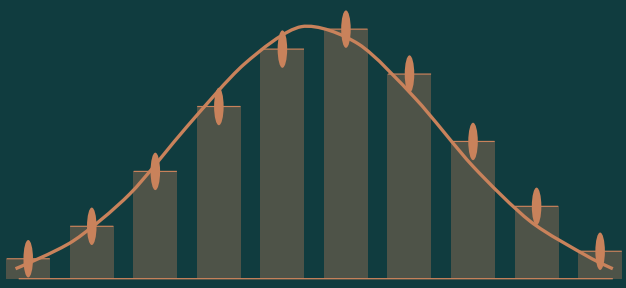
MODELOS, INCERTEZA E APLICAÇÕES

Cálculo de Probabilidade

Fundamentos, variáveis aleatórias, esperança e teoremas-limite

Daniel Miranda Machado



Texto: Daniel Miranda Machado.
Universidade Federal do ABC.
Composição em formato 17×24 cm.
Versão de trabalho – 13 de setembro de 2026.

Licença

Sumário

CAPÍTULO

1

Espaços de Probabilidade

1.1 Introdução

Um modelo determinístico associa condições iniciais e parâmetros a um único resultado. Uma equação diferencial com condição inicial, quando satisfaz as hipóteses de existência e unicidade, é um exemplo típico: fixados os dados, a evolução do sistema fica determinada.

Um modelo probabilístico descreve outra situação. O experimento pode produzir resultados diferentes mesmo quando as condições observáveis parecem idênticas, e o objetivo passa a ser quantificar a incerteza sobre esses resultados. O lançamento de uma moeda ilustra bem essa mudança de ponto de vista. Em princípio, as leis da mecânica determinam a face que ficará para cima; na prática, seria necessário conhecer posição, velocidades, atrito e impacto com precisão inatingível. Em vez de prever cada lançamento, descrevemos a regularidade do conjunto de lançamentos: para uma moeda equilibrada, a proporção de caras tende a se estabilizar perto de $1/2$.

A teoria da probabilidade fornece a linguagem e as regras de cálculo para esses modelos. Começaremos identificando os resultados possíveis e os eventos de interesse; em seguida, estabeleceremos os axiomas que governam suas probabilidades.

1.2 Experimentos probabilísticos

Mesmo sem conhecer antecipadamente o resultado de um experimento, em geral conseguimos descrever o conjunto de todas as saídas possíveis.

Definição 1.1

O **espaço amostral** Ω é o conjunto de todos os resultados possíveis em um determinado experimento.

O espaço amostral pode ser finito, infinito enumerável ou não enumerável. Um conjunto enumerável pode ser disposto em uma lista $x_1, x_2, \ldots$; os naturais, os inteiros e os racionais são exemplos. A reta real, por sua vez, não é enumerável.

Exemplo 1.1.

1. No lançamento de uma moeda, podemos usar $\Omega = \{H, T\}$, onde H representa cara e T, coroa.
2. No lançamento de um dado, $\Omega = \{1,2,3,4,5,6\}$.
3. Ao receber a primeira carta de um baralho comum, o espaço amostral é formado pelas 52 cartas.
4. No jogo de par ou ímpar, se cada pessoa mostra de zero a cinco dedos, podemos registrar apenas a soma e usar $\Omega = \{0,1,2,3,4,5,6,7,8,9,10\}$.
5. No sorteio uniforme de um número do intervalo $[0,1]$, temos $\Omega = [0,1]$.
6. Se retiramos uma lâmpada de uma caixa e observamos o tempo até ela se queimar, esse tempo pode assumir qualquer valor não negativo; podemos usar $\Omega = [0,\infty)$.

◁

Geralmente, porém, não queremos saber apenas qual ponto de Ω ocorreu. Queremos responder perguntas sobre o resultado: o dado foi maior que 4? A carta é de copas? O ponto sorteado caiu perto da origem? Cada pergunta dessas seleciona um subconjunto de Ω. É natural, portanto, chamar esses subconjuntos de *eventos*.

Em um espaço amostral finito podemos, sem dificuldade, admitir todos os subconjuntos. Se os resultados são equiprováveis, a probabilidade de um evento é simplesmente a fração de resultados favoráveis. Mas o que acontece quando Ω é contínuo?

Considere, por exemplo, a escolha uniforme de um ponto no disco unitário. Para uma região A com área bem definida, esperamos que

$$\mathbf{P}(A) = \frac{\text{área de } A}{\pi}.$$

Essa interpretação já sugere uma cautela: em espaços contínuos, não convém tratar arbitrariamente todo subconjunto como evento. Precisamos escolher uma família de conjuntos suficientemente ampla para conter as regiões usuais e, ao mesmo tempo, estável pelas operações que faremos com eventos.

Denotaremos essa família por $\mathcal{F}$. Em espaços finitos ou enumeráveis, podemos tomar $\mathcal{F} = \mathcal{P}(\Omega)$. Em $\mathbb{R}$ ou $\mathbb{R}^n$, trabalharemos com uma família que contém, em particular, os abertos e os fechados. Logo veremos quais propriedades essa família deve satisfazer; elas conduzem naturalmente à noção de σ-álgebra.

Definição 1.2

Dado um espaço amostral Ω e uma família $\mathcal{F}$ de subconjuntos de Ω, um **evento** é um conjunto $A \in \mathcal{F}$. Os eventos com exatamente um elemento são denominados **eventos elementares**.

Cada elemento de Ω será denominado **ponto amostral**, ou simplesmente ponto.

Exemplo 1.2.

1. Considere experimento de jogarmos um dado descrito no exemplo 2.. Nesse caso podemos considerar o evento $\{4,5,6\}$ que é o evento do dado ser maior que 4, ou o evento $\{1,3,5\}$ que é o evento de sair um número ímpar.

2. Se considerarmos o experimento de jogarmos par ou ímpar, descrito no exemplo 4., alguns eventos de importância são $P = \{0,2,4,6,8,10\}$, o evento de sair um número par, e $I = \{1,3,5,7,9\}$, o evento de sair um número ímpar.
   Esse experimento também pode ser representado através do seguinte espaço

amostral:

$$\Omega = \{(i,j) : 0 \leq i \leq 5,\ 0 \leq j \leq 5\},$$

ou seja, os pares ordenados cuja primeira entrada representa o número de dedos colocados pelo primeiro jogador, enquanto a segunda entrada representa o número de dedos colocados pelo do segundo jogador. Nessa representação temos o seguinte evento elementar $\{(1,3)\}$ que representa o fato do primeiro jogador colocar um dedo e o segundo três.
Nessa representação o evento da soma dos dedos colocados ser um número par pode ser representado pelo conjunto:

$$P = \{(i,j) : i + j \text{ é par, com } 0 \leq i \leq 5, 0 \leq j \leq 5\}$$

3. No caso de jogarmos dois dados o espaço amostral pode ser considerado como sendo $\Omega = \{(i,j) : 1 \leq i \leq 6, 1 \leq j \leq 6\}$, ou seja, os pares ordenados cuja primeira entrada representa a saída do primeiro dado, enquanto que a segunda entrada representa a saída do segundo dado. Nesse caso o espaço amostral tem 36 elementos.

