebol/campeonato-paulista/>>. Acesso em: 10 maio 2025.

INTERNATIONAL GEOGEBRA INSTITUTE. **GeoGebra Calculator Suite**. Recurso digital. 2025. Disponível em: <<https://www.geogebra.org/calculator>>. Acesso em: 10 jul. 2025.

KARLIS, Dimitris; NTZOUFRAS, Ioannis. Bayesian modelling of football outcomes: Using the Skellam’s distribution for the goal difference. **IMA Journal of Management Mathematics**, v. 20, n. 2, p. 133–145, 2009. DOI: 10.1093/imaman/dpn026. Disponível em: <<https://doi.org/10.1093/imaman/dpn026>>. Acesso em: 19 set. 2025.

MAHER, Michael J. Modelling association football scores. **Statistica Neerlandica**, v. 36, n. 3, p. 109–118, 1982. DOI: 10.1111/j.1467-9574.1982.tb00782.x.

MARÇAL, Guilherme et al. **Finalizações, posse de bola, goleiros e jogadores mais decisivos: 12 rankings do 1º turno do Brasileirão 2022**. ge.globo. 26 jul. 2022. Disponível em: <<https://ge.globo.com/espiao-estatistico/noticia/2022/07/26/finalizacoes-posse-de-bola-goleiros-e-jogadores-mais-decisivos-veja-12-rankings-do-1o-turno-do-brasileirao.ghml>>. Acesso em: 14 jun. 2025. Inclui ranking de times com maior percentual de posse de bola no 1º turno do Brasileirão 2022.

MEYER, Paul L. **Probabilidade: aplicações à estatística**. 1. ed. Rio de Janeiro: Ao Livro Técnico S.A., 1989. Tradução da obra original *Introductory Probability and Statistical Applications*. Reimpressões: 1970, 1972, 1974.

MORETTIN, Pedro Alberto; BUSSAB, Wilton de Oliveira. **Estatística básica**. 6. ed. São Paulo: Saraiva, 2010. ISBN 978-85-02-08177-2.

MORONEY, M. J. **Facts from Figures**. London: Penguin Books, 1951. (Pelican Books). Pelican Book A236.

ODDS AGORA. **Comparação de Odds para Al Ahly Cairo × Al Ittihad – Copa do Mundo de Clubes FIFA 2023**. 15 dez. 2023. Disponível em:

<<https://www.oddsagora.com.br/football/world/copa-do-mundo-de-clubes-fifa-2023/al-ahly-al-ittihad-AiNoahom/#1X2;2>>. Acesso em: 18 jul. 2025.

ODDSPEDIA. **Paulista Série A1 Odds – Football – Brazil**. 2025. Disponível em:

<<https://oddspedia.com/football/brazil/paulista-serie-a1>>. Acesso em: 2 jun. 2025.

PÓVOA, Luciano et al. **O mercado de apostas esportivas on-line: impactos, desafios para a definição de regras de funcionamento e limites**. Brasília, 2023.

Disponível em: <<https://www12.senado.leg.br/institucional/oconleg/>>. Acesso em: 20 nov. 2024.

ROSS, Sheldon. **Probabilidade: um curso moderno com aplicações**. Tradução: Alberto Resende De Conti. 8. ed. Porto Alegre: Bookman, 2010. 608 p. ISBN 978-85-7780-621-8.

WINDRAWIN. **Estatísticas de resultados finais - Brasileirão Série A**. 2025.

Disponível em: <<https://www.windrawwin.com/br/estatisticas/resultados-final/brasileirao-serie-a/>>. Acesso em: 20 nov. 2024.

ZABALA, Antoni. **A prática educativa: Como ensinar**. Porto Alegre: Artmed, 1998. Trad. Ernani F. da Rocha.

