

ESTATÍSTICA E ANÁLISE DE DADOS CIENTÍFICOS

SARAH G. A. BARBOSA

2024

SUMÁRIO

I Probabilidade, variáveis aleatórias e estatísticas

1

1

TEORIA DA PROBABILIDADE	2
1.1 Experimentos e Eventos	2
1.2 Probabilidade de Eventos	3
1.2.1 Axiomas de Kolmogorov	3
1.2.2 Método Frequentista ou Clássico	4
1.2.3 Método Bayesiano ou Empírico	4
1.2.4 Propriedades Fundamentais da Probabilidade	4
1.3 Probabilidade Condicional	5
1.4 Independência Estatística	6
1.5 Teorema da Probabilidade Total e Teorema de Bayes	8
1.6 Problemas	9

2

VARIÁVEIS ALEATÓRIAS E SUAS DISTRIBUIÇÕES	11
2.1 Variáveis Aleatórias	11
2.2 Funções de Distribuição de Probabilidade	11
2.3 Expectância e Momentos de uma Função de Distribuição	13
2.3.1 A Média e a Média Amostral	13
2.3.2 A Lei dos Grandes Números	14
2.3.3 A Variância e a Variância Amostral	15
2.4 Covariância e Correlação entre Variáveis Aleatórias	15
2.4.1 Distribuição de Probabilidade Conjunta e Momentos de Duas Variáveis Aleatórias	15
2.4.2 Independência Estatística de Variáveis Aleatórias	17
2.5 A Expectância da Variância Amostral e Covariância Amostral	18
2.6 Problemas	20

3

TRÊS DISTRIBUIÇÕES FUNDAMENTAIS: BINOMIAL, GAUSSIANA E POISSON	22
3.1 A Distribuição Binomial	22
3.1.1 Derivação da Distribuição Binomial	22

3.1.2	Momentos da Distribuição Binomial	24
3.2	A Distribuição Gaussiana	25
3.2.1	Derivação da Distribuição Gaussiana a partir da Distribuição Binomial	26
3.2.2	Momentos e Propriedades da Distribuição Gaussiana	28
3.3	A Distribuição de Poisson	29
3.3.1	Derivação da Distribuição de Poisson	29
3.3.2	Momentos e Propriedades da Distribuição de Poisson	30
3.3.3	A Distribuição de Poisson e o Processo de Poisson	30
3.4	Comparação das Distribuições Binomial, Gaussiana e de Poisson	32

RESOLUÇÃO DOS PROBLEMAS

Capítulo 1	33
Capítulo 2	35

REFERÊNCIAS BIBLIOGRÁFICAS

PARTE I:

PROBABILIDADE, VARIÁVEIS ALEATÓRIAS E ESTATÍSTICAS

1

TEORIA DA PROBABILIDADE

1.1 Experimentos e Eventos

Todo experimento tem uma série de resultados possíveis. Por exemplo, o experimento consistindo no lançamento de um dado pode ter seis resultados possíveis, de acordo com o número que aparece após o dado cair. Atribuir probabilidades aos resultados possíveis é uma das principais tarefas da teoria da probabilidade. A seção seguinte apresenta dois métodos para atribuir probabilidades, o método clássico baseado na repetição do experimento e um método baseado no conhecimento empírico do experimento. O fato de haver mais de um método disponível para este propósito não deve ser visto como uma limitação da teoria, mas sim como o fato de que para certas partes da teoria da probabilidade, e ainda mais para a estatística, há um elemento de subjetividade que entra na análise e na interpretação dos resultados. Portanto, é tarefa do estatístico acompanhar quaisquer suposições feitas na análise e levá-las em consideração na interpretação dos resultados. Antes de descrever os métodos para atribuir probabilidades, é necessário desenvolver a terminologia necessária para descrever experimentos e seus resultados.

O **espaço amostral** Ω é definido como o conjunto de todos os resultados possíveis do experimento. No caso do lançamento de um dado, o espaço amostral pode ser escrito como o conjunto dos seis possíveis resultados, $\Omega = \{1, 2, 3, 4, 5, 6\}$. Um **evento** A é um subconjunto de Ω , $A \subset \Omega$, e representa uma série de resultados possíveis para o experimento. Por exemplo, o evento “número par” pode ser representado por $A = \{2, 4, 6\}$, e o evento “número ímpar” como $B = \{1, 3, 5\}$. Diferentes experimentos terão espaços amostrais diferentes que podem ser escritos de maneira equivalente. Para cada experimento, sempre existem dois eventos: o próprio espaço amostral que compreende todos os resultados possíveis e o conjunto vazio que não contém resultados, representado como $A = \emptyset$ e chamado de **evento impossível**.

Eventos são convenientemente estudados usando a teoria elementar dos conjuntos. É útil revisar algumas das propriedades da teoria dos conjuntos que são comumente usadas em probabilidade e estatística. O complementar de um evento A é indicado como A^c , e é o conjunto de todos os resultados possíveis exceto aqueles em A . Por exemplo, o complemento do evento “número ímpar” é o evento “número par”.

Dados dois eventos A e B , a união $C = A \cup B$ é o evento que compreende todos os resultados de A e os de B . No lançamento de um dado, a união de números ímpares e pares é o próprio espaço amostral, consistindo de todos os resultados possíveis. A interseção de dois eventos $C = A \cap B$ é o evento que compreende todos os resultados de A que também são resultados de B . Quando dois eventos não se sobrepõem, $A \cap B = \emptyset$, os eventos são ditos **mutuamente exclusivos**. A união e a interseção podem ser naturalmente estendidas para mais de dois eventos. Os eventos são ilustrados na [Figura 1.1](#).

Finalmente, um número de eventos A_i , para $i = 1, \dots, n$, são ditos uma **partição** do espaço amostral se satisfizerem as seguintes duas propriedades:

$$\begin{cases} A_i \cap A_j = \emptyset, \text{ quando } i \neq j \\ \bigcup_{i=1}^n A_i = \Omega. \end{cases} \quad (1.1)$$

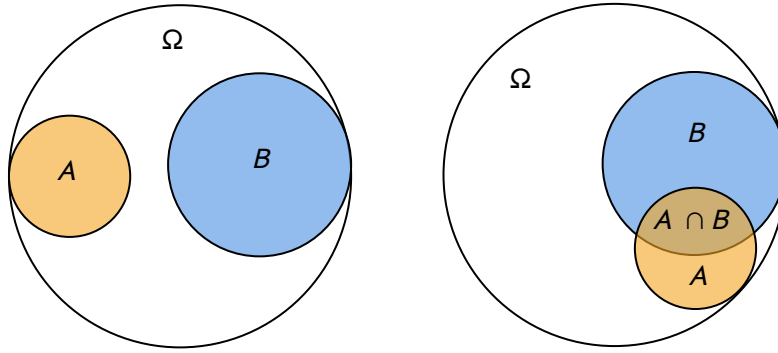


Figura 1.1: Os eventos podem ser representados como subconjuntos do espaço amostral. A probabilidade do evento $P(A \cup B)$ é a soma das duas probabilidades individuais, apenas se os dois eventos forem mutuamente exclusivos. Esta propriedade permite a interpretação da probabilidade como a “área” de um determinado evento dentro do espaço amostral.

Por exemplo, os resultados 1, 2, 3, 4, 5 e 6 para o lançamento de um dado dividem o espaço amostral em uma série de eventos que cobrem todos os resultados possíveis, sem qualquer sobreposição entre eles.

1.2 Probabilidade de Eventos

A probabilidade P de um evento é um número que tem como objetivo descrever as chances de ocorrência de um evento em um único ensaio do experimento. A teoria moderna da probabilidade foi desenvolvida ao longo da primeira metade do século XX com a contribuição de diversos matemáticos, incluindo [Bernshtein \(1917\)](#) e [Kolmogorov \(1950\)](#). Todas essas contribuições levaram a um arcabouço que considera a probabilidade como um número entre 0 e 1, onde $P = 0$ corresponde a um evento impossível, e $P = 1$ a um evento certo. Portanto, a operação de “probabilidade” pode ser pensada como uma função que transforma cada evento possível em um número real entre 0 e 1.

1.2.1 Axiomas de Kolmogorov

O primeiro passo para determinar a probabilidade de um evento é estabelecer um número de regras básicas que capturem o significado da probabilidade. A probabilidade de um evento precisa satisfazer três axiomas definidos por [Kolmogorov \(1950\)](#)¹:

1. A probabilidade de um evento A é um número não negativo, $P(A) \geq 0$;
2. A probabilidade de todos os resultados possíveis, ou espaço amostral, é normalizada para o valor de unidade, $P(\Omega) = 1$;
3. Se A e B são dois eventos mutuamente exclusivos, então

$$P(A \cup B) = P(A) + P(B). \quad (1.2)$$

A [Figura 1.1](#) ilustra a última propriedade usando diagramas de Venn. Os eventos no painel esquerdo são mutuamente exclusivos, e a probabilidade da união é representada pela área de $A \cup B$, ou a soma das duas áreas individuais. Para eventos que não são mutuamente exclusivos, como aqueles no painel direito, essa propriedade não se aplica.

Esses axiomas devem ser considerados as “regras básicas” da teoria da probabilidade, mas não fornecem orientação sobre como atribuir probabilidades aos eventos. Para este fim, existem duas principais abordagens disponíveis. Uma é baseada na repetição dos experimentos um grande número de vezes sob as mesmas condições, e é conhecida como método frequentista ou clássico. A outra é baseada em um conhecimento mais teórico do experimento, mas sem a exigência experimental de um grande número de repetições, e é chamada de método Bayesiano ou empírico.

¹Em seu livro, [Kolmogorov \(1950\)](#) lista um número maior de axiomas devido à necessidade de garantir certas propriedades matemáticas da probabilidade.

1.2.2 Método Frequentista ou Clássico

Considere a realização de um experimento um grande número N de vezes, sob as mesmas condições experimentais. A ocorrência do evento A é indicada como o número $N(A)$. A probabilidade do evento A é dada por

$$P(A) = \lim_{N \rightarrow \infty} \frac{N(A)}{N} \quad (1.3)$$

ou seja, a probabilidade é a frequência relativa de ocorrência de um evento dado a partir de muitas repetições do mesmo experimento. A limitação óbvia dessa definição é a necessidade de realizar o experimento um grande número de vezes. Este requisito não apenas consome tempo, mas também exige que o experimento seja repetível em primeiro lugar, o que pode ou não ser possível. A limitação deste método é evidente ao considerar um lançamento de moeda: não importa o número de lançamentos, a ocorrência de “cara para cima” nunca será exatamente 50%, o que seria esperado com base em um conhecimento empírico do experimento em questão.

1.2.3 Método Bayesiano ou Empírico

Outro método para atribuir probabilidades é usar o conhecimento do experimento, tanto teórico quanto experimental, mas sem a necessidade de dados experimentais extensivos. A probabilidade atribuída a um evento representa o **grau de crença** de que o evento ocorrerá em uma tentativa específica do experimento, e implica um elemento de subjetividade que se tornará mais evidente com o teorema de Bayes. Às vezes, é referida como **probabilidade empírica**, em reconhecimento ao fato de que, às vezes, a probabilidade de um evento é atribuída com base em um conhecimento prático do experimento, embora sem a exigência clássica de repetir o experimento um grande número de vezes. Esse método é nomeado em homenagem ao Reverendo Thomas Bayes, que pioneiramente desenvolveu a teoria da probabilidade (Bayes and Price, 1763).

Exemplo 1.1

No experimento de lançamento de uma moeda, a determinação da probabilidade empírica para os eventos “Cara para cima” ou “Cora para cima” depende do conhecimento de que a moeda é imparcial, e que portanto deve ser verdade que $P(\text{Cara}) = P(\text{Cora})$. Essa afirmação empírica significa o uso do método Bayesiano para determinar probabilidades. Com essa informação, podemos então simplesmente usar os axiomas de Kolmogorov para afirmar que $P(\text{Cara}) + P(\text{Cora}) = 1$, e portanto obter o resultado intuitivo de que $P(\text{Cara}) = P(\text{Cora}) = 1/2$.

1.2.4 Propriedades Fundamentais da Probabilidade

As seguintes propriedades são úteis para atribuir e manipular probabilidades de eventos. Elas são um tanto intuitivas, mas mesmo assim é instrutivo derivá-las dos axiomas de Kolmogorov.

1. A probabilidade do evento nulo é zero, $P(\emptyset) = 0$.
Esta propriedade pode ser derivada começando com os eventos mutuamente exclusivos \emptyset e Ω . Uma vez que a união deles é Ω , segue do terceiro axioma que $P(\Omega) = P(\Omega) + P(\emptyset)$. A partir do segundo axioma, sabe-se que $P(\Omega) = 1$, e com isso, conclui-se que $P(\emptyset) = 0$. A seguinte propriedade é uma generalização desta.
2. A probabilidade do evento complementar A^c satisfaz a propriedade

$$P(A^c) = 1 - P(A). \quad (1.4)$$

Por definição, é verdade que $A \cup A^c = \Omega$, e que A e A^c são mutuamente exclusivos. Usando os segundo e terceiro axiomas, $P(A \cup A^c) = P(A) + P(A^c) = 1$, a partir do qual segue que $P(A^c) = 1 - P(A)$.

3. A probabilidade da união de dois eventos satisfaz a propriedade geral

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \quad (1.5)$$

Esta propriedade generaliza o terceiro axioma de Kolmogorov e pode ser interpretada como o fato de que os resultados na região de sobreposição dos dois eventos devem ser contados apenas uma vez, como ilustrado na [Figura 1.1](#). Primeiramente, perceba que o evento $A \cup B$ pode ser escrito como a união de três conjuntos mutuamente exclusivos,

$$A \cup B = (A \cap B^c) \cup (B \cap A^c) \cup (A \cap B),$$

veja a [Figura 1.1](#). Portanto, usando o terceiro axioma,

$$P(A \cup B) = P(A \cap B^c) + P(B \cap A^c) + P(A \cap B).$$

Em seguida, observe que para qualquer evento A ou B , é verdade que $A = (A \cap B^c) \cup (A \cap B)$, já que $\{B, B^c\}$ é uma partição de Ω . Isso implica que $P(A) = P(A \cap B) + P(A \cap B^c)$ devido ao fato de que os dois conjuntos são novamente mutuamente exclusivos, com uma equação similar para o evento B . Assim, segue-se que,

$$P(A \cup B) = P(A) - P(A \cap B) + P(B) - P(B \cap A) + P(A \cap B) = P(A) + P(B) - P(A \cap B),$$

o que prova a propriedade.