   Nesse caso podemos, por exemplo, considerar o evento F de que a soma das saídas dos dois dados seja maior ou igual a 10,o que é representado pelo conjunto:

$$F = \{(i,j) \in \Omega : i + j \geq 10\} = \{(4,6),(5,5),(5,6),(6,4),(6,5),(6,6)\}.$$

4. No caso de realizarmos a escolha, ao acaso, de um ponto do disco de raio 1 centrado na origem. Então

$$\Omega = \text{ círculo unitário } = \left\{(x,y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\right\}.$$

   Para esse experimento podemos considerar por exemplo os eventos:

   a) $A =$ distância entre o ponto escolhido e a origem é $\leq \frac{1}{2}$;

   b) $B =$ distância entre o ponto escolhido e a origem é ≥ 0.25 ;

◁

Um exemplo particularmente útil é o conjunto vazio, que é o evento sem nenhum elemento. Esse evento será denotado por $\emptyset$ e denominado evento **impossível**. Já o evento espaço amostral Ω será denominado também de evento certo.

Quando consideramos espaços amostrais finitos, o número de elementos de um evento E, será denominado **cardinalidade** ou **tamanho** desse evento e será denotado por $\operatorname{card} E$. No caso do conjunto vazio temos que $\operatorname{card} \emptyset = 0$.

É importante refletirmos sobre o que fizemos até esse ponto. Até esse instante demos significado matemático a dois conceitos fundamentais da teoria da probabilidade: o de espaço amostral e o de evento. Ambos os conceitos foram traduzidos para a linguagem matemática de conjuntos. Falta agora aprofundar um pouco mais essa transcrição, interpretando conceitos da probabilidade na linguagem de conjuntos e interpretando as operações fundamentais em conjuntos em termos probabilísticos.

1.3 Álgebra de eventos

Eventos são, portanto, subconjuntos do espaço amostral E assim podemos usar a linguagem de conjuntos:

Denotamos o fato de um elemento (ou resultado de um experimento) a pertencer a um evento E através da notação $a \in E$ e o fato do elemento a não pertencer a um evento E através da notação $a \notin E$.

Também nesse contexto, diremos que dois eventos E e F são iguais se tiverem os mesmos elementos. Denotaremos tal fato por $E = F$.

Finalmente dizemos que um evento E está contido dentro do evento F, denotado por $E \subset F$, se todos elementos de E forem elementos de F.

Antes de definirmos probabilidade. Vamos apresentar uma família de operações com eventos, que permitem criar novos eventos a partir de eventos dados.

Definição 1.3

Definimos a **união** de dois eventos A e B, que denotaremos por $A \cup B$, como o evento formado pelos elementos de A ou B, i.e., o evento $A \cup B$ é o evento formado pelos elementos que pertencem a pelo menos um dos conjuntos A ou B.

$$A \cup B = \{x : x \in A \text{ ou } x \in B\}.$$

É importante ressaltar que na definição anterior, bem como em geral nos

textos matemáticos, o ou tem conotação inclusiva.

Definição 1.4

Definimos a **intersecção** de dois eventos A e B, que denotaremos por $A \cap B$ como o evento formado pelos elementos de A e B, i.e., o evento $A \cap B$ é formado pelos elementos que pertencem a ambos os conjuntos A e B.

$$A \cap B = \{x : x \in A \text{ e } x \in B\}.$$

Podemos generalizar a união e a intersecção para uma família de eventos (inclusive infinita) de maneira análoga:

Dada uma família de eventos E_1,E_2,..., definimos a união desses eventos, denotada por $\bigcup_{i=1}^{\infty} E_i$, como o evento formado pelos elementos que pertençam a pelo menos um dos eventos E_i Enquanto que a intersecção desses eventos, denotada por $\bigcap_{i=1}^{\infty} E_i$, como o evento formado pelos elementos que pertençam a todos os eventos E_i.

Definição 1.5

Definimos o **complementar de um evento** ou ainda o evento complementar de A, denotado A^{c}, como o conjunto dos elementos de Ω que não estão em A.

$$A^{\mathsf{c}} = \{x : x \in \Omega, x \notin A\}$$

É importante observar que num experimento o evento E ocorre se e somente se o evento E^{c} não ocorrer.

Diremos que dois eventos A e B são **mutuamente excludentes** se é impossível que ocorram simultaneamente ambos os eventos, isto é, se $A \cap B = \emptyset$.

Notação	Probabilidade	Conjunto
Ω	espaço amostral	universo
ω	evento elementar	elemento
$\emptyset$	evento impossível	conjunto vazio
A	evento	subconjunto
A^{c}	evento A não ocorre	
$A \cap B$	ocorre A e B	intersecção
$A \cup B$	ocorre A ou B	união
$A \subset B$	Se ocorre A então ocorre B	inclusão

Tabela 1.1: Conversão entre a linguagem da Probabilidade e da teoria da conjuntos.

As operações de união, intersecção e complemento possuem as seguintes propriedades:

1. Comutatividade: $A \cup B = B \cup A$ e $A \cap B = B \cap A$.
2. Associatividade: $(A \cup B) \cup C = A \cup (B \cup C)$ e $(A \cap B) \cap C = A \cap (B \cap C)$.
3. Distributividade: $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$ e $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$.
4. Leis de De Morgan

$$\left(\bigcup_{i=1}^{n} E_i\right)^{\mathsf{c}} = \bigcap_{i=1}^{n} E_i^{\mathsf{c}}.$$
$$\left(\bigcap_{i=1}^{n} E_i\right)^{\mathsf{c}} = \bigcup_{i=1}^{n} E_i^{\mathsf{c}}.$$

Finalmente vamos estipular as seguintes propriedades para a família de eventos $\mathcal{F}$: essa família deve ser uma σ-álgebra:

Definição 1.6

Uma σ-álgebra $\mathcal{F}$ é uma família de conjuntos satisfazendo:

1. $\Omega \in \mathcal{F}$.
2. Se $A \in \mathcal{F}$, então $A^c \in \mathcal{F}$.
3. Se $A_1, A_2, A_3, \ldots$ é uma sequência em $\mathcal{F}$, então sua união (enumerável) também está em $\mathcal{F}$, i.e., $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$.

Proposição 1.1

Seja $\mathcal{F}$ uma σ-álgebra de subconjuntos de Ω. Então valem as seguintes propriedades:

1. $\varnothing \in \mathcal{F}$
2. $\forall n, \forall A_1, \ldots, A_n \in \mathcal{F}$, temos $\bigcup_{i=1}^{n} A_i \in \mathcal{F}$ e $\bigcap_{i=1}^{n} A_i \in \mathcal{F}$.