Códigos em R da Análise de Dados do Campeonato Paulista de Futebol 2025

```

1
2 # Modelo de Poisson para Corinthians x Palmeiras (Paulistão 2025)
3 # Autor: Rodrigo dos Santos Neri da Silva -- Junho de 2025
4
5 ## 1. Preparar ambiente e dados
6 rm(list = ls())
7
8 jogos <- data.frame(
9   time_casa = c(
10     "palmeiras", "red bull bragantino", "nordeste", "corinthians",
11     "corinthians", "santos", "palmeiras", "sao paulo",
12     "palmeiras", "ponte preta", "corinthians", "guarani",
13     "novorizontino", "palmeiras", "corinthians", "agua santa",
14     "corinthians", "inter de limeira", "portuguesa", "palmeiras",
15     "palmeiras", "corinthians", "mirassol", "sao bernardo",
16     "corinthians", "corinthians", "palmeiras", "palmeiras"
17   ),
18   time_visitante = c(
19     "portuguesa", "corinthians", "palmeiras", "velo clube",
20     "agua santa", "palmeiras", "novorizontino", "corinthians",
21     "red bull bragantino", "corinthians", "nordeste", "palmeiras",
22     "corinthians", "corinthians", "sao bernardo", "palmeiras",
23     "santos", "palmeiras", "corinthians", "sao paulo",
24     "botafogo-sp", "guarani", "palmeiras", "palmeiras",
25     "mirassol", "santos", "sao paulo", "corinthians"
26   ),
27   gols_casa = c(
28     2,1,1,2,2,1,1,3,0,0,2,1,0,1,2,1,2,0,2,0,3,2,2,0,2,2,1,0
29   ),
30   gols_visitante = c(
31     0,2,1,1,1,2,2,1,0,1,1,4,1,1,0,1,1,3,2,0,1,2,3,3,0,1,0,1
32   ),
33   stringsAsFactors = FALSE
34 )
35
36 ## 2. Funções auxiliares
37 soma_gols <- function(time, coluna_times, coluna_gols) {

```

```

38   sum(coluna_gols[coluna_times == time])
39 }
40
41 gols_por_time <- function(time) {
42   marcados <- soma_gols(time, jogos$time_casa, jogos$gols_casa) +
43     soma_gols(time, jogos$time_visitante, jogos$gols_
44       visitante)
45   sofridos <- soma_gols(time, jogos$time_visitante, jogos$gols_casa
46     ) +
47     soma_gols(time, jogos$time_casa, jogos$gols_visitante
48     )
49   partidas <- sum(jogos$time_casa == time) +
50     sum(jogos$time_visitante == time)
51   list(marcados = marcados, sofridos = sofridos, jogos = partidas)
52 }
53
54 ## 3. Taxas médias de ataque e defesa (lambda)
55 times <- c("corinthians", "palmeiras")
56 estatisticas <- lapply(times, gols_por_time)
57
58 lambda_ataque <- sapply(estatisticas, function(x) x$marcados / x$
59   jogos)
60 lambda_defesa <- sapply(estatisticas, function(x) x$sofridos / x$
61   jogos)
62 names(lambda_ataque) <- names(lambda_defesa) <- times
63
64 cat("lambda de ataque (gols por jogo):\n")
65 print(lambda_ataque)
66
67 cat("\nlambda de defesa (gols por jogo):\n")
68 print(lambda_defesa)
69
70 ## 4. Probabilidades de placares
71
72 lambda_cor <- lambda_ataque["corinthians"] * lambda_defesa["
73   palmeiras"]
74 lambda_pal <- lambda_ataque["palmeiras"] * lambda_defesa["
75   corinthians"]
76
77 gols <- 0:10
78 p_cor <- dpois(gols, lambda_cor)
79 p_pal <- dpois(gols, lambda_pal)
80
81 prob_mat <- outer(p_cor, p_pal) * 100
82 dimnames(prob_mat) <- list(corinthians = gols, palmeiras = gols)
83
84 cat("\nMatriz de probabilidades (0-7 gols):\n")
85 print(round(prob_mat[1:8, 1:8], 1))