Exemplo 1.2

Um experimento consiste em sortear um número entre 1 e 100 ao acaso. O evento de interesse C é “sortear um número maior que 50 ou um número ímpar, em uma tentativa específica”. O espaço amostral para este experimento é o conjunto de números $i = 1, \dots, 100$, e a probabilidade de sortear o número i é $P(A_i) = 1/100$, já que cada número tem a mesma probabilidade de ser sorteado. Seja A o evento consistindo de todos os números maiores que 50 e B o evento com todos os números ímpares, é claro que $P(A) = 0.5$ e $P(B) = 0.5$. Os dois eventos se sobrepõem, e o evento $A \cap B$ contém todos os números ímpares maiores que 50, com $P(A \cap B) = 0.25$. Usando a [Equação 1.5](#), a probabilidade de sortear um número maior que 50 ou um número ímpar é

$$P(C) = P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.5 + 0.5 - 0.25 = 0.25.$$

1.3 Probabilidade Condicional

A **probabilidade condicional** descreve a ocorrência de um evento A , sabendo que outro evento B também ocorreu. O condicionamento desempenha um papel proeminente na probabilidade e na estatística, pois a probabilidade de um evento pode depender de outro evento que se sabe ter ocorrido. A probabilidade condicional é indicada como $P(A|B)$ ou “ A dado B ”. A seguinte relação define a probabilidade condicional:

$$P(A \cap B) = P(A|B) \cdot P(B), \quad (1.6)$$

E pode ser expresso de forma equivalente como:

$$P(A|B) = \begin{cases} \frac{P(A \cap B)}{P(B)}, & \text{se } P(B) \neq 0 \\ 0, & \text{se } P(B) = 0 \end{cases} \quad (1.7)$$

Uma justificativa para essa definição é que a ocorrência de B significa que a probabilidade de ocorrência de A é proporcional à probabilidade de ocorrência de $A \cap B$. Além disso, o denominador da probabilidade condicional é $P(B)$, em vez da unidade, porque B é o conjunto de todos os possíveis resultados que se sabe terem acontecido. A situação também é ilustrada no painel direito da [Figura 1.1](#).

Exemplo 1.3

Vamos calcular a probabilidade de obter 8 como a soma de dois lançamentos de um dado, dado que o primeiro lançamento foi um 3. Para isso, é útil definir os seguintes dois eventos:

$$A = \{\text{A soma de dois lançamentos é 8}\},$$

$$B = \{\text{O primeiro lançamento mostra 3}\}.$$

O evento A é dado pelos resultados $(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)$, e como cada combinação tem uma probabilidade de $1/36$, $P(A) = 5/36$. A probabilidade do evento B é $P(B) = 1/6$ já que se relaciona aos resultados de apenas um lançamento do dado. Além disso, o evento $A \cap B$ ocorre se o primeiro lançamento for um 3 e a soma for 8, o que claramente só pode ocorrer se uma sequência de $(3, 5)$ acontecer, assim com probabilidade $P(A \cap B) = 1/36$. De acordo com a definição de probabilidade condicional, a probabilidade de interesse é

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/36}{1/6} = \frac{1}{6},$$

e na verdade apenas a combinação $(5, 3)$ resulta em uma soma de 8. A ocorrência de 3 no primeiro lançamento aumentou portanto a probabilidade de A de $P(A) = 5/36$ para $P(A|B) = 1/6$, já que nem todos os resultados do primeiro lançamento seriam igualmente propícios a uma soma de 8 em dois lançamentos.

1.4 Independência Estatística

O conceito de **independência estatística** entre eventos significa que a ocorrência de um evento não tem influência sobre a ocorrência de outros eventos. Considere, por exemplo, o lançamento de dois dados, um após o outro: o resultado de um dado é independente do outro e os dois lançamentos são ditos estatisticamente independentes. Por outro lado, considere o lançamento de dois dados e que haja interesse no seguinte par de eventos: o primeiro é o resultado do lançamento do dado 1 e o segundo é a soma dos lançamentos dos dados 1 e 2. É claro que o resultado do segundo evento depende do primeiro lançamento e os dois eventos não são independentes.

Dois eventos A e B são ditos estatisticamente independentes se

$$P(A \cap B) = P(A) \cdot P(B). \quad (1.8)$$

Essa definição segue diretamente da [Equação 1.6](#). De fato, se A e B forem estatisticamente independentes, então a probabilidade condicional é $P(A|B) = P(A)$, ou seja, a ocorrência de B não tem influência sobre a ocorrência de A . Portanto, a [Equação 1.8](#) é uma simples consequência da [Equação 1.6](#) quando os dois eventos são independentes. Alguns exemplos ilustram o significado dessa definição.

Exemplo 1.4

Vamos determinar a probabilidade de obter dois 3s ao lançar dois dados. Esse evento pode ser decomposto em dois eventos:

$$A = \{\text{o dado 1 mostra 3, e o dado 2 mostra qualquer número}\},$$

$$B = \{\text{o dado 2 mostra 3, e o dado 1 mostra qualquer número}\}.$$

É natural assumir que $P(A) = 1/6$, $P(B) = 1/6$, e afirmar que os dois eventos A e B são independentes por natureza, já que cada evento envolve um dado diferente, que não tem conhecimento do resultado do outro; o mesmo seria verdade também se o mesmo dado fosse lançado duas vezes. O evento de interesse é $C = A \cap B$, e a definição de probabilidade de dois eventos estatisticamente independentes leva a

$$P(C) = P(A \cap B) = P(A) \cdot P(B) = \frac{1}{36}.$$

Há apenas uma combinação em 36 que resulta em dois 3 consecutivos.

O exemplo acima destaca a importância de uma definição adequada, e às vezes estendida, de um evento. Quanto mais cuidadosa for a descrição do evento e do experimento do qual ele é derivado, mais fácil será fazer cálculos probabilísticos e a avaliação da independência estatística.

Exemplo 1.5

Vamos determinar se os seguintes eventos são estatisticamente independentes entre si:

$A = \{\text{o dado 1 mostra 3 e o dado 2 mostra qualquer número}\},$

$B = \{\text{a soma dos dois dados é 9}\}.$

O procedimento é calcular a probabilidade dos dois eventos e, em seguida, verificar se eles obedecem ou não à [Equação 1.8](#). Este cálculo ilustrará que os dois eventos não são estatisticamente independentes.

O evento A tem uma probabilidade $P(A) = 1/6$; para calcular a probabilidade do evento B , perceba que uma soma de 9 é dada pelas seguintes combinações de resultados dos dois lançamentos: (3, 6), (4, 5), (5, 4) e (6, 3), e portanto $P(B) = 4/36 = 1/9$. O evento $A \cap B$ é a situação em que ambos os eventos A e B ocorrem, o que corresponde à única combinação (3, 6); portanto, $P(A \cap B) = 1/36$. Como

$$P(A) \cdot P(B) = \frac{1}{6} \cdot \frac{1}{9} = \frac{1}{54} \neq P(A \cap B) = \frac{1}{36},$$

os dois eventos não são estatisticamente independentes. Essa conclusão significa que um evento influencia o outro, já que um 3 no primeiro lançamento certamente influencia a possibilidade de ambos os lançamentos terem um total de 9.

Existem duas condições necessárias (mas não suficientes) importantes para a independência estatística entre dois eventos. Essas propriedades podem ajudar a identificar se dois eventos são independentes.

1. Se $A \cap B = \emptyset$, A e B não podem ser independentes, a menos que um deles seja o conjunto vazio. Esta propriedade afirma que deve haver alguma sobreposição entre os dois eventos, caso contrário, não é possível que os eventos sejam independentes. Para A e B serem independentes, deve ser verdade que $P(A \cap B) = P(A) \cdot P(B)$, o que é zero por hipótese. Isso só pode ser verdade se $P(A) = 0$ ou $P(B) = 0$, o que por sua vez significa que $A = \emptyset$ ou $B = \emptyset$ como consequência dos axiomas de Kolmogorov.
2. Se $A \subset B$, então A e B não podem ser independentes, a menos que B seja o espaço amostral inteiro. Esta propriedade afirma que a sobreposição entre dois eventos não pode ser tal que um evento esteja incluído no outro, para que a independência estatística seja possível. Para A e B serem independentes, $P(A \cap B) = P(A) \cdot P(B) = P(A)$, dado que $A \subset B$. Isso só pode ser verdade se $B = \Omega$, já que $P(\Omega) = 1$.

Exemplo 1.6

Considere os seguintes dois eventos:

$A = \{\text{o dado 1 mostra 3 e o dado 2 mostra qualquer número}\},$

$B = \{\text{o dado 1 mostra 3 ou 2 e o dado 2 mostra qualquer número}\}.$

É claro que $A \subset B$, $P(A) = 1/6$ e $P(B) = 2/6 = 1/3$. O evento $A \cap B$ é assim idêntico a A e $P(A \cap B) = 1/6$. Portanto, $P(A \cap B) \neq P(A) \cdot P(B)$ e os dois eventos não são estatisticamente independentes. Este

resultado pode ser facilmente explicado pelo fato de que a ocorrência de A implica a ocorrência de B , o que é uma forte afirmação de dependência entre os dois eventos. A dependência entre os dois eventos também pode ser vista pelo fato de que a não ocorrência de B implica a não ocorrência de A .

1.5 Teorema da Probabilidade Total e Teorema de Bayes

Esta seção descreve dois teoremas que são de grande importância em várias situações práticas.

- **Teorema da Probabilidade Total.** Dado um evento B e um conjunto de eventos A_i que formam uma partição de acordo com as propriedades (1.1), temos que

$$P(B) = \sum_{i=1}^n P(B \cap A_i) = \sum_{i=1}^n P(B|A_i) \cdot P(A_i). \quad (1.9)$$

A primeira equação é verificada imediatamente dado que os $B \cap A_i$ são eventos mutuamente exclusivos tal que $B = \cup_i (B \cap A_i)$. A segunda equação deriva da aplicação da definição de probabilidade condicional.

O teorema da probabilidade total é útil quando a probabilidade de um evento B não pode ser facilmente calculada e é mais fácil calcular as probabilidades condicionais $P(B|A_i)$.

Exemplo 1.7

Considere o evento B consistindo em obter uma soma de 8 em dois lançamentos consecutivos de um dado. Cada lançamento pode ser dividido em 6 eventos A_i representando o dado mostrando i , com $P(A_i) = 1/6$. É claro que $P(B|A_i) = 1/6$ apenas para $i = 2, \dots, 6$ e nulo para $i = 1$, já que não há chance de uma soma de 8 se o primeiro lançamento for um 1. Segue-se que a soma no teorema da probabilidade total leva a:

$$P(B) = \sum_{i=2}^6 P(B|A_i) \cdot P(A_i) = \left(5 \times \frac{1}{6}\right) \times \frac{1}{6} = \frac{5}{36}$$

- **Teorema de Bayes.** Dado um evento B e um conjunto de eventos A_i que formam uma partição de acordo com as propriedades (1.1), temos que

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^n P(B \cap A_i)} \quad (1.10)$$

A prova é uma consequência imediata da definição de probabilidade condicional, [Equação 1.6](#), e do teorema da probabilidade total, [Equação 1.9](#).

O teorema de Bayes é frequentemente escrito em uma forma mais simples ao considerar apenas dois eventos, $A_i = A$ e B . Nesta situação simplificada, o teorema pode ser escrito como:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \quad (1.11)$$

Nesta forma, o teorema de Bayes é apenas uma afirmação de como a ordem de condicionamento entre dois eventos pode ser invertida. T. Bayes apresentou a primeira formulação deste teorema em sua publicação “*An Essay towards solving a Problem in the Doctrine of Chances*” ([Bayes and Price, 1763](#)). Bayes estava especificamente interessado no problema de estimar a probabilidade desconhecida p de um experimento binário que resultou em x sucessos e $n - x$ falhas. A [Equação 1.11](#) pode ser usada para abordar este problema com a seguinte interpretação dos eventos. O experimento B pode ser considerado como os dados coletados em um dado experimento; para o problema de Bayes, foi o fato de x dos n experimentos terem sido um sucesso. O evento A é um modelo que é usado para descrever os dados, no caso de Bayes, a probabilidade desconhecida p . De acordo, as probabilidades envolvidas no teorema de Bayes podem ser interpretadas da seguinte forma:

1. $P(B|A)$ é a probabilidade, ou **verossimilhança** \mathcal{L} , dos dados dado o modelo especificado. Para o problema de Bayes, esta é a probabilidade de ter x sucessos, se p fosse conhecido. Observe como $P(B|A)$ significa que o modelo A é dado, ou conhecido.
2. $P(A)$ é a probabilidade do modelo A , sem nenhum conhecimento dos dados. Este termo é interpretado como uma **probabilidade a priori**, ou o grau de crença de que o modelo é verdadeiro antes das medidas serem feitas. Para Bayes, esta é a probabilidade de que um valor específico de p seja o correto, antes dos dados experimentais (x sucessos em n experimentos) serem coletados. As probabilidades a priori devem ser baseadas no conhecimento quantitativo do experimento, mas também podem refletir a crença subjetiva do analista. Este passo na interpretação do teorema de Bayes explicitamente introduz um elemento de subjetividade que é característico da estatística Bayesiana.
3. $P(B)$ é a probabilidade de coletar o conjunto de dados B . Na prática, essa probabilidade atua como uma constante de normalização e seu valor numérico geralmente não tem consequência prática.
4. Finalmente, $P(A|B)$ é a **probabilidade a posteriori** do modelo após os dados terem sido coletados. A probabilidade a posteriori é o objetivo final da análise, pois descreve a probabilidade do modelo com base na coleta de dados. Para Bayes, esta era a estimativa buscada de p com base nos dados disponíveis.

Esta interpretação do teorema de Bayes é a base da estatística Bayesiana, e pode ser resumida como

$$\text{Probabilidade a posteriori} \propto \text{Verossimilhança} \times \text{Probabilidade a priori}.$$

O teorema de Bayes fornece uma maneira de atualizar o conhecimento a priori dos parâmetros do modelo dados as medidas, levando a estimativas a posteriori dos parâmetros. Uma característica chave da estatística Bayesiana é que o cálculo das probabilidades é baseado em uma probabilidade que pode depender de uma interpretação subjetiva do que é conhecido sobre o experimento antes de quaisquer medições serem feitas. Portanto, deve-se prestar muita atenção à atribuição de probabilidades a priori e ao efeito das posteriores nos resultados finais da análise. A Teoria da Probabilidade de [Jeffreys \(1998\)](#) é uma referência chave para a estatística Bayesiana e a importância das probabilidades a priori.

1.6 Problemas

- 1.1. No lançamento simultâneo de quatro moedas, descreva o espaço amostral e atribua a probabilidade de obter duas caras e duas coroas. Não há distinção entre as moedas.
- 1.2. Ao lançar simultaneamente dois dados independentes, determine a probabilidade de obter um número ímpar no primeiro lançamento ou uma soma total de 9 nos dois lançamentos.
- 1.3. Para um dado lançado, encontre a probabilidade de obter um número par ou maior que 4.
- 1.4. Ao lançar dois dados independentes, demonstre a não independência estatística entre “soma dos dois lançamentos é 8” e “o primeiro lançamento mostra 5”.
- 1.5. No lançamento de dois dados independentes, mostre a independência estatística entre “o primeiro lançamento é par” e “o segundo lançamento é par”.
- 1.6. Em uma caixa com 5 bolas, 3 vermelhas e 2 azuis, calcule (a) a probabilidade de retirar duas bolas vermelhas consecutivas e (b) a probabilidade de retirar duas bolas vermelhas consecutivas, sabendo que o primeiro sorteio foi uma bola vermelha. Assuma que após cada sorteio a bola é substituída na caixa.
- 1.7. Lançando simultaneamente dois dados independentes, calcule (a) a probabilidade do primeiro lançamento ser 1, dado que a soma dos lançamentos foi 5, (b) a probabilidade da soma ser 5, dado que o primeiro lançamento foi 1 e (c) a probabilidade do primeiro lançamento ser 1 e a soma ser 5. Finalmente, (d) confirme com o teorema de Bayes.

- 1.8.** Quatro moedas numeradas de 1 a 4 são lançadas simultaneamente e independentemente. Calcule (a) a probabilidade de obter a sequência ordenada cara-coroa-cara-coroa, (b) a probabilidade dessa sequência sabendo que duas moedas mostram cara e (c) a probabilidade de duas moedas mostrarem cara, sabendo que ocorreu a sequência cara-coroa-cara-coroa.