No caso em que estivermos trabalhando com um espaço E que é $\mathbb{R}$ ou $\mathbb{R}^n$ então a σ-álgebra que consideraremos é a gerada pela família de conjuntos abertos de E é chamada a **σ-álgebra de Borel** de E, denotada por $\mathcal{B}_E$; seus elementos são chamados de conjuntos de Borel. Portanto, $\mathcal{B}_X$ inclui conjuntos abertos, conjuntos fechados (os complementares dos conjuntos abertos), interseções enumeráveis de conjuntos abertos (lembrando que uniões enumeráveis de conjuntos abertos já são abertos), uniões enumeráveis de conjuntos fechados (lembrando que interseções enumeráveis de conjuntos fechados já são fechados) e assim por diante.

1.4 Espaços de probabilidade

Definição 1.7

Um **espaço de probabilidade** é uma tripla $(\Omega, \mathcal{F}, \mathbf{P})$ onde Ω é denominado espaço amostral, $\mathcal{F}$ é uma σ-álgebra de subconjuntos de Ω e $\mathbf{P}$ é uma função que atribui um valor entre 0 e 1 (que denominaremos probabilidade) aos eventos $E \in \mathcal{F}$ e satisfazendo os seguintes axiomas:

Axioma 1. $0 \leq \mathbf{P}(E) \leq 1$.

Axioma 2. $\mathbf{P}(\Omega) = 1$.

Axioma 3. Se $E_1, E_2, \ldots$ forem eventos dois a dois mutuamente excludentes, isto é,

$$E_i \cap E_j = \emptyset \qquad \text{sempre que } i \neq j,$$

então

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

Proposição 1.2

$\mathbf{P}(\emptyset) = 0$.

Demonstração. Para qualquer evento E, podemos escrever $E = E \cup \emptyset$. Como os eventos E e $\emptyset$ são mutuamente excludentes, decorre do axioma 3, que $\mathbf{P}(E) = \mathbf{P}(E \cup \emptyset) = \mathbf{P}(E) + \mathbf{P}(\emptyset)$. E assim, a conclusão do teorema é imediata ∎

Proposição 1.3

$\mathbf{P}(E^{\mathsf{c}}) = 1 - \mathbf{P}(E)$.

Demonstração. Como $\Omega = E \cup E^{\mathsf{c}}$, com E e E^{c} mutuamente excludentes então, usando os axiomas 2 e 3, temos que:

$$1 = \mathbf{P}(\Omega) = \mathbf{P}(E \cup E^{\mathsf{c}})$$

$$1 = \mathbf{P}(E) + \mathbf{P}(E^{\mathsf{c}})$$

∎

Proposição 1.4

Se E e F forem dois eventos quaisquer, então

$$\mathbf{P}(E \cup F) = \mathbf{P}(E) + \mathbf{P}(F) - \mathbf{P}(E \cap F)$$

Demonstração. Como $E \cup F = E \cup (F \cap E^{\mathsf{c}})$ temos que:

$$\mathbf{P}(E \cup F) = \mathbf{P}(E) + \mathbf{P}(F \cap E^{\mathsf{c}}) .$$

Como $F = (E \cap F) \cup (F \cap E^{\mathsf{c}})$ temos que

$$\mathbf{P}(F) = \mathbf{P}(E \cap F) + \mathbf{P}(F \cap E^{\mathsf{c}}) .$$

Subtraindo a segunda igualdade da primeira, temos:

$$\mathbf{P}(E \cup F) - \mathbf{P}(F) = \mathbf{P}(E) - \mathbf{P}(E \cap F) .$$

■

Generalizando temos:

Proposição 1.5 — Princípio da inclusão-exclusão

Sejam E_1,E_2,...,E_n eventos. Então

$$\begin{aligned}\mathbf{P}(E_1 \cup E_2 \cup \cdots \cup E_n) = &\sum_{i=1}^{n} \mathbf{P}(E_i) - \sum_{i_1<i_2} \mathbf{P}(E_{i_1} \cap E_{i_2}) + \cdots \\ &+ (-1)^{r+1} \sum_{i_1<i_2<\cdots<i_r} \mathbf{P}(E_{i_1} \cap E_{i_2} \cap \cdots \cap E_{i_r}) + \cdots \\ &+ (-1)^{n+1} \mathbf{P}(E_1 \cap E_2 \cap \cdots \cap E_n).\end{aligned}$$

A soma

$$\sum_{i_1<i_2<\cdots<i_r} \mathbf{P}(E_{i_1} \cap E_{i_2} \cap \cdots \cap E_{i_r})$$

é feita sobre todos os $\binom{n}{r}$ subconjuntos possíveis de tamanho r do conjunto $\{1,2,\ldots,n\}$.

Em palavras, a proposição afirma que a probabilidade da união de n eventos é igual à soma das probabilidades individuais desses eventos, menos a soma das probabilidades das interseções dois a dois, mais a

soma das probabilidades das interseções três a três, e assim por diante, alternando os sinais.

Proposição 1.6

Se $A \subset B \Rightarrow \mathbf{P}(B - A) = \mathbf{P}(B) - \mathbf{P}(A)$

Definição 1.8

Seja $A_n, n \geq 1$, uma sequência de eventos. Essa sequência é dita

1. **crescente** se $A_1 \subset A_2 \subset \cdots \subset A_n \subset A_{n+1} \subset \cdots$. Nesse caso, definimos
$$\lim_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} A_n.$$
2. **decrescente** se $A_1 \supset A_2 \supset \cdots \supset A_n \supset A_{n+1} \supset \cdots$. Nesse caso, definimos
$$\lim_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} A_n.$$

Proposição 1.7

Sejam A_1,A_2,$\cdots$ eventos tais que $A_n \downarrow \emptyset$, ou seja, $A_1 \supset A_2 \supset A_3 \supset \cdots$ e ainda o $\lim_{n\to\infty} A_n = \emptyset$, então $\mathbf{P}(A_n) \to 0$.

Demonstração. Sejam A_1,A_2,$\cdots$ eventos tais que $A_n \downarrow \emptyset$, ou seja, $A_1 \supset A_2 \supset A_3 \supset \cdots$

Como $A_1 \supset A_2 \supset A_3 \supset \cdots$ então

$$A_1 = (A_1 - A_2) \cup (A_2 - A_3) \cup \cdots = \bigcup_{i=1}^{\infty} (A_i - A_{i+1}).$$

Observe que os conjuntos $A_i - A_{i+1}$ são conjuntos disjuntos, pois a sequência é uma sequência decrescente. Pelo axioma 3 temos que

$$\mathbf{P}(A_1) = \mathbf{P}\left(\bigcup_{i=1}^{\infty} (A_i - A_{i+1})\right) = \sum_{i=1}^{\infty} \mathbf{P}(A_i - A_{i+1}).$$

Logo por P5 $\mathbf{P}(A_i - A_{i+1}) = \mathbf{P}(A_i) - \mathbf{P}(A_{i+1})$, e portanto

$$\mathbf{P}(A_1) = \lim_{n\to\infty} \sum_{i=1}^{n-1} \mathbf{P}(A_i - A_{i+1}).$$

Note que os termos da somatória vão se cancelando restando apenas o primeiro e o último, assim

$$\mathbf{P}(A_1) = \lim_{n\to\infty} \mathbf{P}(A_1) - \mathbf{P}(A_n) = \mathbf{P}(A_1) - \lim_{n\to\infty} \mathbf{P}(A_n) \Rightarrow \lim_{n\to\infty} \mathbf{P}(A_n) = 0.$$

Portanto $\mathbf{P}(A_n) \to 0$.