```

```
82 prob_cor <- round(sum(prob_mat[lower.tri(prob_mat)]), 1)
83 prob_emp <- round(sum(diag(prob_mat)), 1)
84 prob_pal <- round(sum(prob_mat[upper.tri(prob_mat)]), 1)
85
86 ## 5. Cálculo das odds
87 odds <- data.frame(
88   categoria = c("Corinthians", "Empate", "Palmeiras"),
89   Odds = (1 / c(prob_cor, prob_emp, prob_pal)) * 100
90 )
91
92 cat("\nOdds sugeridas (em %):\n")
93 print(odds)
94
95 ## 6. Gráfico de pizza das probabilidades
96 if (!requireNamespace("ggplot2", quietly = TRUE)) {
97   install.packages("ggplot2")
98 }
99 library(ggplot2)
100
101 df <- data.frame(
102   categoria = c("Corinthians", "Empate", "Palmeiras"),
103   percent = c(prob_cor, prob_emp, prob_pal)
104 )
105
106 ggplot(df, aes(x = "", y = percent, fill = categoria)) +
107   geom_bar(width = 1, stat = "identity") +
108   coord_polar("y") +
109   geom_text(aes(label = sprintf('%.1f%%', percent)),
110             position = position_stack(vjust = 0.5)) +
111   labs(title = "Probabilidades - Corinthians x Palmeiras") +
112   theme_void() +
113   theme(legend.position = "right")
```

Listing A.1. Script completo do modelo.

Apêndice B

Códigos em R da Análise de Dados do Campeonato Brasileiro de Futebol – Série A 2025

```
1
2 # Modelo de Maher (1982) - Distribuicao de Poisson Independente
3
4 #
5
6 -----
7
8 # 0) Preparacao do ambiente
9 #
10 -----
11
12 rm(list = ls())
13 options(encoding = "UTF-8")
14
15 # Carregamento de pacotes necessarios
16 library(dplyr)
17 library(ggplot2)
18 library(lubridate)
19
20 #
21 -----
22
23 # 1) Leitura e filtragem dos dados
24 #
25 -----
26
27 # Definicao do caminho relativo para o arquivo de dados CSV
28 file_path <- file.path("dados", "BRA.csv")
29
30 #
31 -----
32
33 # Leitura do arquivo em UTF-8
34 jogos_raw <- read.csv(
35   file      = file_path,
36   header    = TRUE,
37   stringsAsFactors = FALSE,
38   fileEncoding = "UTF-8"
```

```

27 )
28
29 # Preparacao dos dados dos ultimos 2 anos
30 matches <- jogos_raw %>%
31   mutate(data = dmy(Date)) %>%
32   filter(!is.na(HG) & !is.na(AG)) %>%
33   transmute(
34     data          = data,
35     mandante      = tolower(Home),
36     visitante     = tolower(Away),
37     gols_mandante = HG,
38     gols_visitante = AG
39   ) %>%
40   filter(
41     data >= Sys.Date() - years(2),
42     data <= Sys.Date()
43   )
44
45 cat(sprintf("Total de jogos processados: %d\n", nrow(matches)))
46
47 #
48 -----
49
50 # 2) Estimacao do modelo de Maher
51 #
52 -----
53
54 fit_maher <- function(data, tol = 1e-8, max_iter = 200) {
55   times <- sort(unique(c(data$mandante, data$visitante)))
56   n <- length(times)
57   gols_marcados <- setNames(rep(0, n), times)
58   gols_sofridos <- setNames(rep(0, n), times)
59
60   for (i in seq_len(nrow(data))) {
61     m <- data$mandante[i]
62     v <- data$visitante[i]
63     gols_marcados[m] <- gols_marcados[m] + data$gols_mandante[i]
64     gols_marcados[v] <- gols_marcados[v] + data$gols_visitante[i]
65     gols_sofridos[m] <- gols_sofridos[m] + data$gols_visitante[i]
66     gols_sofridos[v] <- gols_sofridos[v] + data$gols_mandante[i]
67   }
68
69   a <- rep(1, n); names(a) <- times
70   b <- rep(1, n); names(b) <- times
71   k <- 1
72
73   for (it in seq_len(max_iter)) {
74     a_old <- a; b_old <- b; k_old <- k
75
76     denom_att <- sapply(times, function(t) {
77       sum(b[data$visitante[data$mandante == t]] * k^2) +