2 VARIÁVEIS ALEATÓRIAS E SUAS DISTRIBUIÇÕES

2.1 Variáveis Aleatórias

Uma variável aleatória é uma quantidade de interesse cujo valor verdadeiro é desconhecido. Para obter informações sobre uma variável aleatória, é necessário projetar e conduzir experimentos. É inerente a qualquer experimento que a variável aleatória de interesse nunca será conhecida exatamente. Em vez disso, a variável será caracterizada por uma **função de distribuição de probabilidade**, que determina qual é a probabilidade de que um determinado valor da variável aleatória ocorra. Repetir a medição geralmente aumenta o conhecimento da distribuição da variável: essa é a razão para querer medir a quantidade o máximo possível.

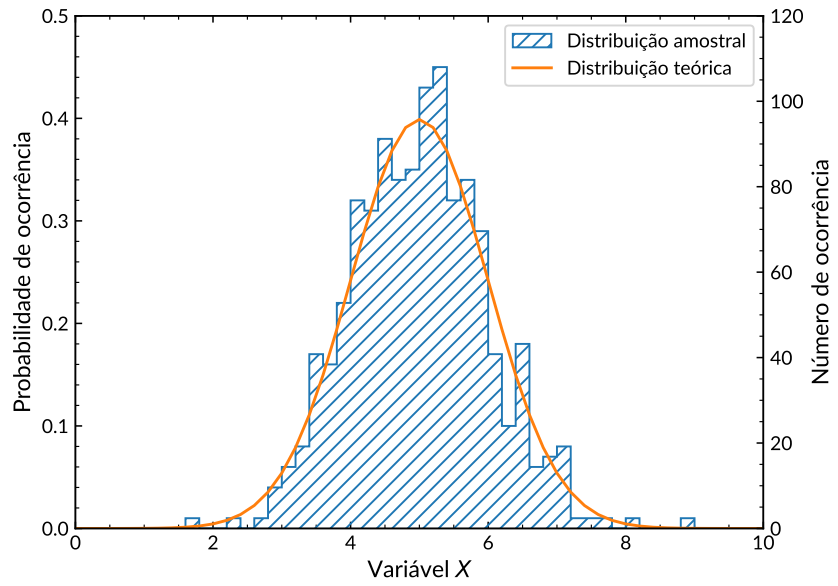
Exemplos de variáveis aleatórias são a massa da Terra, a altura da Torre Eiffel ou a voltagem de uma tomada elétrica. A natureza aleatória de praticamente todas as quantidades reside principalmente no fato de que nenhuma quantidade é conhecida exatamente por nós sem realizar um experimento, e que nenhum experimento é perfeito devido a limitações práticas ou até teóricas. Entre as razões práticas estão, por exemplo, limitações na precisão do aparelho de medição. As razões teóricas dependem da natureza da variável. Por exemplo, a medição da posição e velocidade de uma partícula subatômica é limitada pelo princípio da incerteza de Heisenberg (Heisenberg, 1927), que proíbe um conhecimento exato de ambas as quantidades mesmo na presença de um aparelho de medição perfeito.

O método geral para obter informações sobre uma variável aleatória X começa com um conjunto de medições x_i , garantindo que as medições sejam realizadas sob as mesmas condições experimentais. A partir dessas medições, obtém-se uma distribuição da frequência de ocorrência de todos os valores de X conhecida como a **distribuição amostral da variável**, que descreve a distribuição empírica dos valores coletados no experimento (Figura 2.1). Espera-se também que uma variável aleatória tenha uma distribuição teórica, por exemplo, Gaussiana, uniforme, etc., de acordo com a natureza da própria variável e o método de medição. Esta distribuição teórica é referida como **distribuição populacional** e representa a crença de que existe uma descrição ideal da variável aleatória. Como será mostrado nos capítulos seguintes, espera-se que a distribuição amostral se torne a distribuição populacional se um número infinito de medições for realizado, de tal forma que a aleatoriedade associada a um pequeno número de medições seja eliminada.

2.2 Funções de Distribuição de Probabilidade

É conveniente definir uma função que descreve a probabilidade de ocorrência da variável aleatória. Variáveis aleatórias discretas são descritas por uma **função de massa de probabilidade** (FMP) $f(x_i)$, onde $f(x_i)$ representa a probabilidade da variável ter um valor exato de x_i . Variáveis contínuas são descritas por uma **função de densidade de probabilidade** (FDP) $f(x)$, de modo que $f(x) dx$ é a probabilidade da variável estar no intervalo $[x, x + dx]$. Para simplicidade, a maioria das propriedades será ilustrada para variáveis contí-

Figura 2.1: Distribuição amostral de uma variável aleatória X a partir de 500 medições, obtida combinando as medições em intervalos de largura igual ($\Delta x = 0.2$). A forma da distribuição teórica depende da natureza do experimento e do número de medições, e neste exemplo é representada pela curva laranja.



nuas, com a compreensão de que elas também se aplicam a variáveis discretas com modificações simples, como a substituição de uma integral por uma soma. Em princípio, também é possível ter variáveis aleatórias mais complexas que são parcialmente contínuas e parcialmente discretas. Um tratamento mais completo das variáveis aleatórias pode ser encontrado em qualquer livro-texto dedicado ao assunto, como os de [Ross \(2019\)](#) ou [Siegrist \(2019\)](#). As funções de distribuição de probabilidade têm as seguintes propriedades-chave:

1. Elas são normalizadas para 1. Para variáveis contínuas, isso significa que

$$\int_{-\infty}^{\infty} f(x) dx = 1. \quad (2.1)$$

Para variáveis definidas em um subconjunto dos números reais, por exemplo, apenas valores $x \geq 0$ ou em um intervalo finito, $f(x)$ é definido como zero fora do domínio de definição da função. Para variáveis discretas, a integral é substituída por uma soma sobre todos os valores que a variável pode assumir.

2. A distribuição de probabilidade nunca pode ser negativa, $f(x) \geq 0$. Isso é uma consequência do axioma de Kolmogorov, que requer que a probabilidade seja não negativa.
3. A **função de distribuição acumulada** (FDA) $F(x)$, ou simplesmente **função de distribuição**,

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(\tau) d\tau, \quad (2.2)$$

representa a probabilidade de que a variável tenha um valor menor ou igual a x . $F(x)$ é uma função não decrescente de x que começa em zero e tem seu maior valor de um. Uma função relacionada é a **função de sobrevivência** $S(x) = 1 - F(x)$, representando a probabilidade de $X > x$.

Exemplo 2.1

A variável aleatória exponencial segue a FDP:

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0, \quad (2.3)$$

onde λ é um parâmetro que deve ser positivo. Portanto, a FDP é $f(x) = 0$ para valores negativos da variável. A FDA é dada por:

$$F(x) = \int_{-\infty}^x \lambda e^{-\lambda \tau} d\tau = \int_{-\infty}^0 \lambda e^{-\lambda \tau} d\tau + \int_0^x \lambda e^{-\lambda \tau} d\tau = 0 + \lambda \left(-\frac{1}{\lambda} e^{-\lambda \tau} \right)_0^x = 1 - e^{-\lambda x}. \quad (2.4)$$

Na [Figura 2.2](#) estão representadas a FDP $f(x)$ e a FDA $F(x)$ para uma variável exponencial com $\lambda = 0.5$.

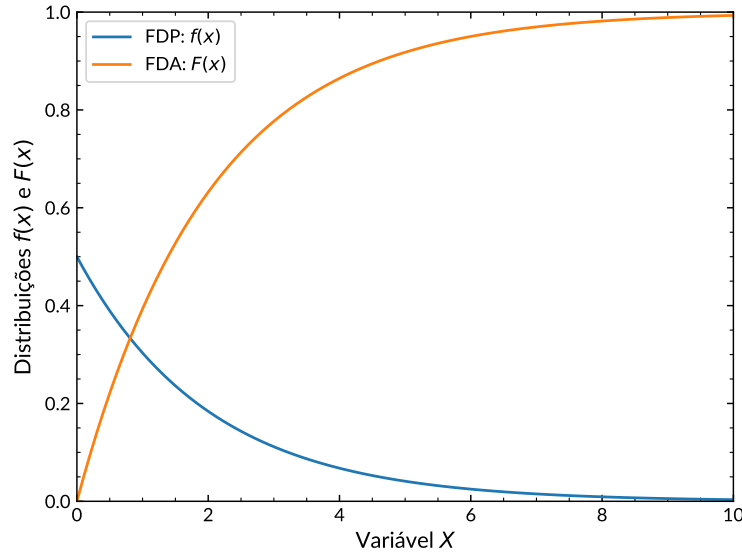


Figura 2.2: A função de distribuição $f(x)$ e a função de distribuição acumulada $F(x)$ para uma variável exponencial com $\lambda = 0.5$.

2.3 Expectância e Momentos de uma Função de Distribuição

A função de densidade de probabilidade $f(x)$ ou a função de distribuição $F(x)$ fornece uma descrição completa da variável aleatória X . É conveniente encontrar algumas quantidades que descrevem os recursos mais importantes da distribuição. A **expectância** (ou **valor esperado**) de uma função determinística $g(x)$ da variável aleatória é definida por

$$\mathbb{E}[g(X)] = \int g(x)f(x) dx. \quad (2.5)$$

O significado dessa expressão é que cada valor $g(x)$ é ponderado de acordo com a probabilidade de ocorrência de x , que é $f(x)$. A expectância é um operador linear, ou seja, satisfaz a propriedade

$$\mathbb{E}[ag_1(x) + bg_2(x)] = a \mathbb{E}[g_1(x)] + b \mathbb{E}[g_2(x)]$$

devido à propriedade de linearidade da integral.

O **momento** de ordem n é definido como

$$\mu_n = \mathbb{E}[X^n] = \int x^n f(x) dx. \quad (2.6)$$

Portanto, o momento μ_n é a expectância da função $g(x) = x^n$. É possível demonstrar, embora além do escopo deste livro, que o conhecimento dos momentos de todas as ordens é suficiente para determinar unicamente a função de distribuição ([Wilks, 1947](#)). Este é um fato importante, pois transfere o problema de determinar a função de distribuição para o de determinar pelo menos alguns de seus momentos. Além disso, várias funções de distribuição têm apenas alguns momentos diferentes de zero, o que torna a tarefa ainda mais gerenciável.

Os momentos de uma distribuição são quantidades teóricas que podem ser calculadas a partir de $f(x)$, sem qualquer referência a medidas. A seguir, são descritos dois dos momentos mais comumente utilizados, a média e a variância, e as quantidades amostrais que os aproximam, a média amostral e a variância amostral.

2.3.1 A Média e a Média Amostral

O momento de primeira ordem também é conhecido como **média** ou expectância da variável aleatória,

$$\mu = \mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x) dx. \quad (2.7)$$

A média μ é um número característico que representa um valor médio de X , obtido ponderando todos os valores possíveis da variável por sua função de distribuição. Portanto, a média é uma medida muito simples e conveniente da variável aleatória.

Para estimar a média da variável aleatória X , considere N medidas x_i , com $i = 1, \dots, N$. Cada medida x_i pode ser considerada como uma amostra retirada aleatoriamente da distribuição populacional de X . A **média amostral** é definida como

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i. \quad (2.8)$$

É evidente que diferentes amostras de tamanho N não resultarão no mesmo valor de \bar{x} , devido à aleatoriedade das medidas. Isso indica que \bar{x} não é um número fixo, mas é ele próprio uma variável aleatória que deve ser escrita como

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i.$$

Esta nova equação é formalmente idêntica à (2.8) exceto pelo uso de letras maiúsculas, indicando que as quantidades são variáveis aleatórias, não apenas números. Portanto, é razoável perguntar se essa nova variável aleatória \bar{X} tem a mesma expectância que a variável populacional X em si. Essa pergunta pode ser facilmente respondida considerando que as variáveis X_i são distribuídas de maneira idêntica da mesma forma que X , já que representam medidas aleatórias de X . A expectância de \bar{X} é então:

$$\mathbb{E}[\bar{X}] = \mathbb{E}\left[\frac{1}{N} \sum_{i=1}^N X_i\right] = \frac{1}{N} \sum_{i=1}^N \mathbb{E}[X_i] = \frac{N\mu}{N} = \mu.$$

A média amostral tem a mesma expectância que a variável populacional X , e portanto \bar{X} é considerado um **estimador não tendencioso** de X . Esta é uma propriedade muito desejável, dado que a média amostral foi projetada exatamente com o propósito de estimar a variável populacional X com suas medidas. Também se espera que, à medida que o número de medidas N aumenta, a média amostral se aproxime cada vez mais da média populacional com maior precisão. Isso é mostrado na seção seguinte.

2.3.2 A Lei dos Grandes Números

A **lei dos grandes números** afirma que a média amostral converge para a média populacional conforme o tamanho da amostra aumenta, e é um dos teoremas fundamentais da probabilidade. Considere N variáveis aleatórias X_i que são identicamente distribuídas com μ sendo a média comum delas. A lei pode ser enunciada como:

$$\lim_{N \rightarrow \infty} \frac{X_1 + \dots + X_N}{N} = \mu, \quad (2.9)$$

Isso implica que a média amostral \bar{X} tende para a média μ , que é um número determinístico e não uma variável aleatória. A Equação 2.9 é uma afirmação muito forte porque mostra que, assintoticamente, a soma das variáveis aleatórias se torna uma constante igual à média amostral das N variáveis, ou N medidas. Embora não haja indicação sobre quão grande N deve ser para alcançar esse objetivo, isso ainda é um resultado-chave para estabelecer o comportamento assintótico das variáveis aleatórias. É útil apontar que existem diferentes versões desta lei, de acordo com o método pelo qual essa convergência é estabelecida. Propriedades matemáticas desta lei e sua derivação podem ser encontradas em livros de teoria da probabilidade, como Kolmogorov (1950); Ross (2019).

Essa lei também tem consequências importantes para funções de variáveis aleatórias. Dada uma função $g(x)$, seria conveniente estimar seu valor esperado $\mathbb{E}[g(X)]$ a partir das N medidas das variáveis X_i . De acordo com a lei dos grandes números, é possível dizer que

$$\lim_{N \rightarrow \infty} \frac{g(X_1) + \dots + g(X_N)}{N} = \mathbb{E}[g(X)]. \quad (2.10)$$

Essa expressão afirma que um grande número de medidas das variáveis X_i pode ser usado para medir $\mathbb{E}[g(X)]$, contornando completamente a função de densidade de probabilidade da função $g(X)$. Essa propriedade será útil ao estudar a distribuição de funções de uma variável aleatória.

2.3.3 A Variância e a Variância Amostral

A **variância** é a expectância do quadrado da diferença entre X e sua média,

$$\text{Var}(X) = \mathbb{E}[(X - \mu)^2] = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx \quad (2.11)$$

e geralmente é indicada como $\text{Var}(X) = \sigma^2$. A raiz quadrada da variância é referida como o **desvio padrão** σ , e é uma medida comum da diferença média de uma determinada medida x_i da média da variável aleatória.