Proposição 1.8 — Continuidade da Probabilidade

Seja $A_n, n \geq 1$, uma sequência de eventos crescente ou decrescente então vale a seguinte propriedade, conhecida como continuidade da probabilidade:

$$\lim_{n\to\infty} \mathbf{P}(A_n) = \mathbf{P}(\lim_{n\to\infty} A_n).$$

Demonstração. Primeiramente vamos considerar o caso em que $A_n \downarrow A$, ou seja, $A_{n+1} \subset A_n$ para qualquer $n \in \mathbb{N}$ e $\bigcap_{n\geq 1} A_n = A$. Assim sendo, temos que $\mathbf{P}(A_{n+1}) \leq \mathbf{P}(A_n)$, pois $A_{n+1} \subset A_n$.

Além disso, por propriedades de conjunto temos que $A_n - A \downarrow \emptyset$, o que implica que

$$\mathbf{P}(A_n - A) \to 0.$$

Logo

$$\mathbf{P}(A_n - A) = \mathbf{P}(A_n) - \mathbf{P}(A) \Rightarrow \mathbf{P}(A_n) - \mathbf{P}(A) \to 0 \Rightarrow \mathbf{P}(A_n) \to \mathbf{P}(A)$$

mas a sequência $\{\mathbf{P}(A_n)\}_{n\in\mathbb{N}}$ é decrescente, logo $\mathbf{P}(A_n) \downarrow \mathbf{P}(A)$

1.5 Probabilidade em espaços equiprováveis finitos

Uma das família de espaços de probabilidade mais simples e úteis são os espaços de probabilidade no qual todo evento elementar tem a mesma proba-

bilidade, i.e., $\mathbf{P}(w_1) = \mathbf{P}(w_2), \forall w_1, w_2 \in \Omega$. Esses espaços são denominados espaço de **probabilidade uniforme**.

Quando $\Omega = \{w_1, \ldots, w_n\}$ é um espaço de probabilidade uniforme temos que

$$\mathbf{P}(w_i) = \frac{1}{n} = \frac{1}{\operatorname{card}\Omega}$$

e logo

$$\mathbf{P}(E) = \sum_{w_i \in E} \mathbf{P}(w_i) = \sum_{i \in E} \frac{1}{\operatorname{card}\Omega} = \frac{\operatorname{card} E}{\operatorname{card}\Omega}.$$

Exemplo 1.3.

Qual a probabilidade de tiramos duas caras jogando uma moeda três vezes?

◁

Se denotarmos cara por H e coroa por T, temos que o espaço amostral nesse caso pode ser representado por:

$$\{(H,H,H),(H,H,T),(H,T,H),(T,H,H),(H,T,T),(T,H,T),(T,T,H),$$

$(T,T,T)\}$ Este espaço possui 2^3 elementos igualmente prováveis.

O evento "tirar duas caras" tem 4 elementos:

$$\{(H,H,H),(H,H,T),(H,T,H),(T,H,H)\}$$

e logo temos que a probabilidade de tirarmos duas caras é $\frac{4}{8} = \frac{1}{2}$

Exemplo 1.4.

Qual a probabilidade de tirarmos 12 jogando dois dados?

◁

Poderíamos considerar nesse caso que o espaço amostral fosse constituído pela soma dos valores dos dados sendo assim $\{2,3,4,\ldots,11,12\}$. Mas, se considerássemos esse espaço amostral, os eventos elementares não teriam a mesma probabilidade pois para tiramos 12 temos que tirar 6 em ambos os dados enquanto que para tirarmos 10 temos três possibilidades (4 e 6), (5 e 5) ou (6 e 4) para o primeiro e segundo dado respectivamente.

Nesse caso é muito mais interessante considerar o espaço amostral como $\{(i,j) : 1 \leq i \leq 6, 1 \leq j \leq 6\}$, ou seja, os pares ordenados cuja primeira entrada representa a saída do primeiro dado, enquanto a segunda entrada representa a saída do segundo dado. Nesse caso o espaço amostral tem 36 elementos igualmente prováveis. E nesse caso a probabilidade de tirarmos 12 é $\frac{1}{36}$.

Exemplo 1.5.
Qual a probabilidade de tirarmos mais que 9 jogando dois dados?

◁

Nesse caso podemos, por exemplo, considerar o evento de que a soma dos dois dados seja maior que 10, que é representado pelo conjunto $\{(i,j) : i+j > 10\} = \{(4,6),(5,5),(5,6),(6,4),(6,5),(6,6)\}$. Esse conjunto tem 6 elementos e assim a probabilidade de tirarmos mais que 10 é $\frac{6}{36} = \frac{1}{6}$

Exemplo 1.6.
Dois dados são lançados. Qual é a probabilidade de que a soma das faces seja igual a 8?

◁

Ao lançarmos dois dados, é importante escolher um espaço amostral em que todos os resultados elementares tenham a mesma probabilidade. Por isso, não devemos considerar apenas os possíveis valores da soma. O mais adequado é considerar os pares ordenados

$$\Omega = \{(i,j) : 1 \leq i \leq 6,\ 1 \leq j \leq 6\},$$

em que i representa o resultado do primeiro dado e j representa o resultado do segundo dado.

Esse espaço amostral possui $6 \cdot 6 = 36$ elementos igualmente prováveis. O evento "a soma das faces é igual a 8" é formado pelos pares

$$\{(2,6),(3,5),(4,4),(5,3),(6,2)\}.$$

Logo, há 5 casos favoráveis entre 36 possíveis. Portanto,

$$\mathbf{P}(\text{soma igual a } 8) = \frac{5}{36}.$$

Exemplo 1.7.
Três bolas são retiradas aleatoriamente de uma urna contendo 7 bolas brancas e 4 bolas pretas. Qual é a probabilidade de que duas bolas sejam brancas e uma seja preta?

◁

Neste problema, a ordem em que as bolas são retiradas não importa. Assim, podemos trabalhar diretamente com subconjuntos de 3 bolas escolhidos entre as 11 bolas da urna. O número total de escolhas possíveis é

$$\binom{11}{3}.$$

Queremos que, entre as três bolas escolhidas, exatamente duas sejam brancas e uma seja preta. Para isso, devemos escolher 2 das 7 bolas brancas e 1 das 4 bolas pretas. O número de escolhas favoráveis é, portanto,

$$\binom{7}{2}\binom{4}{1}.$$

Como todos os subconjuntos de 3 bolas são igualmente prováveis, temos

$$\mathbf{P}(\text{duas brancas e uma preta}) = \frac{\binom{7}{2}\binom{4}{1}}{\binom{11}{3}} = \frac{21 \cdot 4}{165} = \frac{28}{55}.$$

Exemplo 1.8.