```

```

74     sum(b[data$mandante[data$visitante == t]])
75   })
76   a <- gols_marcados / pmax(denom_att, 1e-8)
77   a <- a * n / sum(a)
78
79   denom_def <- sapply(times, function(t) {
80     sum(a[data$mandante[data$visitante == t]] * k^2) +
81     sum(a[data$visitante[data$mandante == t]])
82   })
83   b <- gols_sofridos / pmax(denom_def, 1e-8)
84
85   num_k <- sum(data$gols_mandante)
86   den_k <- sum(a[data$mandante] * b[data$visitante])
87   k <- sqrt(num_k / max(den_k, 1e-8))
88
89   delta <- max(abs(c(a - a_old, b - b_old, k - k_old)), na.rm =
90     TRUE)
91   if (!is.na(delta) && delta < tol) break
92 }
93 list(ataque = a, defesa = b, vantagem_casa = k, iteracoes = it)
94 }
95
96 fit <- fit_maher(matches)
97 cat("Modelo treinado.\n")
98
99 #
100 -----
101 # 3) Funcao de previsao
102 #
103 -----
104 pred_maher <- function(mandante, visitante, fit, max_gols = 10) {
105   if (!all(c(mandante, visitante) %in% names(fit$ataque))) {
106     stop("Times nao encontrados no ajuste do modelo.")
107   }
108   a <- fit$ataque; b <- fit$defesa; k <- fit$vantagem_casa
109   lambda_m <- k^2 * a[mandante] * b[visitante]
110   lambda_v <- a[visitante] * b[mandante]
111   probs <- outer(
112     dpois(0:max_gols, lambda_m),
113     dpois(0:max_gols, lambda_v)
114   )
115   dimnames(probs) <- list(mandante = 0:max_gols, visitante = 0:max_gols)
116   probs
117 }
118 #
119 -----

```

```

118 # 4) Exemplo de uso
119 #
-----
120 probs_exemplo <- pred_maher("flamengo rj", "palmeiras", fit)
121 P_mandante <- sum(probs_exemplo[lower.tri(probs_exemplo)])
122 P_empate <- sum(diag(probs_exemplo))
123 P_visitante <- sum(probs_exemplo[upper.tri(probs_exemplo)])
124
125 cat(sprintf("Probabilidades: Mandante=%.1f%% | Empate=%.1f%% |
126             Visitante=%.1f%%\n",
127             100 * P_mandante, 100 * P_empate, 100 * P_visitante))
128 #
-----
129 # 5) Calculo de odds justas e grafico
130 #
-----
131 odds <- 1 / c(P_mandante, P_empate, P_visitante)
132 names(odds) <- c("Mandante", "Empate", "Visitante")
133
134 df_plot <- data.frame(
135   categoria = names(odds),
136   prob      = round(100 * c(P_mandante, P_empate, P_visitante), 1)
137 )
138
139 ggplot(df_plot, aes(x = "", y = prob, fill = categoria)) +
140   geom_bar(width = 1, stat = "identity") +
141   coord_polar(theta = "y") +
142   geom_text(aes(label = paste0(prob, "%")), position = position_
143             stack(vjust = 0.5)) +
144   theme_void() +
145   labs(title = "Probabilidades Justas - Flamengo x Palmeiras", fill
146         = NULL) +
147   theme(plot.title = element_text(hjust = 0.5))

```

Listing B.1. Script completo do modelo.