A principal razão para definir a média da diferença de uma medida de sua média em termos de um momento de segunda ordem é que a expectância da diferença $X - \mu$ é sempre zero:

$$\mathbb{E}[X - \mu] = \mathbb{E}[X] - \mathbb{E}[\mu] = \mu - \mu = 0$$

Portanto, a diferença de uma variável aleatória não é comumente utilizada em estatísticas, uma vez que sua expectância é nula. Observe que as dimensões físicas dos momentos da ordem n são aquelas da variável elevada à potência n . Por exemplo, se X for medida em metros, a variância é medida em metros quadrados (m^2), enquanto o desvio padrão tem as mesmas dimensões que a variável e sua média.

Usando a propriedade linear da expectância, é simples mostrar que a seguinte propriedade se aplica:

$$\text{Var}(X) = \mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2 - 2X\mu + \mu^2] = \mathbb{E}[X^2] - 2\mu^2 + \mu^2 = \mathbb{E}[X^2] - \mu^2. \quad (2.12)$$

Essa relação é muito conveniente para calcular a variância a partir dos momentos de primeira e segunda ordem. Outra propriedade útil da variância, que decorre do fato de que a variância é um momento de segunda ordem, é

$$\text{Var}(aX) = a^2 \cdot \text{Var}(X), \quad (2.13)$$

onde a é uma constante. O desvio e a variância são momentos calculados em relação à média, e são referidos como **momentos centrais**.

A **variância amostral** é definida como

$$s^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2, \quad (2.14)$$

uma nova variável aleatória que se destina a estimar a variância populacional σ^2 por meio de N medições aleatórias. O fator $N - 1$, e não N , no denominador é necessário para garantir que a variância da amostra seja um estimador não tendencioso da variância populacional. Uma prova dessa propriedade é fornecida na [Seção 2.5](#).

2.4 Covariância e Correlação entre Variáveis Aleatórias

É comum medir mais de uma variável aleatória em um experimento dado. As variáveis frequentemente estão relacionadas entre si, e, portanto, é necessário definir uma medida de como uma variável afeta a medição das outras. Considere a medição tanto do comprimento de um lado de um quadrado quanto de sua área; é claro que as duas quantidades estão relacionadas de tal forma que a mudança de uma quantidade afeta a outra da mesma maneira, ou seja, uma mudança positiva no comprimento do lado resulta em uma mudança positiva na área. Nesse caso, o comprimento e a área serão ditos ter uma correlação positiva. Esta seção introduz conceitos-chave para estudar duas ou mais variáveis aleatórias simultaneamente.

2.4.1 Distribuição de Probabilidade Conjunta e Momentos de Duas Variáveis Aleatórias

Quando duas variáveis são medidas ao mesmo tempo, deseja-se saber a probabilidade de um par de medidas dado para as duas variáveis. Essa informação é fornecida pela **função de distribuição de probabilidade conjunta**, indicada como $h(x, y)$, onde $h(x, y) dx dy$ é a probabilidade de que as duas variáveis X e Y estejam em um intervalo bidimensional de tamanho $dx dy$ em torno do valor (x, y) . Essa função bidimensional

pode ser representada experimentalmente por meio de sua distribuição amostral, da mesma forma que as distribuições unidimensionais.

Normalmente, é conveniente descrever o comportamento de uma variável de cada vez, mesmo que o experimento envolva mais de uma variável. O processo de considerar apenas uma variável de cada vez é chamado **marginalização**. A função de densidade de probabilidade marginal de X , ou seja, a função de distribuição de X isoladamente, é obtida a partir da função de distribuição de probabilidade conjunta como:

$$f_X(x) = \int_{-\infty}^{\infty} h(x, y) dy, \quad (2.15)$$

com uma equação equivalente para a distribuição marginal de Y . A expectância de X é definida como:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x h(x, y) dx dy = \mu_x, \quad (2.16)$$

e similarmente a expectância de Y é igual a:

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y h(x, y) dx dy = \mu_y.$$

A combinação dessas duas equações leva ao resultado:

$$\mathbb{E}[X + Y] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x + y) h(x, y) dx dy = \mathbb{E}[X] + \mathbb{E}[Y],$$

o que significa que a expectância da soma de duas variáveis é igual à soma das expectâncias. Este é um resultado conveniente que se aplica a todas as variáveis aleatórias e pode ser generalizado para mais de duas variáveis, sendo geralmente referido como a propriedade linear da expectância. Em particular, é útil ressaltar que a propriedade linear se aplica independentemente da independência estatística entre as variáveis, que é introduzida mais adiante nesta seção.

A variância é definida de forma similar como:

$$\text{Var}(X) = \mathbb{E}[(X - \mu_x)^2] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_x)^2 h(x, y) dx dy = \sigma_x^2. \quad (2.17)$$

Essas equações reconhecem o fato de que a outra variável, neste caso Y , é de fato parte do experimento, mas é considerada não interessante para o cálculo em questão. Portanto, a variável não interessante é integrada (ou somada) sobre, ponderada pela sua função de distribuição de probabilidade.

A **covariância** de duas variáveis aleatórias é definida como:

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mu_x)(Y - \mu_y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_x)(y - \mu_y) h(x, y) dx dy \quad (2.18)$$

e geralmente é representada como $\text{Cov}(X, Y) = \sigma_{xy}^2$. A covariância é a expectância do produto dos desvios das duas variáveis, e a ordem das duas variáveis é irrelevante, de modo que $\text{Cov}(X, Y) = \text{Cov}(Y, X)$. A covariância é positiva se, em média, um desvio positivo de X for acompanhado por um desvio positivo de Y , ou se dois desvios negativos forem prováveis de ocorrer simultaneamente, de modo que o integrando seja uma quantidade positiva. Se, por outro lado, as duas variáveis tendem a ter desvios de sinais opostos, a covariância será negativa. A covariância, assim como a média e a variância, é uma quantidade populacional que pode ser calculada a partir da distribuição teórica das variáveis aleatórias. A covariância também pode ser calculada desenvolvendo o lado esquerdo da [Equação 2.18](#),

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY - X\mathbb{E}[Y] - \mathbb{E}[X]Y + \mathbb{E}[X]\mathbb{E}[Y]] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y],$$

e possui as seguintes propriedades lineares:

$$\text{Cov}(X + a, Y + b) = \text{Cov}(X, Y) \quad \text{e} \quad \text{Cov}(aX + b, cY + d) = ac \cdot \text{Cov}(X, Y),$$

o que pode ser visto imediatamente a partir da definição de covariância e da propriedade linear da expectativa. Outra propriedade útil é:

$$\text{Cov}(X + Z, Y) = \text{Cov}(X, Y) + \text{Cov}(Z, Y),$$

que pode ser comprovada expandindo as expectativas, e pode ser generalizada para mais de duas variáveis.

A **covariância amostral** para uma coleção de N pares de medidas é calculada como:

$$s_{xy}^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y}), \quad (2.19)$$

usando uma equação similar à variância amostral. É possível mostrar que o denominador $(N-1)$ é necessário para garantir que a covariância amostral seja um estimador não tendencioso da covariância populacional. Essa derivação é fornecida na [Seção 2.5](#).

A **correlação** é simplesmente uma versão normalizada da covariância:

$$\text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \frac{\sigma_{xy}^2}{\sigma_x \sigma_y}, \quad (2.20)$$

e pode ser indicada como $\text{Cor}(X, Y) = \rho$. A **desigualdade de Cauchy-Schwarz** é frequentemente declarada como:

$$\mathbb{E}[|XY|] \leq \sqrt{(\mathbb{E}[X^2])\mathbb{E}[Y^2]}], \quad (2.21)$$

e pode ser usada para mostrar que $-1 \leq \rho \leq 1$. Quando o coeficiente de correlação é zero, as duas variáveis são consideradas **não correlacionadas**. O **coeficiente de correlação amostral** r é calculado como:

$$r = \frac{s_{xy}^2}{s_x s_y} = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^N (y_i - \bar{y})^2}}, \quad (2.22)$$

onde s_{xy}^2 é a covariância amostral e s_x^2 e s_y^2 são as variâncias amostrais das duas variáveis, e os denominadores $(N-1)$ tanto da covariância quanto das variâncias foram cancelados. A desigualdade de Cauchy-Schwarz pode ser usada novamente para mostrar que $-1 \leq r \leq 1$, assim como para a correlação populacional.

2.4.2 Independência Estatística de Variáveis Aleatórias

A independência entre eventos foi descrita e quantificada no [Capítulo 1](#), onde foi mostrado que dois eventos são independentes apenas quando a probabilidade de ambos os eventos ocorrerem (ou sua interseção) é o produto das probabilidades individuais. O conceito é estendido aqui para variáveis aleatórias, definindo duas variáveis aleatórias como independentes se e somente se a função de distribuição de probabilidade conjunta puder ser fatorada na seguinte forma:

$$h(x, y) = f(x) \cdot g(y), \quad (2.23)$$

onde as duas funções $f(x)$ e $g(y)$ se tornam as funções de distribuição de probabilidade das duas variáveis aleatórias. Quando duas variáveis são independentes, a função de distribuição de probabilidade marginal de cada variável, obtida por marginalização sobre a outra variável (ver [Equação 2.15](#)), é

$$f_X(x) = \int_{-\infty}^{\infty} f(x) \cdot g(y) dy = f(x), \quad (2.24)$$

mostrando que $f(x)$ é de fato a função de distribuição marginal de X . Um resultado idêntico também é aplicável a $g(y)$.

É importante ressaltar que independência entre variáveis aleatórias e falta de correlação não são propriedades equivalentes. Independência, que é uma propriedade das funções de distribuição (2.23), é uma propriedade muito mais forte do que a falta de correlação, que é baseada em uma afirmação que envolve apenas momentos. Pode ser provado que independência implica falta de correlação, mas não o contrário.

Podemos demonstrar isso calculando a covariância de duas variáveis aleatórias independentes, com função de distribuição conjunta $h(x, y)$. A covariância é dada por:

$$\sigma_{xy} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_x)(y - \mu_y)h(x, y) dx dy = \int_{-\infty}^{\infty} (x - \mu_x)f(x) dx \int_{-\infty}^{\infty} (y - \mu_y)g(y) dy = 0,$$

onde cada integral se anula como a expectância da diferença de uma variável aleatória.

Como um contra-exemplo do fato de que variáveis dependentes podem ter um fator de correlação não nulo, considere o caso de uma variável aleatória X com uma distribuição $f(x)$ que é simétrica em torno da origem, e outra variável $Y = X^2$. Elas não podem ser independentes pois estão relacionadas funcionalmente, mas será mostrado que a covariância delas é zero. A simetria em torno de zero implica $\mu_x = 0$. A média de Y é $\mathbb{E}[Y] = \mathbb{E}[X^2] = \sigma_x^2$ já que a média de X é nula (veja Equação 2.12). A partir disso, a covariância é dada por:

$$\text{Cov}(X, Y) = \mathbb{E}[X(Y - \sigma_x^2)] = \mathbb{E}[X^3 - X\sigma_x^2] = \mathbb{E}[X^3] = 0$$

devido à simetria de $f(x)$. Portanto, as duas variáveis X e X^2 são não correlacionadas, mas não são independentes.

A variância da soma de duas ou mais variáveis independentes possui uma propriedade muito conveniente que é útil apresentar neste ponto. Considere a soma de N variáveis independentes, $X = X_1 + \dots + X_N$. A variância de X possui a seguinte propriedade aditiva:

$$\text{Var}(X) = \sum_{i=1}^N \text{Var}(X_i).$$

Para provar isso, considere a função de distribuição conjunta de N variáveis independentes que pode ser fatorada de acordo com uma generalização simples da Equação 2.23,

$$h(x_1, \dots, x_N) = f_1(x_1) \dots f_N(x_N),$$

onde cada $f_i(x_i)$ representa a função de distribuição marginal da variável X_i , com média populacional μ_i . A variância da soma pode ser escrita como

$$\text{Var}(X) = \mathbb{E}[(X - \mu_X)^2] = \mathbb{E}[(X_1 + \dots + X_N - (\mu_1 + \dots + \mu_N))^2],$$

onde μ_X é a média da soma das variáveis, que é sempre igual à soma das médias. O quadrado pode ser expandido como

$$\text{Var}(X) = \sum_{i=1}^N \mathbb{E}[(X_i - \mu_i)^2] + 2 \sum_{i>j} \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)]$$

onde as expectâncias de todos os termos de produto cruzado na segunda soma são a covariância entre variáveis, que são todas nulas devido à independência delas. Isso leva à propriedade linear da variância, que se aplica apenas no caso de variáveis independentes.

2.5 A Expectância da Variância Amostral e Covariância Amostral

A variância amostral e a covariância amostral são valores que calculamos usando várias medições independentes de uma ou duas variáveis aleatórias. Assim como a média amostral, elas também são tratadas como variáveis aleatórias, o que significa que têm uma distribuição com seus momentos correspondentes. Por exemplo, para a média amostral, sua expectância foi calculada diretamente como $\mathbb{E}[x] = \mu$, onde μ é

a média populacional da variável X . Isso mostra que a média amostral é um estimador imparcial da média populacional.

Considerações semelhantes precisam ser aplicadas à variância amostral e à covariância amostral, e assegurar se suas expectâncias concordam com as quantidades populacionais que foram projetadas para aproximar. As medidas que contribuem para essas variáveis amostrais são do tipo x_i , ou pares (x_i, y_i) , para $i = 1, \dots, N$. As N medidas são amostras independentes da variável X ou do par de variáveis X, Y . Essa propriedade de independência entre as amostras é fundamental para entender a expectância das variáveis amostrais em questão. As medidas x_i são ditas ser **variáveis independentes e identicamente distribuídas** (frequentemente abreviadas como IID), ou seja, amostras independentes da mesma distribuição populacional $f(x)$ com média populacional μ . O mesmo se aplica ao caso bidimensional de pares (x_i, y_i) , que são amostras independentes de uma distribuição bidimensional populacional $h(x, y)$.

A seguinte derivação mostra que a expectância da variância amostral, conforme definida na Equação 2.14, é igual à variância populacional da variável X , e que a expectância da covariância amostral, conforme definida na Equação 2.19, é igual à covariância populacional. Essas propriedades mostram que a variância amostral e a covariância amostral são estimadores não enviesados das quantidades populacionais correspondentes. O fator de $N - 1$ no denominador deriva principalmente do fato de que as médias amostrais foram estimadas a partir dos dados, introduzindo uma fonte adicional de incerteza. Pode-se pensar nesta situação da seguinte maneira: as N medidas, seja de uma única variável ou de um par de variáveis, são usadas primeiro para estimar a(s) média(s) amostral(is), reduzindo assim o número efetivo ou independente de medidas para $N - 1$.