Um comitê de 5 pessoas deve ser selecionado de um grupo com 8 homens e 7 mulheres. Qual é a probabilidade de que o comitê seja formado por 2 homens e 3 mulheres?

◁

Como a ordem das pessoas no comitê não importa, o espaço amostral é formado por todos os subconjuntos de 5 pessoas escolhidos entre as 15 pessoas disponíveis. Portanto, o número total de comitês possíveis é

$$\binom{15}{5}.$$

Para formar um comitê com exatamente 2 homens e 3 mulheres, devemos escolher 2 dos 8 homens e 3 das 7 mulheres. Assim, o número de comitês favoráveis é

$$\binom{8}{2}\binom{7}{3}.$$

Logo,

$$\mathbf{P}(2 \text{ homens e } 3 \text{ mulheres}) = \frac{\binom{8}{2}\binom{7}{3}}{\binom{15}{5}} = \frac{28 \cdot 35}{3003} = \frac{140}{429}.$$

Exemplo 1.9.

Uma urna contém n bolas, das quais uma é especial. Se k bolas são retiradas aleatoriamente, qual é a probabilidade de que a bola especial seja escolhida?

◁

Como estamos retirando k bolas de uma urna com n bolas, e a ordem de retirada não importa, o número total de escolhas possíveis é

$$\binom{n}{k}.$$

Agora contamos as escolhas favoráveis. Para que a bola especial seja escolhida, ela deve estar entre as k bolas retiradas. Depois de incluir essa bola, faltam escolher as outras $k-1$ bolas entre as $n-1$ bolas restantes. Assim, o número de casos favoráveis é

$$\binom{n-1}{k-1}.$$

Portanto,

$$\mathbf{P}(\text{a bola especial é escolhida}) = \frac{\binom{n-1}{k-1}}{\binom{n}{k}}.$$

Usando a fórmula das combinações, obtemos

$$\frac{\binom{n-1}{k-1}}{\binom{n}{k}} = \frac{k}{n}.$$

Assim, a probabilidade de que a bola especial seja retirada é

$$\frac{k}{n}.$$

Exemplo 1.10.

Uma mão de pôquer consiste em 5 cartas. Se as cartas tiverem valores consecutivos e não forem todas do mesmo naipe, dizemos que a mão é um *straight*. Qual é a probabilidade de que uma mão de pôquer seja um *straight*?

◁

Uma mão de pôquer é um conjunto de 5 cartas escolhidas entre as 52 cartas do baralho. Portanto, o número total de mãos possíveis é

$$\binom{52}{5}.$$

Para formar um *straight*, primeiro escolhemos uma sequência de 5 valores consecutivos. No baralho usual, há 10 possibilidades para essa sequência.

Depois, para cada uma dessas sequências, cada carta pode aparecer em qualquer um dos 4 naipes. Isso daria inicialmente

$$10 \cdot 4^5$$

possibilidades.

No entanto, precisamos excluir os casos em que as 5 cartas são todas do mesmo naipe. Esses casos são *straight flushes*, e não apenas *straights*. Para cada sequência de valores, existem 4 mãos desse tipo, uma para cada naipe. Assim, o número de *straights* é

$$10(4^5 - 4).$$

Logo,

$$\mathbf{P}(\text{straight}) = \frac{10(4^5 - 4)}{\binom{52}{5}} \approx 0{,}0039.$$

Exemplo 1.11.

Uma mão de pôquer consiste em 5 cartas. Um *full house* ocorre quando a mão possui três cartas de um mesmo valor e duas cartas de outro valor. Qual é a probabilidade de obter um *full house*?

◁

O número total de mãos possíveis é

$$\binom{52}{5}.$$

Para formar um *full house*, escolhemos primeiro o valor da trinca. Há 13 escolhas possíveis. Depois escolhemos o valor do par, que deve ser diferente do valor da trinca; para isso, há 12 escolhas possíveis.

Escolhido o valor da trinca, devemos escolher 3 das 4 cartas daquele valor:

$$\binom{4}{3}.$$

Do mesmo modo, escolhido o valor do par, devemos escolher 2 das 4 cartas daquele valor:

$$\binom{4}{2}.$$

Portanto, o número de mãos favoráveis é

$$13 \cdot 12 \binom{4}{3} \binom{4}{2}.$$

Assim,

$$\mathbf{P}(\text{full house}) = \frac{13 \cdot 12 \binom{4}{3}\binom{4}{2}}{\binom{52}{5}} \approx 0{,}0014.$$

Exemplo 1.12.

Trinta pessoas estão em uma sala. Qual é a probabilidade de que ao menos duas delas façam aniversário no mesmo dia?

◁

Vamos ignorar anos bissextos e supor que todos os dias do ano sejam igualmente prováveis para o nascimento de uma pessoa.

É mais simples calcular primeiro a probabilidade do evento complementar, isto é, a probabilidade de que todas as 30 pessoas façam aniversário em dias diferentes.

A primeira pessoa pode fazer aniversário em qualquer um dos 365 dias. Para que a segunda faça aniversário em um dia diferente, ela deve escolher um dos 364 dias restantes. A terceira deve escolher um dos 363 dias restantes, e assim por diante.

Portanto,

$$\mathbf{P}(\text{todos fazem aniversário em dias diferentes}) = \frac{365 \cdot 364 \cdot 363 \cdots 336}{365^{30}}.$$

Logo,

$$\mathbf{P}(\text{ao menos duas pessoas fazem aniversário no mesmo dia}) = 1 - \frac{365 \cdot 364 \cdot 363 \cdots 336}{365^{30}}.$$

Aproximadamente,

$$\mathbf{P}(\text{ao menos duas pessoas fazem aniversário no mesmo dia}) \approx 0{,}706.$$

Assim, em uma sala com 30 pessoas, a chance de haver ao menos duas pessoas com aniversário no mesmo dia é de aproximadamente 70,6%.

Exemplo 1.13.

Um baralho de 52 cartas é embaralhado e as cartas são viradas uma de cada vez até que o primeiro ás apareça. É mais provável que a carta virada logo após o primeiro ás seja o rei de copas ou o dois de paus?

◁

As duas probabilidades são iguais. Embora isso possa parecer pouco intuitivo à primeira vista, o argumento de contagem mostra que nenhuma das duas cartas tem vantagem.

Vamos calcular, por exemplo, a probabilidade de que o rei de copas apareça logo após o primeiro ás. Retiramos inicialmente o rei de copas do baralho e ordenamos as outras 51 cartas. Em qualquer ordenação dessas 51 cartas, existe uma única posição em que podemos inserir o rei de copas para que ele apareça imediatamente depois do primeiro ás.