Estimação de Parâmetros

C.1 Introdução

A inferência estatística, que se dedica a tirar conclusões de uma população através de uma amostra da população, apresenta alguns processos, como:

- **Amostragem:** conhecer e definir a população, bem como escolher uma amostra que a represente; por fim, obter as informações relevantes dessa amostra selecionada.
- **Estimação de parâmetro:** tem por objetivo estimar as características ou parâmetros da população. Essa estimação pode ser pontual ou por intervalos de confiança. Nesta seção, abordaremos a estimação pontual, vide por exemplo, Morettin e Bussab (2010).

Definição C.1. Estimador. Seja θ um parâmetro desconhecido de uma variável aleatória X e $(X_1, X_2, \dots, X_n)$ uma amostra aleatória de tamanho n de X . Chama-se *estimador* de θ qualquer estatística

$$T = g(X_1, \dots, X_n),$$

isto é, *qualquer função (mensurável) dos dados amostrais* destinada a inferir θ . A realização numérica $t = g(x_1, \dots, x_n)$, obtida após observar a amostra, é denominada *estimativa* de θ .

C.2 Estimador de Máxima Verossimilhança

O método de máxima verossimilhança, também conhecido na literatura como *Maximum Likelihood Estimation* (MLE), tem como princípio fundamental a escolha do parâmetro populacional que mais provavelmente teria gerado a amostra observada. Em outras palavras, o parâmetro é estimado de forma a maximizar a verossimilhança dos dados.

Definição C.2. A função de verossimilhança para uma amostra $x_1, x_2, \dots, x_n$ é definida por:

$$L(\theta; x_1, \dots, x_n) = f(x_1; \theta)f(x_2; \theta) \dots f(x_n; \theta) = \prod_{i=1}^n f(x_i; \theta), \quad (\text{C.1})$$

sendo uma função do parâmetro θ , enquanto os dados observados são fixos. O **estimador de máxima verossimilhança** (EMV) é o valor $\hat{\theta}$ que maximiza essa função:

$$\hat{\theta} = \arg \max_{\theta} L(\theta; x_1, \dots, x_n).$$

Em muitos casos, como na distribuição de Poisson, o parâmetro λ pode ser estimado por meio do método da máxima verossimilhança. O procedimento usual consiste em tomar o logaritmo da função de verossimilhança (log-verossimilhança), derivá-la em relação ao parâmetro θ e igualar a derivada a zero para encontrar o valor que a maximiza.

C.3 Estimando o Parâmetro λ de uma Poisson

Vamos utilizar o método de máxima verossimilhança para estimar o parâmetro λ da função de probabilidade dada pela expressão (3.22).

Seja um conjunto de observações $x_1, x_2, \dots, x_n$ que segue uma distribuição de Poisson com parâmetro λ . A função de verossimilhança é dada por:

$$L(\lambda) = \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{x_i}}{x_i!}$$

Reescrevendo a função de verossimilhança, temos:

$$L(\lambda) = e^{-n\lambda} \lambda^{\sum_{i=1}^n x_i} \prod_{i=1}^n \frac{1}{x_i!}$$

Usamos o logaritmo natural da função de verossimilhança para facilitar a derivação, obtendo a função log-verossimilhança:

$$\log L(\lambda) = -n\lambda + \left(\sum_{i=1}^n x_i \right) \log(\lambda) - \sum_{i=1}^n \log(x_i!)$$

A seguir, derivamos a função log-verossimilhança em relação ao parâmetro λ e igualamos a derivada a zero:

$$\frac{\partial \log L(\lambda)}{\partial \lambda} = -n + \frac{\sum_{i=1}^n x_i}{\lambda} = 0$$

Isolando λ , o estimador de máxima verossimilhança é dado por:

$$\hat{\lambda} = \frac{\sum_{i=1}^n X_i}{n} = \bar{X}$$

De onde se conclui que o estimador de máxima verossimilhança para o parâmetro λ da distribuição de Poisson é a média amostral, representada por $\bar{X}$.