Vamos então obter a expectância da variância amostral da seguinte maneira:

$$\begin{aligned}\mathbb{E}[s^2] &= \mathbb{E}\left[\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2\right] \\ &= \left(\frac{1}{N-1}\right) \mathbb{E}\left[\sum_{i=1}^N (x_i - \mu + \mu - \bar{x})^2\right] \\ &= \left(\frac{1}{N-1}\right) \mathbb{E}\left[\sum_{i=1}^N (x_i - \mu)^2 + \sum_{i=1}^N (\mu - \bar{x})^2 + 2(\mu - \bar{x}) \sum_{i=1}^N (x_i - \mu)\right]\end{aligned}$$

O termo $\mathbb{E}[\sum_{i=1}^N (\mu - \bar{x})^2]$ é N vezes a variância da média amostral. Esta variância pode ser calculada observando que as N medidas x_i são, na verdade, variáveis aleatórias independentes com a mesma distribuição e mesma média μ . A propriedade linear da variância, que só se aplica a variáveis independentes, leva a:

$$\text{Var}(\bar{x}) = \frac{\text{Var}(x_1 + \dots + x_n)}{N^2} = \frac{\sigma^2}{N},$$

onde σ^2 é a variância populacional de cada uma das variáveis X_i . O último termo na equação é $\sum_{i=1}^N (x_i - \mu) = N(\bar{x} - \mu)$, portanto:

$$\begin{aligned}\mathbb{E}[s^2] &= \frac{1}{N-1} \left(\mathbb{E}\left[\sum_{i=1}^N (x_i - \mu)^2\right] + N \mathbb{E}[(\mu - \bar{x})^2] + 2N \mathbb{E}[(\mu - \bar{x})(\bar{x} - \mu)] \right) \\ &= \frac{1}{N-1} \left[N\sigma^2 + N \left(\frac{\sigma^2}{N}\right) - 2N \mathbb{E}[(\mu - \bar{x})^2] \right].\end{aligned}$$

Como $\mathbb{E}[(\mu - \bar{x})^2]$ é novamente a variância da média amostral, a expectância da variância amostral é finalmente calculada como:

$$\mathbb{E}[s^2] = \frac{1}{N-1} \left[N\sigma^2 + N \left(\frac{\sigma^2}{N}\right) - 2N \left(\frac{\sigma^2}{N}\right) \right] = \sigma^2.$$

Antes de calcular a expectância da covariância amostral, é necessário calcular a covariância entre as duas médias amostrais \bar{x} e \bar{y} . De acordo com as propriedades da covariância:

$$\text{Cov}(\bar{x}, \bar{y}) = \frac{1}{N^2} \left(\sum_{i=1}^N \text{Cov}(x_i, y_i) + 2 \sum_{i \neq j} \text{Cov}(x_i, y_j) \right).$$

A independência entre pares de medidas significa que $\text{Cov}(x_i, y_j) = 0$ quando $i \neq j$. Além disso, como $\text{Cov}(x_i, y_i) = \sigma_{xy}^2$ é a covariância populacional, obtém-se o resultado:

$$\text{Cov}(\bar{x}, \bar{y}) = \frac{\sigma_{xy}^2}{N}.$$

A covariância amostral pode ser escrita como:

$$s_{xy}^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{N-1} \left(\sum_{i=1}^N x_i y_i - N \bar{x} \bar{y} \right).$$

Uma vez que,

$$\mathbb{E}[x_i y_i] = \sigma_{xy}^2 + \mu_x \mu_y \quad \text{e} \quad \mathbb{E}[\bar{x} \bar{y}] = \text{Cov}(\bar{x}, \bar{y}) + \mathbb{E}[\bar{x}] \mathbb{E}[\bar{y}],$$

a expectância da covariância amostral é finalmente calculada como:

$$\begin{aligned} \mathbb{E}[s_{xy}^2] &= \left(\frac{1}{N-1} \right) \mathbb{E} \left[\sum_{i=1}^N x_i y_i - N \bar{x} \bar{y} \right] \\ &= \left(\frac{1}{N-1} \right) [N(\sigma_{xy}^2 + \mu_x \mu_y) - N(\text{Cov}(\bar{x}, \bar{y}) + \mathbb{E}[\bar{x}] \mathbb{E}[\bar{y}])] \\ &= \left(\frac{1}{N-1} \right) \left[N(\sigma_{xy}^2 + \mu_x \mu_y) - N \left(\frac{\sigma_{xy}^2}{N} + \mu_x \mu_y \right) \right] \\ &= \sigma_{xy}^2. \end{aligned}$$

Isso mostra que a covariância amostral é um estimador não enviesado da covariância populacional.

2.6 Problemas

2.1. Considere a distribuição exponencial

$$f(x) = \lambda e^{-\lambda x},$$

onde $\lambda > 0$ e $x \geq 0$. Mostre que a distribuição é devidamente normalizada e calcule a média e variância.

2.2. Considere a média amostral como uma variável aleatória definida por

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i,$$

onde x_i são variáveis aleatórias independentes e idênticas com média μ e variância σ^2 . Mostre que a variância de \bar{x} é igual a σ^2/N .

2.3. Nos experimentos pioneiros conduzidos por J.J. Thomson no final do século XIX, a intenção era medir a relação entre a massa e a carga de uma nova entidade subatômica, posteriormente denominada elétron. Este experimento não apenas introduziu um método revolucionário, mas também transformou nossa compreensão da física e das ciências naturais, ao demonstrar que essa nova partícula era consideravelmente mais leve do que o portador de carga conhecido na época, o próton. A abordagem experimental envolveu a deflexão de raios catódicos negativamente carregados por um campo magnético H em um tubo de vácuo (Thomson, 1897). Thomson procurou determinar a massa m dessas partículas carregadas. Assim, o experimento foi baseado na medição de quantidades específicas: W , a energia cinética das partículas; $Q = Ne$, a carga elétrica total das partículas (onde N é o número de partículas e e é a carga elementar); e $I = HR$, onde R é o raio de curvatura da trajetória das partículas no campo magnético H . Alguns dos dados experimentais coletados por Thomson são apresentados

nas Tabela 2.1 e 2.2, onde “gás” se refere ao tipo de gás utilizado nos tubos de vácuo. Utilizando esses conjuntos de dados para os Tubos 1 e 2 separadamente, determine as médias amostrais e variâncias das variáveis aleatórias W/Q e I , bem como a covariância e os coeficientes de correlação entre W/Q e I .

Gás	W/Q	I	m/e	ν
Ar	4.6×10^{11}	230	0.57×10^{-7}	4×10^9
Ar	1.8×10^{12}	350	0.34×10^{-7}	1×10^{10}
Ar	6.1×10^{11}	230	0.43×10^{-7}	5.4×10^9
Ar	2.5×10^{12}	400	0.32×10^{-7}	1.2×10^{10}
Ar	5.5×10^{11}	230	0.48×10^{-7}	4.8×10^9
Ar	1×10^{12}	285	0.4×10^{-7}	7×10^9
Ar	1×10^{12}	285	0.4×10^{-7}	7×10^9
Hidrogênio	6×10^{12}	205	0.35×10^{-7}	6×10^9
Hidrogênio	2.1×10^{12}	460	0.5×10^{-7}	9.2×10^9
Ácido carbônico	8.4×10^{11}	260	0.4×10^{-7}	7.5×10^9
Ácido carbônico	1.47×10^{12}	340	0.4×10^{-7}	8.5×10^9
Ácido carbônico	3.0×10^{12}	480	0.39×10^{-7}	1.3×10^{10}

Tabela 2.1: Dados das medições de Thomson do Tubo 1.

Gás	W/Q	I	m/e	ν
Ar	2.8×10^{11}	175	0.53×10^{-7}	3.3×10^9
Ar	2.8×10^{11}	175	0.47×10^{-7}	4.1×10^9
Ar	3.5×10^{11}	181	0.47×10^{-7}	3.8×10^9
Hidrogênio	2.8×10^{11}	175	0.53×10^{-7}	3.3×10^9
Ar	2.5×10^{11}	160	0.51×10^{-7}	3.1×10^9
Ácido carbônico	2.0×10^{11}	148	0.54×10^{-7}	2.5×10^9
Ar	1.8×10^{11}	151	0.63×10^{-7}	2.3×10^9
Hidrogênio	2.8×10^{11}	175	0.53×10^{-7}	3.3×10^9
Hidrogênio	4.4×10^{11}	201	0.46×10^{-7}	4.4×10^9
Ar	2.5×10^{11}	176	0.61×10^{-7}	2.8×10^9
Ar	4.2×10^{11}	200	0.48×10^{-7}	4.1×10^9

Tabela 2.2: Dados das medições de Thomson do Tubo 2.

2.4. Ainda usando o experimento de J.J. Thomson, verifique a afirmação de que

O valor de m/e é independente da natureza do gás.

Você pode fazer isso calculando a média amostral e o desvio padrão para as medições em cada gás (ar, hidrogênio e ácido carbônico), e então testando se as três medições concordam entre si dentro de seus desvios padrão.

2.5. Calcule a covariância amostral e o coeficiente de correlação para o seguinte conjunto de dados: $(0, 2)$, $(2, 5)$, $(1, 4)$, $(3, 1)$.

3 TRÊS DISTRIBUIÇÕES FUNDAMENTAIS: BINOMIAL, GAUSSIANA E POISSON

3.1 A Distribuição Binomial

Um experimento **binário** só pode ter dois resultados possíveis, que podem ser interpretados como **sucesso** ou **falha**, como por exemplo o lançamento de uma moeda. Mesmo experimentos complexos com um maior número de resultados possíveis podem ser descritos como binários, quando se está simplesmente interessado na ocorrência de um evento específico A , ou sua não ocorrência, A^c . Por exemplo, o lançamento de um dado pode ser interpretado como um experimento binário se estivermos interessados na ocorrência do número seis, ou na ocorrência de qualquer outro número. É de fundamental importância em estatística determinar as propriedades dos experimentos binários, e a distribuição do número de sucessos quando o experimento é repetido várias vezes sob as mesmas condições experimentais. Uma variável aleatória binária é geralmente referida como **variável de Bernoulli**.

3.1.1 Derivação da Distribuição Binomial

Considere um experimento binário caracterizado por uma probabilidade de sucesso p e, portanto, uma probabilidade de falha $q = 1 - p$. As probabilidades p e q são determinadas de acordo com a teoria da probabilidade e são assumidas como conhecidas para o experimento em questão. Quando o experimento é repetido N vezes nas mesmas condições experimentais, é interessante calcular a probabilidade de obter n sucessos em N tentativas. A ordem em que os n sucessos ocorrem não é relevante; por exemplo, considere lançar uma moeda quatro vezes e estar interessado na probabilidade de exatamente duas dessas jogadas mostrarem cara. O cálculo da probabilidade binomial segue estes passos:

1. **Probabilidade de uma sequência ordenada.** A probabilidade de ter n sucessos e, portanto, $N - n$ falhas ocorrendo em uma ordem específica é dada por:

$$P(\text{sequência específica de } n \text{ sucessos}) = p^n \times q^{N-n}. \quad (3.1)$$

Este resultado pode ser visto usando a propriedade de independência entre os N eventos, de modo que as probabilidades individuais podem simplesmente ser multiplicadas.

2. **Número de sequências ordenadas ou permutações.** Comece contando quantas sequências ordenadas existem que têm n sucessos em N tentativas. Para começar, cada uma das N tentativas pode resultar no “primeiro” sucesso, e portanto existem N possibilidades para qual tentativa será o primeiro sucesso. Continuando para o “segundo” sucesso, sobram apenas $N - 1$ possibilidades para qual tentativa será o segundo sucesso, e assim por diante. Este método de contagem de sequências rotula cada sucesso como primeiro, segundo, etc., e leva ao número de permutações de n sucessos em N tentativas:

$$\text{Perm}(n, N) = N \cdot (N - 1) \cdot (N - n + 1) = \frac{N!}{(N - n)!}. \quad (3.2)$$

Exemplo 3.1

Considere o caso de $n = 2$ sucessos em $N = 4$ tentativas. De acordo com a [Equação 3.2](#), o número de permutações é $4!/(4 - 2)! = 4!/2! = 12$. As 12 sequências ordenadas que resultam em 2 sucessos em 4 tentativas estão listadas na [Tabela 3.1](#). O símbolo S_1 denota o “primeiro sucesso” e S_2 o “segundo sucesso”. Considere, por exemplo, as linhas 5 e 8: ambas representam a mesma situação em que as tentativas 2 e 3 resultam em sucesso. Na realidade, elas não são sequências diferentes, mas simplesmente um resultado do método de contagem de sequências ordenadas no tempo.

Sequência	Número de tentativas				Sequência	Número de tentativas			
	1	2	3	4		1	2	3	4
1	S_1	S_2	-	-	7	S_2	-	S_1	-
2	S_1	-	S_2	-	8	-	S_2	S_1	-
3	S_1	-	-	S_2	9	-	-	S_1	S_2
4	S_2	S_1	-	-	10	S_2	-	-	S_1
5	-	S_1	S_2	-	11	-	S_2	-	S_1
6	-	S_1	-	S_2	12	-	-	S_2	S_1

Tabela 3.1: Ilustração das permutações (sequências ordenadas) de 2 sucessos em 4 tentativas

3. **Número de sequências não ordenadas ou combinações.** Como fica claro a partir do exemplo anterior, o número de permutações não é exatamente o número procurado, uma vez que não importa qual sucesso é rotulado como primeiro, segundo, etc. De acordo com a [Equação 3.2](#), no caso de $n = N$, há $n!$ maneiras de ordenar n sucessos entre si. Portanto, o número de permutações (ordenadas no tempo) precisa ser dividido por $n!$ para evitar a contagem dupla de sequências equivalentes. Fica claro, portanto, que o número de **combinações** de n sucessos em N tentativas é:

$$C(n, N) = \frac{\text{Perm}(n, N)}{n!} = \frac{N!}{n!(N - n)!} = \binom{N}{n}. \quad (3.3)$$

O número de combinações é o número de sequências possíveis de n sucessos em N tentativas. Esse número, é chamado de **coeficiente binomial** e é indicado pelo símbolo entre parênteses. O coeficiente binomial é usado na expansão binomial:

$$(p + q)^N = \sum_{n=0}^N \binom{N}{n} p^n q^{N-n}. \quad (3.4)$$

Exemplo 3.2

Continue a considerar o caso de 2 sucessos em 4 tentativas. Existem $2! = 2$ maneiras de ordenar os 2 sucessos entre si (ou um ou o outro é o primeiro sucesso). Portanto, o número de combinações de 2 sucessos em 4 tentativas é $4!/[2!(4 - 2)!] = 6$, e não 12. Conforme indicado acima, de fato, cada sequência tinha sua sequência “gêmea” listada separadamente, e a [Equação 3.3](#) conta corretamente apenas sequências diferentes.

De acordo com os resultados obtidos acima, o que resta a ser feito é usar a probabilidade de cada sequência ([3.1](#)) e multiplicá-la pelo número de combinações na [Equação 3.3](#) para obter a probabilidade geral de ter n sucessos em N tentativas. Isso leva à função massa de probabilidade (FMP):

$$P_N(n) = \binom{N}{n} p^n q^{N-n}, \quad n = 0, \dots, N, \quad (3.5)$$

conhecida como a **distribuição binomial**. Esta distribuição descreve a probabilidade de n sucessos em N tentativas de um experimento binário com probabilidade de sucesso p . A Equação 3.4 mostra que a distribuição binomial é devidamente normalizada. Exemplos da distribuição binomial são mostrados na Figura 3.1.