Como existem 51! ordenações das demais cartas e 52! ordenações possíveis do baralho completo, temos

$$\mathbf{P}(\text{rei de copas logo após o primeiro ás}) = \frac{51!}{52!} = \frac{1}{52}.$$

O mesmo raciocínio vale para o dois de paus. Retiramos o dois de paus, ordenamos as outras 51 cartas e o inserimos imediatamente após o primeiro ás. Assim,

$$\mathbf{P}(\text{dois de paus logo após o primeiro ás}) = \frac{51!}{52!} = \frac{1}{52}.$$

Portanto, nenhuma das duas cartas é mais provável. As duas têm a mesma probabilidade de aparecer logo após o primeiro ás.

Exemplo 1.14.

Suponha que n homens deixem seus chapéus no centro de uma sala. Os chapéus são misturados e, em seguida, cada homem escolhe aleatoriamente um chapéu. Qual é a probabilidade de que nenhum homem escolha seu próprio chapéu?

◁

Esse é o chamado problema dos chapéus. Vamos numerar os homens e os chapéus de 1 a n, de modo que o chapéu i pertença ao homem i.

Depois que os chapéus são misturados, cada distribuição dos chapéus entre os homens pode ser vista como uma permutação dos números $1,2,\ldots,n$. Como todas as permutações são igualmente prováveis, o espaço amostral tem $n!$ resultados.

Chamemos de E_i o evento em que o homem i escolhe o seu próprio chapéu. Queremos calcular a probabilidade de que nenhum homem escolha seu próprio chapéu, isto é,

$$\mathbf{P}(E_1^c \cap E_2^c \cap \cdots \cap E_n^c).$$

É mais simples calcular primeiro a probabilidade do evento complementar: pelo menos um homem escolhe seu próprio chapéu. Esse evento é

$$E_1 \cup E_2 \cup \cdots \cup E_n.$$

Para calcular essa probabilidade, usamos o princípio da inclusão-exclusão.

Se fixarmos um homem, por exemplo o homem i, e exigirmos que ele escolha seu próprio chapéu, então os outros $n-1$ chapéus podem ser distribuídos livremente entre os outros homens. Logo,

$$\mathbf{P}(E_i) = \frac{(n-1)!}{n!}.$$

Como há n escolhas possíveis para esse homem, a primeira contribuição na inclusão-exclusão é

$$n\frac{(n-1)!}{n!} = 1.$$

Agora, se fixarmos dois homens e exigirmos que ambos escolham seus próprios chapéus, então os outros $n-2$ chapéus podem ser distribuídos livremente. Assim,

$$\mathbf{P}(E_i \cap E_j) = \frac{(n-2)!}{n!}.$$

Como há $\binom{n}{2}$ maneiras de escolher esses dois homens, a contribuição dos pares é

$$\binom{n}{2}\frac{(n-2)!}{n!} = \frac{1}{2!}.$$

De modo geral, se fixarmos k homens e exigirmos que todos eles escolham seus próprios chapéus, os demais $n-k$ chapéus podem ser distribuídos de

$$(n-k)!$$

maneiras. Portanto,

$$\mathbf{P}(E_{i_1} \cap \cdots \cap E_{i_k}) = \frac{(n-k)!}{n!}.$$

Como há $\binom{n}{k}$ maneiras de escolher esses k homens, a contribuição correspondente é

$$\binom{n}{k}\frac{(n-k)!}{n!} = \frac{1}{k!}.$$

Pelo princípio da inclusão-exclusão, a probabilidade de que pelo menos um homem escolha seu próprio chapéu é

$$\mathbf{P}(E_1 \cup \cdots \cup E_n) = 1 - \frac{1}{2!} + \frac{1}{3!} - \cdots + (-1)^{n+1}\frac{1}{n!}.$$

Portanto, a probabilidade de que nenhum homem escolha seu próprio chapéu é

$$\begin{aligned}\mathbf{P}(E_1^c \cap \cdots \cap E_n^c) &= 1 - \mathbf{P}(E_1 \cup \cdots \cup E_n) \\ &= 1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \cdots + (-1)^n\frac{1}{n!}.\end{aligned}$$

Assim,

$$\mathbf{P}(\text{nenhum homem escolhe seu próprio chapéu}) = \sum_{k=0}^{n}\frac{(-1)^k}{k!}.$$

Quando n é grande, essa soma se aproxima de

$$e^{-1} \approx 0{,}3679.$$

Portanto, para um número grande de homens, a probabilidade de que ninguém recupere o próprio chapéu é aproximadamente 36,8%.

1.6 Probabilidade na reta real

Se Ω é um conjunto infinito, suas probabilidades podem não ser determinadas em termos das probabilidades dos eventos elementares. Isso é especialmente verdadeiro quando Ω representa o conjunto de pontos em uma linha ou em um espaço n-dimensional. Esses exemplos são de extrema relevância, já que muitas aplicações podem ser descritas em termos de eventos nesse espaço. Neste contexto, exploraremos a determinação de probabilidades utilizando a reta real como ilustração.

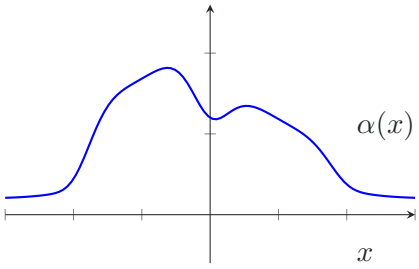
Consideremos Ω como o conjunto de todos os números reais. Aqui, os eventos são subconjuntos de pontos na reta real. Demonstra-se que é impossível atribuir probabilidades a todos os subconjuntos de Ω de modo a satisfazer os axiomas de probabilidade. Para construir um espaço de probabilidade na reta real, consideramos como eventos todos os intervalos $x_1 \leq x \leq x_2$ e suas uniões e interseções enumeráveis. Esses eventos formam a σ-álgebra de Borel.

Essa σ-álgebra inclui intervalos, pontos e os conjuntos usados nas aplicações usuais. Para especificar uma probabilidade na reta, basta conhecer $\mathbf{P}((-\infty,x])$

para todo x, desde que esses valores satisfaçam as propriedades de uma função de distribuição. Eles determinam as probabilidades de todos os eventos de Borel.

Suponha que $\alpha(x)$ seja uma função tal que

$$\int_{-\infty}^{\infty} \alpha(x)dx = 1 \quad \text{e } \alpha(x) \geq 0$$



Definimos a probabilidade do evento $\{x \leq x_i\}$ pela integral

$$\mathbf{P}\left(x \leq x_i\right) = \int_{-\infty}^{x_i} \alpha(x)dx \tag{1.1}$$

Isto especifica as probabilidades dos eventos de Borel. Por exemplo,

$$\mathbf{P}\left(x_1 < x \leq x_2\right) = \int_{x_1}^{x_2} \alpha(x)dx$$

Na verdade, os eventos $\{x \leq x_1\}$ e $\{x_1 < x \leq x_2\}$ são mutuamente excludentes e sua união é igual a $\{x \leq x_2\}$. Portanto

$$\mathbf{P}\left(x \leq x_1\right) + \mathbf{P}\left(x_1 < x \leq x_2\right) = \mathbf{P}\left(x \leq x_2\right)$$

Como a integral de uma densidade sobre um intervalo cujo comprimento tende a zero também tende a zero, cada ponto tem probabilidade nula: $\mathbf{P}(\{x_2\}) = 0$. Ainda assim, $\mathbf{P}(\mathbb{R}) = 1$. Não há contradição: o axioma da aditividade refere-se a uniões enumeráveis, enquanto a reta é uma união não enumerável de pontos.

Exemplo 1.15.

Uma substância radioativa é selecionada em $t = 0$ e o tempo t de emissão de uma partícula é observado. Este processo define um experimento cujos resultados são todos pontos no eixo t positivo. Este experimento pode ser considerado um caso especial do experimento da linha real se assumirmos que Ω é todo o eixo t e todos os eventos no eixo negativo têm probabilidade zero.

Suponha então que a função $\alpha(t)$ seja dada por

$$\alpha(t) = ce^{-ct}U(t), \qquad c > 0, \qquad U(t) = \begin{cases} 1, & t \geq 0, \\ 0, & t < 0. \end{cases}$$

Inserindo em 1.1, concluímos que a probabilidade de uma partícula ser emitida no intervalo de tempo $(0, t_0)$ é igual

$$c \int_0^{t_0} e^{-ct} dt = 1 - e^{-ct_0}$$

◁

Exemplo 1.16.

Uma chamada telefônica ocorre uniformemente no intervalo $(0,T)$. Isso significa que a probabilidade de ocorrer entre 0 e t_0 é t_0/T. Para $0 \leq t_1 \leq t_2 \leq T$,

$$\mathbf{P}\left(t_1 \leq t \leq t_2\right) = \frac{t_2 - t_1}{T}$$

Este é novamente um caso especial da Equação 1.1 com $\alpha(t) = 1/T$ para $0 \leq t \leq T$ e 0 caso contrário.

◁

Ao atravessar este capítulo

Os axiomas transformam descrições de experimentos em cálculos: primeiro se define o espaço amostral, depois os eventos e, por fim, uma probabilidade compatível com complementaridade, uniões e limites.

1.7 Exercícios

Exercício 1.1 — Operações com eventos

Em uma turma, 42 estudantes cursam Cálculo, 31 cursam Álgebra e 18 cursam ambas. Entre 60 estudantes, determine as probabilidades de escolher alguém que: (a) curse ao menos uma disciplina; (b) curse somente Cálculo; (c) não curse nenhuma das duas.

Exercício 1.2 — Contagem

Uma senha contém quatro algarismos distintos seguidos de duas letras distintas. Os algarismos podem incluir zero, mas a primeira posição não pode ser zero. Escolhida uma senha uniformemente entre as admissíveis, calcule a probabilidade de ela começar por um algarismo par e terminar por uma vogal.

Exercício 1.3 — Inclusão–exclusão

Três sensores detectam um sinal com probabilidades 0,72, 0,68 e 0,61. As probabilidades das detecções simultâneas por pares são 0,50, 0,44 e 0,41, e a detecção pelos três tem probabilidade 0,35. Calcule a probabilidade de ao menos um sensor detectar o sinal.

Exercício 1.4 — Continuidade da probabilidade

Se $A_1 \subset A_2 \subset \cdots$ e $\mathbf{P}(A_n) = 1 - 3^{-n}$, determine $\mathbf{P}(\bigcup_{n\geq 1} A_n)$. Formule e resolva a versão correspondente para uma sequência decrescente B_n com $\mathbf{P}(B_n) = \frac{1}{2} + 4^{-n}$.

Exercício 1.5 — Modelo geométrico

Um ponto é escolhido uniformemente no retângulo $[0{,}4] \times [0{,}3]$. Calcule a probabilidade de sua distância à origem ser menor que 2 e a probabilidade de $x + y \leq 3$. Represente as duas regiões.

CAPÍTULO

2

Probabilidade condicional e independência

2.1 Probabilidade condicional

Probabilidades mudam quando recebemos informação. Se sabemos que um evento B ocorreu, resultados fora de B deixam de ser possíveis. Mas como devemos redistribuir as probabilidades dentro de B?

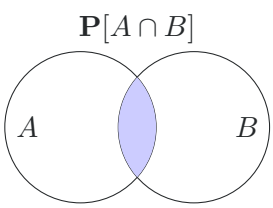
Há duas exigências naturais. Primeiro, B deve passar a ter probabilidade 1. Segundo, as proporções internas devem ser preservadas: se, antes de condicionar, $A \cap B$ era três vezes mais provável que $C \cap B$, essa razão não deve mudar apenas porque descobrimos que estamos em B.

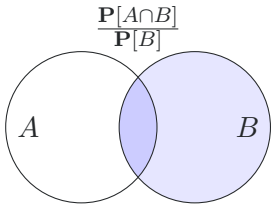
Isso praticamente determina a operação que precisamos. Para medir A depois de observar B, conservamos apenas a parte compatível, $A \cap B$, e renormalizamos pela massa total de B.

Definição 2.1 — Probabilidade Condicional

Dado um espaço de probabilidade $(\Omega,\mathcal{F},\mathbf{P})$ e eventos $A,B \in \mathcal{F}$ com $\mathbf{P}(B) > 0$, a **probabilidade condicional** de A dado B é

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}$$





Observe que as propriedades que comentamos acima são respeitadas por essa definição.

- $\mathbf{P}(A|B)$ é proporcional a $\mathbf{P}(A \cap B)$;
- $\mathbf{P}(B|B) = 1$;
- Para $A, C \in \mathcal{F}$ com $\mathbf{P}(C \cap B) > 0$, vale

$$\frac{\mathbf{P}(A \mid B)}{\mathbf{P}(C \mid B)} = \frac{\dfrac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}}{\dfrac{\mathbf{P}(C \cap B)}{\mathbf{P}(B)}} = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(C \cap B)}.$$

Frequentemente pensamos na probabilidade condicional como a probabilidade quando restringimos ao evento B, tomando este como o novo espaço amostral.

Teorema 2.1

Dado um espaço de probabilidade $(\Omega,\mathcal{F},\mathbf{P})$ e $B \in \mathcal{F}$ com $\mathbf{P}(B) > 0$, então $(B,\mathcal{F}_B,\mathbf{P}(\,\cdot \mid B))$ é um espaço de probabilidade, onde $\mathcal{F}_B = \{A \cap B : A \in \mathcal{F}\}$.

Exemplo 2.1 — Inspeção automática com tempo máximo.

Uma inspeção automática pode terminar antes do prazo ou, se a análise não for concluída antecipadamente, permanecer ativa até o tempo máximo de 40 minutos. Seja T sua duração. Suponha que

$$\mathbf{P}(T < t) = \frac{t}{60}, \qquad 0 \le t \le 40.$$

Sabendo que a inspeção ainda está em andamento após 30 minutos, qual é a probabilidade de ela permanecer ativa até os 40 minutos?