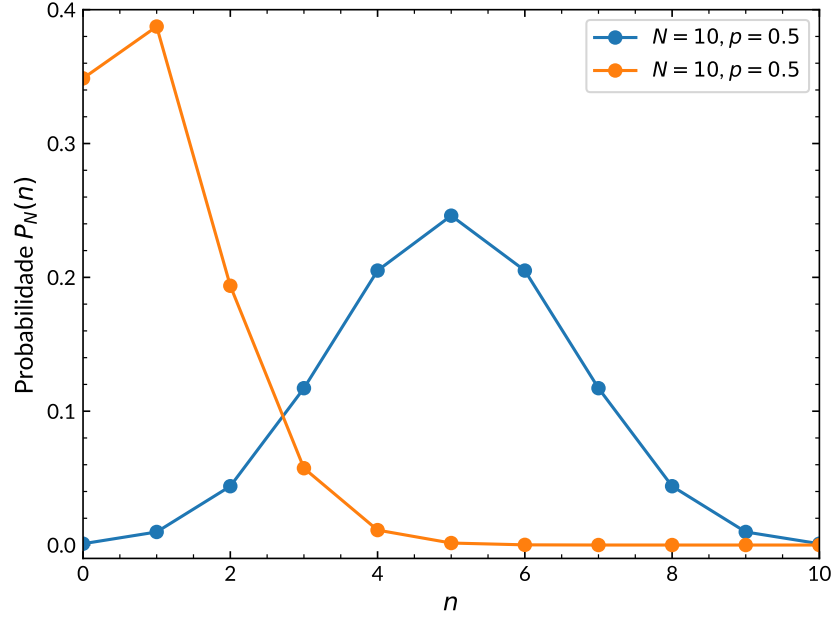


Figura 3.1: Distribuições binomiais de exemplo para $N = 10$ e $p = 0.5$, $p = 0.1$. A função é definida apenas para inteiros não negativos $0 \leq n \leq N$.

3.1.2 Momentos da Distribuição Binomial

Os momentos de primeira e segunda ordem de uma variável X distribuída binomialmente são dados por:

$$\mathbb{E}[X] = \mu = pN \quad (3.6a)$$

$$\mathbb{E}[X^2] = \mu^2 + pqN \quad (3.6b)$$

Vamos provar essas equações começando pela média,

$$\mathbb{E}[X] = \sum_{n=0}^N n P_N(n) = \sum_{n=0}^N \binom{N}{n} n p^n q^{N-n} = \sum_{n=0}^N \binom{N}{n} \left[p \frac{\partial}{\partial p} \right] p^n q^{N-n};$$

O operador linear $p \frac{\partial}{\partial p}$ pode ser aplicado a toda a soma, levando a (considere a substituição da Equação 3.4 aqui):

$$\mathbb{E}[X] = p \frac{\partial}{\partial p} \sum_{n=0}^N \binom{N}{n} p^n q^{N-n} = p \frac{\partial}{\partial p} (p+q)^N = pN(p+q)^{N-1} = pN,$$

onde o último passo se deve ao fato de que $p+q=1$. A derivação para o momento $\mathbb{E}[X^2]$ é similar:

$$\mathbb{E}[X^2] = \sum_{n=0}^N n^2 P_N(n) = \sum_{n=0}^N \binom{N}{n} n^2 p^n q^{N-n}.$$

Observe que, se $np^n = p \frac{\partial}{\partial p} p^n$, a seguinte expressão é válida:

$$n^2 p^n = p \frac{\partial}{\partial p} (np^n) = \left(p \frac{\partial}{\partial p} \right) \left(p \frac{\partial}{\partial p} \right) p^n = \left(p \frac{\partial}{\partial p} \right)^2 p^n.$$

Assim,

$$\mathbb{E}[X^2] = \sum_{n=0}^N \binom{N}{n} \left(p \frac{\partial}{\partial p} \right)^2 p^n q^{N-n} = \left(p \frac{\partial}{\partial p} \right)^2 (p+q)^N = p \frac{\partial}{\partial p} [pN(p+q)^{N-1}].$$

Desenvolvendo,

$$\begin{aligned} \mathbb{E}[X^2] &= p [N(p+q)^{N-1} + pN(N-1)(p+q)^{N-2}] = pN + (pN)^2 - p^2N \\ &= (pN)^2 + pN(1-p) = (pN)^2 + pqN, \end{aligned}$$

onde novamente usamos o fato de que $p+q=1$.

Segue-se que a variância da distribuição binomial é dada por:

$$\sigma^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = (pN)^2 + pqN - (pN)^2 = pqN. \quad (3.7)$$

As equações (3.6) e (3.7) descrevem as características mais importantes da distribuição binomial, mostradas na Figura 3.1 para o caso de $N=10$. A média é naturalmente dada pelo produto do número de tentativas N e a probabilidade de sucesso p em cada uma das tentativas.

Exemplo 3.3

Uma companhia aérea sabe que 5% das pessoas que fazem reservas não aparecerão no portão de embarque. Para um voo com capacidade para 50 passageiros, a companhia vendeu 52 passagens e quer determinar a probabilidade de que haverá um assento disponível para cada passageiro que aparecer. Utilizando a distribuição binomial, onde cada passageiro tem uma probabilidade $p=0.95$ de aparecer, queremos calcular a probabilidade de que no máximo 50 dos 52 passageiros compareçam ($P(X \leq 50)$). Definindo X como a variável aleatória que representa o número de passageiros que aparecem, temos $X \sim \text{Binomial}(N=52, p=0.95)$. A probabilidade desejada é dada por:

$$P(X \leq 50) = 1 - P(X=52) - P(X=51),$$

onde, usando a fórmula da distribuição binomial (3.5),

$$\begin{aligned} P(X=52) &= \binom{52}{52} (0.95)^{52} (0.05)^0 = (0.95)^{52} \\ P(X=51) &= \binom{52}{51} (0.95)^{51} (0.05)^1 = 52 \cdot (0.95)^{51} \cdot 0.05. \end{aligned}$$

Calculando esses valores, obtemos:

$$P(X \leq 50) = 1 - (0.95)^{52} - 52 \cdot (0.95)^{51} \cdot 0.05 \approx 0.741.$$

Assim, a probabilidade de que haverá um assento disponível para cada passageiro que aparecer é aproximadamente 74.1%. Consequentemente, a companhia aérea está assumindo um risco de 25.9% de ter um voo com *overbooking*.

3.2 A Distribuição Gaussiana

A distribuição Gaussiana, frequentemente referida como **distribuição normal**, desempenha um papel especial na estatística. Uma variável aleatória contínua X é dita ter uma distribuição Gaussiana se sua função densidade de probabilidade (FDP) for

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{(x-\mu)^2}{2\sigma^2} \right] \quad (3.8)$$

A função de distribuição é definida para todos os valores reais, e sua forma é determinada pelos dois parâmetros μ e σ^2 , que também representam a média e a variância da variável aleatória (veja a Figura 3.2). Uma Gaussiana com parâmetros μ e σ^2 é frequentemente referida como $\mathcal{N}(\mu, \sigma^2)$. É útil mostrar que a distribuição Gaussiana pode ser considerada um caso especial da distribuição binomial, no caso de um grande número de experimentos realizados.

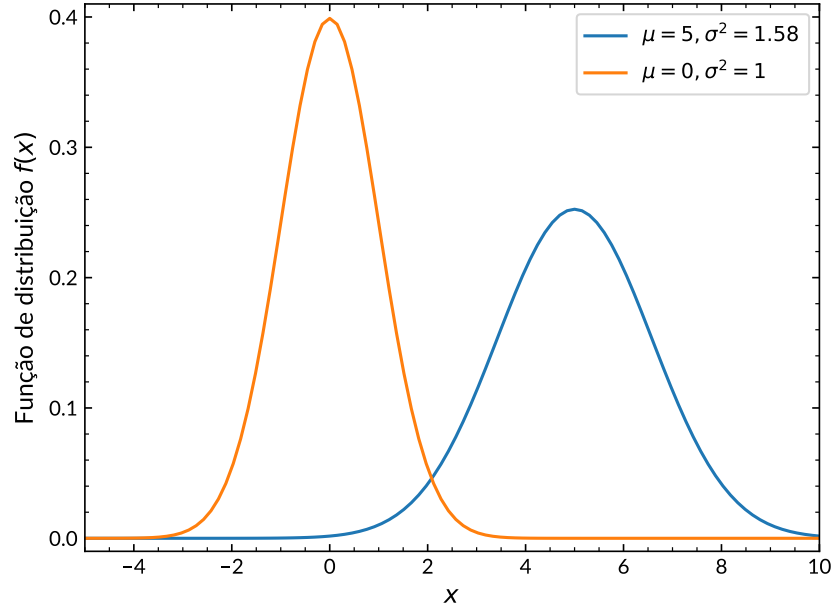


Figura 3.2: Distribuições Gaussianas para valores selecionados da média e variância. A curva Gaussiana em azul tem a mesma média e variância que uma distribuição binomial com $N = 10$ e $p = 0.5$ (mostrada na Figura 3.1).

3.2.1 Derivação da Distribuição Gaussiana a partir da Distribuição Binomial

Quando o tamanho da amostra N é grande, a distribuição binomial, representada na Equação 3.5, assume uma forma mais simplificada. Ter uma expressão analítica alternativa para a distribuição binomial é vantajoso, especialmente considerando a complexidade numérica associada ao cálculo de fatoriais para números grandes. Como destacado na Figura 3.1, a distribuição binomial atinge um máximo em torno de $n = Np$. Esta seção demonstra que a distribuição binomial pode ser aproximada por:

$$P_N(n) \approx \frac{1}{\sqrt{2\pi Npq}} \exp \left[-\frac{(n - Np)^2}{2Npq} \right], \quad (3.9)$$

quando $N \gg 1$, e para valores da variável que estão próximos ao pico da distribuição.

Vamos começar expandindo o logaritmo da probabilidade binomial como uma série de Taylor na vizinhança do valor de pico \tilde{n} ,

$$\ln P_N(n) = \ln P_N(\tilde{n}) + \sum_{k=1}^{\infty} \frac{B_k}{k!} \Delta n^k,$$

onde $\Delta n = n - \tilde{n}$ é a desvio em relação ao valor de pico e

$$B_k = \left. \frac{\partial^k \ln P_N(n)}{\partial n^k} \right|_{n=\tilde{n}}.$$

Como, por hipótese, \tilde{n} é um ponto de máximo, $B_1 = 0$. Negligenciando termos de ordem superior à segunda, obtém-se a aproximação

$$\ln P_N(n) \approx \ln P_N(\tilde{n}) + \frac{1}{2} B_2 \Delta n^2,$$

onde B_2 é negativo, já que $n = \tilde{n}$ é um ponto de máximo. Segue-se que,

$$P_N(n) \approx P_N(\tilde{n}) \exp \left(-\frac{|B_2| \Delta n^2}{2} \right).$$

Negligenciar termos de ordem superior em Δn significa que a aproximação será particularmente precisa em regiões onde Δn é pequeno, ou seja, próximo ao pico da distribuição. Longe do pico, a aproximação não será tão precisa. Para calcular $|B_2|$, façamos

$$\ln P_N(n) = \ln \left\{ \left[\frac{N!}{n!(N-n)!} \right] p^n q^{N-n} \right\} = \ln N! - \ln n! - \ln(N-n)! + n \ln p + (N-n) \ln q,$$

e trate n como uma variável contínua. Esta aproximação é razoável quando n são números grandes, em particular quando a média $Np \gg 1$ e para valores próximos da média. O logaritmo de $n!$ pode ser aproximado pela **fórmula de Stirling**,

$$\ln n! = n \ln n - n + O(\ln n),$$

onde O significa que, para todos os valores suficientemente grandes de n , a diferença entre $\ln n!$ e $n \ln n - n$ será no máximo proporcional ao logaritmo. Assim, a derivada dessa expressão pode ser aproximada com segue,

$$\frac{\partial \ln n!}{\partial n} \approx \ln n + n \left(\frac{1}{n} \right) - 1 = \ln n.$$

A partir disso, segue-se que a primeira derivada da função de probabilidade, como esperado, é zero no valor de pico,

$$\left. \frac{\partial \ln P_N(n)}{\partial n} \right|_{n=\tilde{n}} = -\ln n + \ln(N-n) + \ln p - \ln q \Big|_{n=\tilde{n}} = \ln \left[\frac{(N-n)p}{nq} \right] \Big|_{n=\tilde{n}} = 0$$

de modo que o resultado familiar de,

$$(N - \tilde{n})p = \tilde{n}q \Rightarrow \tilde{n}(p + q) = Np \Rightarrow \tilde{n} = Np,$$

é obtido. Isso leva ao cálculo da segunda derivada,

$$B_2 = \left. \frac{\partial^2 \ln P_N(n)}{\partial n^2} \right|_{n=\tilde{n}} = \frac{\partial}{\partial n} \ln \left[\frac{(N-n)p}{nq} \right] \Big|_{n=\tilde{n}} = \frac{\partial}{\partial n} [\ln(N-n) - \ln n] \Big|_{n=\tilde{n}} = -\frac{1}{N-n} - \frac{1}{n} \Big|_{n=\tilde{n}},$$

onde,

$$B_2 = -\frac{1}{N-\tilde{n}} - \frac{1}{\tilde{n}} = -\frac{1}{N(1-p)} - \frac{1}{Np} = -\frac{1}{Npq},$$

em que usamos o fato que $p + q = 1$. Finalmente, a constante de normalização $P(\tilde{n})$ pode ser calculada fazendo uso da integral

$$\int_{-\infty}^{\infty} e^{-ax^2} dx = \sqrt{\frac{\pi}{a}}.$$

impondo a condição de normalização da função de distribuição de probabilidade,

$$P_N(\tilde{n}) \int_{-\infty}^{\infty} \exp \left(-\frac{|B_2| \Delta n^2}{2} \right) d\Delta n = P_N(\tilde{n}) \sqrt{\frac{2\pi}{|B_2|}} = 1,$$

levando a (lembre-se que $|B_2| = 1/(Npq)$),

$$P_N(\tilde{n}) = \frac{1}{\sqrt{2\pi Npq}}.$$

Assim, a aproximação da distribuição binomial para grandes valores de n é, portanto:

$$P_N(n) \approx \frac{1}{\sqrt{2\pi Npq}} \exp \left[-\frac{(n - Np)^2}{2Npq} \right].$$

Temos, portanto, que a média de uma distribuição binomial é $\mu = Np$ e a variância é $\sigma^2 = Npq$. Assim, a aproximação da binomial para grandes valores de n pode ser reescrita como se segue,

$$P_N(n) \approx \frac{1}{\sqrt{2\pi Npq}} \exp \left[-\frac{(n - \mu)^2}{2\sigma^2} \right]. \quad (3.10)$$

que é a forma padrão da distribuição Gaussiana quando n é uma variável contínua.

3.2.2 Momentos e Propriedades da Distribuição Gaussiana

Os parâmetros μ e σ^2 são, respectivamente, a média e a variância da distribuição Gaussiana. Esses resultados seguem da derivação da distribuição Gaussiana a partir da binomial e podem ser confirmados por cálculo direto das expectativas a partir da [Equação 3.8](#). Também pode ser provado que momentos centrais de ordem ímpar são zero, uma vez que a Gaussiana é simétrica em relação à média. Dado seu amplo uso na estatística, é importante quantificar a “largura efetiva” da distribuição Gaussiana em torno de sua média. Para determinar a probabilidade de que uma variável Gaussiana tenha valores em um intervalo de $\pm z\sigma$,

$$A(z) = \int_{\mu-z\sigma}^{\mu+z\sigma} f(x) dx = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mu-z\sigma}^{\mu+z\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx,$$

podemos começar simplificando essa integral realizando uma mudança de variável. Definimos $t = (x-\mu)/\sigma$, onde $x = \mu + t\sigma$ e $dx = \sigma dt$. Os limites da integral também mudam: quando $x = \mu - z\sigma$, $t = -z$, e quando $x = \mu + z\sigma$, $t = z$. Substituindo na integral, obtemos,

$$A(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-z}^z \exp\left[-\frac{(\mu+t\sigma-\mu)^2}{2\sigma^2}\right] \sigma dt = \frac{1}{\sqrt{2\pi}} \int_{-z}^z \exp\left(-\frac{t^2}{2}\right) dt. \quad (3.11)$$

Veja que probabilidade da variável esteja dentro de $\pm 1\sigma$ da média é $A(1) = 0.683$, ou 68.3%. Este intervalo da variável é também referido como um **intervalo de confiança** com 68,3% de probabilidade (intervalos de confiança serão estudados em detalhe em capítulos posteriores). A correspondência entre o intervalo $\pm 1\sigma$ e a faixa que abrange 68.3% da probabilidade aplica-se estritamente apenas à distribuição Gaussiana, para a qual o valor de σ é definido pela função de distribuição. No entanto, é prática comum calcular o intervalo de 68.3% (às vezes arredondado para 68%) mesmo para aquelas variáveis aleatórias que não seguem estritamente uma distribuição Gaussiana, referindo-se a ele como o intervalo de 1σ . As probabilidades associadas aos intervalos característicos de uma variável Gaussiana são apresentadas na [Tabela 3.2](#).

Tabela 3.2: Probabilidade associada a intervalos característicos de uma distribuição gaussiana.

Intervalo em torno da média	Probabilidade Integrada (%)
$\pm 1\sigma$	68.27
$\pm 2\sigma$	95.45
$\pm 3\sigma$	99.73
$\pm 4\sigma$	99.99
$\pm 5\sigma$	≥ 99.9999
FWHM (ou $\pm 1.18\sigma$)	76.10

Todas as distribuições Gaussianas podem ser obtidas a partir da distribuição padrão $\mathcal{N}(0, 1)$ através de uma simples mudança de variável. Se X é uma variável aleatória distribuída como $\mathcal{N}(\mu, \sigma^2)$ e Z uma Gaussiana padrão $\mathcal{N}(0, 1)$, então a relação entre Z e X é dada por

$$Z = \frac{X - \mu}{\sigma} \quad (\text{ou } X = \mu + \sigma Z). \quad (3.12)$$

A variável Z também é referida como o **z-score** associado à variável X .

A distribuição cumulativa de uma variável normal $\mathcal{N}(0, 1)$ é definida pela seguinte integral:

$$B(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z \exp\left(-\frac{t^2}{2}\right) dt = \frac{1}{2} + \frac{A(z)}{2}, \quad (3.13)$$

e descreve a probabilidade integrada até o valor z , onde $A(z)$, conforme a [Equação 3.11](#), representa a probabilidade integrada entre $\pm z$.

A **meia largura à meia altura**, algumas vezes referida como **HWHM** (do inglês *half width at half maximum*), é definida como a distância entre o ponto de máximo da função de distribuição de probabilidade (em $x = \mu$) e o ponto onde a distribuição alcança metade do pico. Dessa forma, ao considerar $x = \mu$ na [Equação 3.8](#), obtemos

$$f(\mu) = \frac{1}{\sigma\sqrt{2\pi}},$$

Como queremos encontrar o ponto onde a distribuição alcança metade deste valor máximo, multiplicamos essa expressão por $1/2$ e igualamos esse resultado a [Equação 3.8](#),

$$\frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] = \frac{1}{2\sigma\sqrt{2\pi}}.$$

Após algum algebrismo na expressão acima, encontramos

$$\text{HWHM} = |x - \mu| = \sigma\sqrt{2\ln 2} \approx 1.18\sigma,$$

o que significa que o ponto de meia altura está ligeiramente além de um desvio padrão da média, de ambos os lados da média. Da mesma forma, a **largura total à meia altura**, ou **FWHM** (do inglês *full width at half maximum*), é definida como a faixa completa entre os dois pontos de meia altura, e é aproximadamente 2.36σ .

3.3 A Distribuição de Poisson

A distribuição de Poisson descreve a probabilidade de ocorrência de eventos em experimentos de contagem quando o resultado possível é um número inteiro. A distribuição é, portanto, discreta e pode ser derivada como um caso limite da distribuição binomial.

3.3.1 Derivação da Distribuição de Poisson

A distribuição binomial possui outra aproximação útil quando a probabilidade de sucesso é pequena, $p \ll 1$. Nesse caso, o número de resultados positivos é muito menor do que o número de tentativas, $n \ll N$, e a função fatorial pode ser aproximada como:

$$N! = N(N-1) \cdots (N-n+1) \cdot (N-n)! \approx N^n (N-n)!.$$

O termo q^{N-n} pode ser aproximado usando:

$$\ln q^{N-n} = \ln(1-p)^{N-n} = (N-n) \ln(1-p) \approx -p(N-n) \approx -pN,$$

levando a:

$$q^{N-n} \approx e^{-pN}$$

Essas duas aproximações podem ser usadas na [Equação 3.5](#) para obter:

$$P_N(n) = \frac{N!}{n!(N-n)!} p^n q^{N-n} \approx \frac{N^n (N-n)!}{n!(N-n)!} p^n e^{-pN} = \frac{(pN)^n}{n!} e^{-pN}. \quad (3.14)$$

Como pN é a média da distribuição, a aproximação torna-se:

$$P_N(n) = \frac{\mu^n}{n!} e^{-\mu}. \quad (3.15)$$

conhecida como **distribuição de Poisson**. Esta função descreve a probabilidade de obter n resultados positivos, ou contagens, quando o número esperado de resultados é μ . Pode-se ver imediatamente que a distribuição está devidamente normalizada, já que:

$$\sum_{n=0}^{\infty} \frac{\mu^n}{n!} = e^{\mu}.$$

Uma característica fundamental desta distribuição é que ela é descrita por apenas um parâmetro, a média μ , em contraste com a distribuição Gaussiana que tinha dois parâmetros. Isso não significa que a distribuição de Poisson não tenha variância – nesse caso, não seria uma variável aleatória – mas que a variância pode ser escrita como uma função da média, como será mostrado a seguir.

3.3.2 Momentos e Propriedades da Distribuição de Poisson

A Equação 3.15 perdeu sua referência ao experimento binomial, restando apenas a média $\mu = Np$ como parâmetro. Usando a definição de média e variância, é fácil provar que uma variável aleatória X que segue uma distribuição de Poisson tem um valor esperado $\mathbb{E}[X] = \mu$ e uma variância $\text{Var}(X) = \mu$. O fato de que a média é igual à variância pode ser visto usando os valores para a binomial, $\mu = Np$ e $\sigma^2 = Npq$; como $p \ll 1$, $q \approx 1$, e $\mu \approx \sigma^2$. Como resultado, a distribuição de Poisson tem apenas um parâmetro que é igual tanto à média quanto à variância.

A média e a variância de uma distribuição de Poisson também podem ser calculadas diretamente a partir da função de massa de probabilidade. A média pode ser calculada da seguinte forma:

$$\mathbb{E}[X] = \sum_{n=0}^N n P_N(n) = e^{-\mu} \sum_{n=0}^N n \frac{\mu^n}{n!} = e^{-\mu} \left(\mu \frac{d}{d\mu} \right) \left(\sum_{n=0}^N \frac{\mu^n}{n!} \right) = \mu.$$

O momento de segunda ordem é:

$$\mathbb{E}[X^2] = \sum_{n=0}^N n^2 P_N(n) = e^{-\mu} \sum_{n=0}^N n^2 \frac{\mu^n}{n!} = e^{-\mu} \left(\mu \frac{d}{d\mu} \right) \left(\sum_{n=0}^N n \frac{\mu^n}{n!} \right) = e^{-\mu} \mu \left[\frac{d}{d\mu} (\mu e^{\mu}) \right] = \mu + \mu^2.$$

onde a equação para a média foi usada na derivação. Portanto, a variância é:

$$\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mu + \mu^2 - \mu^2 = \mu.$$

A distribuição de Poisson é interpretada como a probabilidade de ocorrência de n contagens quando a média esperada das contagens é μ . Isso faz da distribuição de Poisson a principal ferramenta estatística para todos os experimentos de *contagem*. Ao contrário da distribuição binomial, que é limitada a valores inteiros $n \leq N$, a distribuição de Poisson é definida para qualquer número inteiro não negativo. A referência ao número total de eventos possíveis (N) e a probabilidade de ocorrência de cada evento (p) foi perdida, restando apenas a média μ para descrever a principal propriedade do experimento de contagem.

Como pode ser visto na Figura 3.3, a distribuição de Poisson não é simétrica em relação à média, e a distribuição torna-se mais simétrica para valores maiores da média. Como em todas as distribuições discretas, só é significativo calcular a probabilidade em um ponto específico ou para um conjunto de pontos, e não para um intervalo de pontos, como no caso das distribuições contínuas. Além disso, a média da própria distribuição pode ser um número não inteiro, e ainda assim o resultado do experimento descrito pela distribuição de Poisson só pode assumir valores inteiros.

Exemplo 3.4

Considere uma fonte astronômica conhecida por produzir fótons, que são geralmente detectados por um dado detector na quantidade de $\mu = 2.5$ em um dado intervalo de tempo. A probabilidade de detectar $n = 4$ fótons em um dado intervalo de tempo é, portanto:

$$P(n = 4; \mu = 2.5) = \frac{\mu^n}{n!} e^{-\mu} = \frac{2.5^4}{4!} e^{-2.5} = 0.134.$$

A razão para uma probabilidade aparentemente grande de obter uma medida que difere da média esperada é simplesmente devido à natureza estatística do processo de detecção.

3.3.3 A Distribuição de Poisson e o Processo de Poisson

Uma justificativa mais formal para a interpretação da distribuição de Poisson como a distribuição de experimentos de contagem vem do **processo de Poisson**. Embora um tratamento completo deste assunto esteja além do escopo deste livro, uma breve descrição de processos estocásticos servirá para fortalecer a

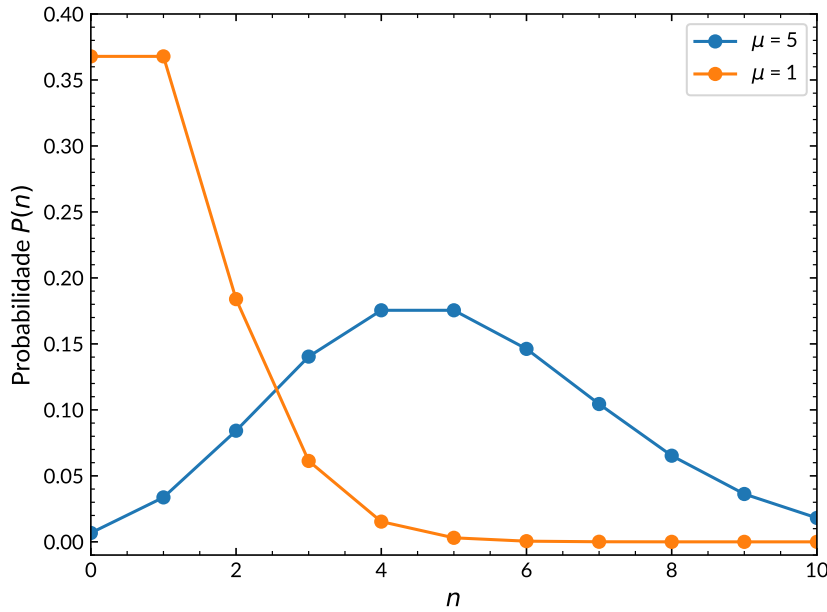


Figura 3.3: Distribuições de Poisson amostrais para valores selecionados da média. A função de massa de probabilidade de Poisson é definida para todos os valores $n \geq 0$.

interpretação de 3.15, que é um dos fundamentos da estatística. Mais detalhes sobre processos estocásticos podem ser encontrados, por exemplo, no livro-texto de Ross (2019).

Um **processo de contagem estocástico** $\{N(t), t > 0\}$ é uma sequência de variáveis aleatórias $N(t)$, em que t indica o tempo, e $N(t)$ é uma variável aleatória que indica o número de eventos que ocorreram até o tempo t . O processo estocástico pode ser pensado como a repetição do experimento de “contar a ocorrência de um determinado evento” em vários momentos t e $N(t)$ é o resultado do experimento. O **processo de Poisson com taxa λ** é um tipo particular de processo estocástico, com as seguintes propriedades:

1. $N(0) = 0$, o que significa que no tempo 0 não há contagens detectadas.
2. O processo tem **incrementos independentes**, o que significa que $N(s+t) - N(s)$ é independente de $N(s)$. Esta propriedade significa que eventos que ocorrem após o tempo s não são influenciados pelos que ocorreram antes.
3. O processo tem **incrementos estacionários**, ou seja, a distribuição do número de eventos em um intervalo de tempo s depende apenas da duração do próprio intervalo de tempo.
4. $P(N(h) = 1) = \lambda h + O(h)$, onde $O(h)$ é uma função com a propriedade que

$$\lim_{h \rightarrow 0} \frac{O(h)}{h} = 0.$$

5. $P(N(h) \geq 2) = O(h)$. As duas últimas propriedades significam que a probabilidade de obter uma contagem depende do valor finito λ , enquanto é improvável que dois ou mais eventos ocorram em um curto intervalo de tempo.

Pode-se mostrar que, sob essas hipóteses, o número de eventos $N(t)$ registrado em qualquer intervalo de comprimento t é distribuído segundo Poisson,

$$P\{N(s+t) - N(s) = n\} = \frac{(\lambda t)^n}{n!} e^{-\lambda t} \quad (3.16)$$

Isso mostra que a distribuição de Poisson deve ser interpretada como a distribuição da ocorrência de n eventos durante um intervalo de tempo t , sob a hipótese de que a taxa de ocorrência de eventos é λ . Esta interpretação é idêntica à fornecida acima, dado que $\mu = \lambda t$ é a média das contagens nesse intervalo de tempo.

3.4 Comparação das Distribuições Binomial, Gaussiana e de Poisson

Uma comparação entre as distribuições binomial, gaussiana e de Poisson com a mesma média é ilustrada na Figura 3.4.

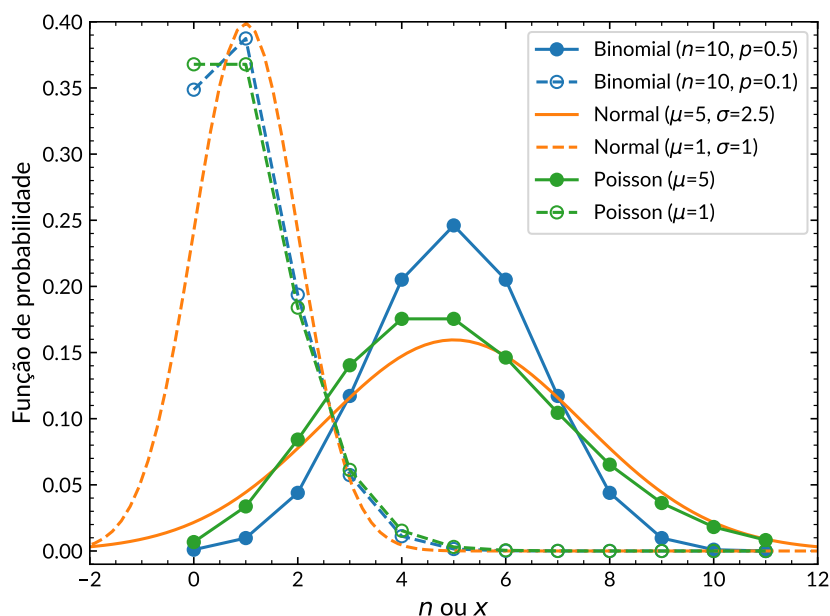


Figura 3.4: Comparação das distribuições de probabilidade binomial, gaussiana e de Poisson. As curvas à esquerda são para variáveis aleatórias com uma média esperada de $\mu = 1$ e à direita para $\mu = 5$.

A média e a variância de uma distribuição binomial são determinadas pela escolha do número de tentativas N e pela probabilidade de sucesso p do experimento binário, com $\mu = Np$ e $\sigma^2 = Npq$. No caso da distribuição gaussiana, a média e a variância podem ser escolhidas independentemente como os dois parâmetros da função de distribuição de probabilidade. A distribuição de Poisson, por outro lado, tem a característica especial de ter o mesmo valor para a média e a variância, na quantidade do único parâmetro μ na distribuição. Os dois conjuntos de curvas na Fig. 3.4 compartilham a mesma média, mas não é possível impor simultaneamente a mesma variância. Por exemplo, a distribuição de Poisson com $\mu = 5$ tem uma variância maior do que a binomial com $N = 10$ e $p = 0,5$, que tem uma variância de $\sigma^2 = 2,5$.

É de importância prática discutir a aproximação da distribuição de Poisson com uma distribuição normal de mesma média e variância. Para valores pequenos da média, a distribuição de Poisson é assimétrica em torno da média, como ilustrado na Fig. 3.4. À medida que a média aumenta, a distribuição torna-se progressivamente mais simétrica e os intervalos de confiança em torno da média se aproximam daqueles de uma distribuição gaussiana, como ilustrado na Fig. 3.5. Intervalos de confiança são calculados como intervalos em torno da média que englobam uma dada probabilidade. Para a distribuição gaussiana, os intervalos de confiança são sempre simétricos em torno da média. Para valores pequenos da média de Poisson, os intervalos de confiança são deslocados para valores maiores da variável aleatória, comparados à gaussiana, e tornam-se progressivamente mais próximos dos intervalos gaussianos à medida que a média aumenta. Dado que a distribuição de Poisson tem valores inteiros, a forma irregular das curvas de intervalo de confiança de Poisson (em vermelho na Fig. 3.5) é explicada pela exigência de que os intervalos também sejam limitados por inteiros. Por exemplo, o intervalo de uma variável de Poisson com $\mu = 1$ que engloba 90

A aproximação de uma distribuição normal também é mostrada pela curva sólida, indicando a diferença fracional das distribuições de Poisson e normal no valor de pico de $x = \mu$, como uma indicação da forma relativa das duas curvas. Como resultado, a aproximação de uma distribuição de Poisson com uma gaussiana de mesma média e variância torna-se cada vez mais precisa à medida que a média aumenta.

RESOLUÇÃO DOS PROBLEMAS

Capítulo 1

- 1.1. Considere o lançamento de quatro moedas iguais em sequência, onde Cara é representada por C e Coroa por K. O espaço amostral para esse experimento é:

$$\Omega = \{CCCC, CCKC, CCKC, CKCC, KCCC, CCKK, CKCK, KCKC, CKKC, KKCC, KCCK, KKKC, KKCK, KCKK, KKKK\}.$$

Assumindo que as moedas são imparciais, cada evento tem a mesma probabilidade de $1/16$. A probabilidade de obter duas caras e duas coroas, ou seja, os eventos $\{CCKK, CKCK, KCKC, CKKC, KKCC, KCCK\}$ é $6/16 = 3/8$.

- 1.2. Considere os eventos:

$$A = \{\text{primeiro lançamento mostra número ímpar}\}, \quad e \quad B = \{\text{soma dos dois dados é 9}\}.$$

O evento A ocorre nos resultados 1, 3, e 5 (como cada evento tem uma probabilidade $1/6$), assim $P(A) = 1/2$. Para calcular $P(B)$, os pares que somam 9 são (3, 6), (4, 5), (5, 4), e (6, 3), portanto $P(B) = 4/36 = 1/9$. A probabilidade da união de A e B deve excluir a interseção deles para evitar duplicações, que são os lançamentos (3, 6) e (5, 4), onde $P(A \cap B) = 1/8$. Portanto, considerando o terceiro axioma de Kolmogorov generalizado (Equação 1.5), temos que a probabilidade do primeiro lançamento mostra um número ímpar ou a soma dos dois dados é 9 é

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{1}{2} + \frac{1}{9} - \frac{1}{18} = \frac{5}{18}.$$

- 1.3. Considere os eventos:

$$A = \{\text{obter número par}\}, \quad e \quad B = \{\text{obter número maior que 4}\}.$$

O evento A ocorre nos resultados 2, 4, e 6, assim $P(A) = 3/6 = 1/2$. Para calcular $P(B)$, os resultados maiores que 4 são 5 e 6, portanto $P(B) = 2/6 = 1/3$. A interseção de A e B ocorre apenas em 6, assim $P(A \cap B) = 1/6$. Assim, pelo terceiro axioma de Kolmogorov generalizado (Equação 1.5), temos:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{1}{2} + \frac{1}{3} - \frac{1}{6} = \frac{2}{3}.$$

- 1.4. Considere os eventos:

$$A = \{\text{soma dos dois lançamentos é 8}\}, \quad e \quad B = \{\text{o primeiro lançamento mostra 5}\}.$$

O evento A ocorre para as seguintes combinações de resultados: $(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)$ e portanto $P(A) = 5/36$. O evento B tem uma probabilidade $P(B) = 1/6$. O evento $A \cap B$ tem probabilidade $P(A \cap B) = 1/36$ (situação em que ambos os eventos ocorrem, no caso, quando ocorrer o resultado $(5, 3)$). Como

$$P(A) \cdot P(B) = \frac{5}{36} \cdot \frac{1}{6} = \frac{5}{1296} \neq P(A \cap B) = \frac{1}{36},$$

os dois eventos não são estatisticamente independentes.

1.5. Considere os seguintes eventos:

$$A = \{\text{o primeiro lançamento é par}\}, \quad e \quad B = \{\text{o segundo lançamento é par}\}.$$

O evento A ocorre para 2, 4, 6 (bem como o evento B) e portanto $P(A) = 3/6 = 1/2$ (analogamente, $P(B) = 1/2$). O evento $A \cap B$, que ocorre para as seguintes combinações de resultados: $(2, 2), (2, 4), (2, 6), (4, 2), (4, 4), (4, 6), (6, 2), (6, 4)$, tem probabilidade $P(A \cap B) = 9/36 = 1/4$. Como

$$P(A) \cdot P(B) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} = P(A \cap B) = \frac{1}{4},$$

os dois eventos são estatisticamente independentes.

- 1.6.** A probabilidade de retirar uma bola vermelha é $p = 3/5$, já que 3 das 5 bolas são vermelhas. (a) Os sorteios consecutivos são independentes e, portanto, a probabilidade de duas bolas vermelhas consecutivas é $p = 3/5 \cdot 3/5 = 9/25$, já que os sorteios são feitos com reposição. (b) A condição de que o primeiro sorteio foi uma bola vermelha é irrelevante para o resultado do segundo sorteio, e portanto a probabilidade é apenas $p = 3/5$. A solução também pode ser obtida usando a definição de probabilidade condicional (Equação 1.6) e a definição de independência estatística (Equação 1.8),

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A) \cdot P(B)}{P(A)} = P(B) = \frac{3}{5}$$

onde A é o evento de uma bola vermelha no primeiro sorteio e B é o evento de uma bola vermelha no segundo sorteio.

- 1.7.** A probabilidade do dado der 1 (evento A) é $P(A) = 1/6$, e a probabilidade da soma ser 5 (evento B) é $P(B) = 4/36$, já que as sequências que resultam em uma soma de 5 são apenas quatro: $(2, 3), (3, 2), (4, 1), (1, 4)$. (a) A probabilidade de A dado B é,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/36}{4/36} = \frac{1}{4},$$

já que $P(A \cap B) = 1/36$ representa a probabilidade da única combinação $(1, 4)$ ocorrer. (b) A probabilidade de B dado A é,

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{1/36}{1/6} = \frac{1}{6},$$

(c) A probabilidade do primeiro lançamento ser 1 e a soma ser 5 é simplesmente,

$$P(A \cap B) = P(A|B) \cdot P(B) = \frac{1}{4} \times \frac{4}{36} = \frac{1}{36}.$$

(d) Usando a Equação 1.11, podemos mostrar que,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{1/6 \times 1/6}{4/36} = \frac{1}{4}.$$

o teorema de Bayes é verificado.

- 1.8.** (a) Uma sequência específica tem uma probabilidade de $P(A) = (1/2)^4 = 1/16$, já que os quatro lançamentos são independentes, e cada um tem uma probabilidade de $1/2$ de cara. (b) Precisamos contar o número de sequências que têm 2 moedas caindo cara (evento B). Usando o resultado do Problema 1.1., vemos que existem 6 dessas sequências, e portanto a probabilidade é $P(B) = 6/16 = 3/8$ por contagem direta. O evento B inclui A , portanto a probabilidade da interseção é $P(A \cap B) = 1/16$. A probabilidade de uma sequência cara-coroa-cara-coroa dado que duas moedas caíram com a face para cima é

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/16}{3/8} = \frac{1}{6}$$

(c) Para isso, podemos usar o teorema de Bayes (Equação 1.11),

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{1/6 \times 3/8}{1/16} = 1.$$

O conhecimento da sequência cara-coroa-cara-coroa de fato implica a certeza de que 2 moedas caíram cara.

Capítulo 2

- 2.1.** Para mostrar que a distribuição exponencial é devidamente normalizada, façamos:

$$\int_0^{\infty} f(x) dx = \lambda \int_0^{\infty} e^{-\lambda x} dx = \lambda \left(-\frac{1}{\lambda} e^{-\lambda x} \right)_0^{\infty} = 1.$$

A média μ pode ser calculada usando a Equação 2.7:

$$\mu = \int_0^{\infty} x f(x) dx = \lambda \int_0^{\infty} x e^{-\lambda x} dx = \lambda \left(-\frac{1}{\lambda} x e^{-\lambda x} \right)_0^{\infty} + \frac{1}{\lambda} \int_0^{\infty} e^{-\lambda x} dx = \frac{1}{\lambda}$$

A variância σ^2 pode ser calculada usando a Equação 2.12:

$$\sigma^2 = \mathbb{E}[x^2] - \mu^2 = \lambda \int_0^{\infty} x^2 e^{-\lambda x} dx - \left(\frac{1}{\lambda} \right)^2 = \lambda \left(-\frac{1}{\lambda} x^2 e^{-\lambda x} \right)_0^{\infty} + \frac{2}{\lambda} \int_0^{\infty} x e^{-\lambda x} dx - \left(\frac{1}{\lambda} \right)^2 = \frac{1}{\lambda^2}.$$

- 2.2.** A variância da média amostral \bar{x} pode ser calculada da seguinte maneira:

$$\text{Var}(\bar{x}) = \text{Var} \left(\frac{1}{N} \sum_{i=1}^N x_i \right).$$

Usando a propriedade fornecida na Equação 2.13, ou seja, $\text{Var}(aX) = a^2 \cdot \text{Var}(X)$, temos,

$$\text{Var}(\bar{x}) = \left(\frac{1}{N} \right)^2 \text{Var} \left(\sum_{i=1}^N x_i \right).$$

Considerando o fato de que $\text{Var}(x + y) = \text{Var}(x) + \text{Var}(y)$, podemos reescrever a expressão anterior como:

$$\text{Var}(\bar{x}) = \left(\frac{1}{N} \right)^2 \left(\sum_{i=1}^N \text{Var}(x_i) \right).$$

Dado que $\text{Var}(x_i) = \sigma^2$ para $i = 1, 2, \dots, N$, obtemos que,

$$\text{Var}(\bar{x}) = \left(\frac{1}{N} \right)^2 \left(\sum_{i=1}^N \sigma^2 \right) = \left(\frac{1}{N} \right)^2 (N\sigma^2) = \frac{\sigma^2}{N}.$$

2.3. Utilizando os conjuntos de dados para o Tubo 1 e o Tubo 2 (disponíveis em `../data/thomson_tubo1.dat` e `../data/thomson_tubo2.dat`, respectivamente), podemos calcular a média amostral (Equação 2.8), variância (Equação 2.14), covariância (Equação 2.19) e coeficiente de correlação (Equação 2.22). As fórmulas correspondentes foram implementadas em Python e podem ser encontradas em `../utils/estatisticas.py`. Os resultados obtidos após a aplicação dessas fórmulas são os seguintes:

- Tubo 1:
 - Média de W/Q : 13.3
 - Variância de W/Q : 71.5
 - Média de I : 312.9
 - Variância de I : 8715.7
 - Covariância entre W/Q e I : 759.1
 - Coeficiente de correlação entre W/Q e I : 0.96
- Tubo 2:
 - Média de W/Q : 2.92
 - Variância de W/Q : 0.67
 - Média de I : 174.3
 - Variância de I : 286.2
 - Covariância entre W/Q e I : 13.2
 - Coeficiente de correlação entre W/Q e I : 0.95

Cabe ressaltar que, por conveniência, cada medição de W/Q está faltando um fator de 10^{11} . A aplicação das fórmulas correspondentes pode ser vista no Jupyter Notebook `../docs/notebooks/cap02.ipynb`.

2.4. Ao combinar todas as medições feitas no ar (14 medições), no hidrogênio (5) e no ácido carbônico (4), os resultados são os seguintes:

- Ar: $m/e = 0.47 \pm 0.09$ (o desvio padrão é 0.09);
- Hidrogênio: $m/e = 0.47 \pm 0.08$;
- Ácido carbônico: $m/e = 0.43 \pm 0.07$.

Todas as medições estão dentro do desvio padrão umas das outras, e a afirmação é portanto verificada. As medições estão em unidades de 10^{-7} . A aplicação das fórmulas correspondentes pode ser vista no Jupyter Notebook `../docs/notebooks/cap02.ipynb`.

2.5. A covariância amostral é calculada como -0.33 e o coeficiente de correlação amostral como -0.14 . A aplicação das fórmulas correspondentes pode ser vista no Jupyter Notebook `../docs/notebooks/cap02.ipynb`.

REFERÊNCIAS BIBLIOGRÁFICAS

- Bayes, M. and Price, M. An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S. *Philosophical Transactions of the Royal Society of London Series I*, 53:370–418, January 1763.
- Bernshtein, S. N. An attempt to axiomatize the foundations of the theory of probability (in russian). *Soobshch. Khar'k. matem. ob-va*, 15:209–274, 1917.
- Heisenberg, W. Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik. *Zeitschrift für Physik*, 43(3–4):172–198, March 1927. doi: 10.1007/BF01397280.
- Jeffreys, H. *Theory of Probability*. International series of monographs on physics. Clarendon Press, 1998.
- Kolmogorov, A. *Foundations of the Theory of Probability*. Chelsea Publishing Company, 1950.
- Pearson, K. and Lee, A. On the laws of inheritance in man: I. inheritance of physical characters. *Biometrika*, 2(4):357–462, 1903.
- Ross, S. *Introduction to Probability Models*. Elsevier Science, 2019.
- Siegrist, K. Random: Probability, mathematical statistics, stochastic processes. <https://www.randomservices.org/random/>, 2019.
- Thomson, J. J. Xl. cathode rays. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 44(269):293–316, 1897. doi: 10.1080/14786449708621070.
- Wilks, S. *Mathematical Statistics*. Princeton University Press, 1947.