Solução. Seja F o evento de a inspeção atingir o tempo máximo. Como

$$\mathbf{P}(T < 40) = \frac{40}{60} = \frac{2}{3},$$

temos

$$\mathbf{P}(F) = \mathbf{P}(T = 40) = 1 - \frac{2}{3} = \frac{1}{3}.$$

Por outro lado, o evento de a inspeção ainda estar em andamento após 30 minutos é $\{T \geq 30\}$, cuja probabilidade é

$$\mathbf{P}(T \geq 30) = 1 - \mathbf{P}(T < 30) = 1 - \frac{30}{60} = \frac{1}{2}.$$

Como $F \subseteq \{T \geq 30\}$, segue que

$$\mathbf{P}(F \mid T \geq 30) = \frac{\mathbf{P}(F)}{\mathbf{P}(T \geq 30)} = \frac{1/3}{1/2} = \frac{2}{3}.$$

O ponto importante é que a distribuição de T possui uma parte contínua antes de 40 minutos e uma massa de probabilidade no próprio tempo máximo: toda análise que não termina antes fica acumulada em $T = 40$.

◁

Exemplo 2.2.

Suponha que, em uma determinada cidade, 40% da população seja composta por homens e 60% por mulheres. Suponha também que 50% dos homens sejam fumantes e que 30% das mulheres sejam fumantes.

Se uma pessoa é escolhida aleatoriamente e sabemos que ela é fumante, qual é a probabilidade de que essa pessoa seja homem?

Solução.

Vamos denotar por H o evento "a pessoa escolhida é homem" e por F o evento "a pessoa escolhida é mulher". Além disso, seja S o evento "a pessoa escolhida é fumante".

As informações do enunciado podem ser escritas como

$$\mathbf{P}(H) = 0{,}40, \qquad \mathbf{P}(F) = 0{,}60,$$

e

$$\mathbf{P}(S \mid H) = 0{,}50, \qquad \mathbf{P}(S \mid F) = 0{,}30.$$

Queremos calcular

$$\mathbf{P}(H \mid S),$$

isto é, a probabilidade de que a pessoa seja homem, sabendo que ela é fumante.

Pela definição de probabilidade condicional,

$$\mathbf{P}(H \mid S) = \frac{\mathbf{P}(H \cap S)}{\mathbf{P}(S)}.$$

Primeiro, calculamos a probabilidade de a pessoa escolhida ser homem e fumante:

$$\mathbf{P}(H \cap S) = \mathbf{P}(H)\mathbf{P}(S \mid H) = 0{,}40 \cdot 0{,}50 = 0{,}20.$$

Agora precisamos calcular $\mathbf{P}(S)$, isto é, a probabilidade de que uma pessoa escolhida ao acaso seja fumante. Uma pessoa fumante pode ser homem ou mulher. Assim,

$$S = (S \cap H) \cup (S \cap F),$$

e essa união é disjunta. Portanto,

$$\mathbf{P}(S) = \mathbf{P}(S \cap H) + \mathbf{P}(S \cap F).$$

Já sabemos que

$$\mathbf{P}(S \cap H) = 0{,}20.$$

Além disso,

$$\mathbf{P}(S \cap F) = \mathbf{P}(F)\mathbf{P}(S \mid F) = 0{,}60 \cdot 0{,}30 = 0{,}18.$$

Logo,

$$\mathbf{P}(S) = 0{,}20 + 0{,}18 = 0{,}38.$$

Portanto,

$$\mathbf{P}(H \mid S) = \frac{0{,}20}{0{,}38} = \frac{20}{38} = \frac{10}{19} \approx 0{,}526.$$

Assim, sabendo que a pessoa escolhida é fumante, a probabilidade de que ela seja homem é aproximadamente

$$52{,}6\%.$$

◁

2.2 Probabilidade total e fórmula de Bayes

A probabilidade condicional também pode ser lida ao contrário. Muitas vezes conhecemos $\mathbf{P}(B \mid A)$, mas a pergunta interessante é $\mathbf{P}(A \mid B)$. Há alguma maneira de inverter o condicionamento?

A resposta vem da mesma intersecção escrita de dois modos:

$$\mathbf{P}(A \cap B) = \mathbf{P}(A \mid B)\mathbf{P}(B) = \mathbf{P}(B \mid A)\mathbf{P}(A).$$

Quando $\mathbf{P}(B) > 0$, basta isolar $\mathbf{P}(A \mid B)$.

Teorema 2.2 — Fórmula de Bayes

Dados eventos A e B, tais que $\mathbf{P}(B) \neq 0$ então:

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(B|A)\,\mathbf{P}(A)}{\mathbf{P}(B)}$$

Exemplo 2.3.

Considere um teste para uma doença rara. Suponha que o teste tenha sensibilidade de 98%, isto é, ele acusa resultado positivo em 98% das pessoas que realmente possuem a doença. Suponha também que sua especificidade seja de 98%, isto é, ele acusa resultado negativo em 98% das pessoas saudáveis.

Além disso, suponha que apenas 0,5% da população possua a doença. Se uma pessoa escolhida aleatoriamente testa positivo, qual é a probabilidade de que ela realmente esteja doente?

Solução.

Vamos denotar por D o evento "a pessoa está doente" e por $+$ o evento "o teste deu positivo". Queremos calcular

$$\mathbf{P}(D \mid +),$$

isto é, a probabilidade de a pessoa estar doente sabendo que o teste deu positivo.

Pelas informações do enunciado, temos

$$\mathbf{P}(D) = 0{,}005, \qquad \mathbf{P}(D^c) = 0{,}995,$$

e também

$$\mathbf{P}(+ \mid D) = 0{,}98.$$

Como a especificidade do teste é 98%, a probabilidade de falso positivo é

$$\mathbf{P}(+ \mid D^c) = 1 - 0{,}98 = 0{,}02.$$

Pelo Teorema de Bayes,

$$\mathbf{P}(D \mid +) = \frac{\mathbf{P}(+ \mid D)\mathbf{P}(D)}{\mathbf{P}(+ \mid D)\mathbf{P}(D) + \mathbf{P}(+ \mid D^c)\mathbf{P}(D^c)}.$$

Substituindo os valores, obtemos

$$\mathbf{P}(D \mid +) = \frac{0{,}98 \cdot 0{,}005}{0{,}98 \cdot 0{,}005 + 0{,}02 \cdot 0{,}995}.$$

Portanto,

$$\mathbf{P}(D \mid +) = \frac{0{,}0049}{0{,}0049 + 0{,}0199} = \frac{0{,}0049}{0{,}0248} \approx 0{,}1976.$$

Assim, mesmo após um teste positivo, a probabilidade de que a pessoa esteja realmente doente é de aproximadamente

$$19{,}76\%.$$

Esse resultado pode parecer surpreendente à primeira vista. A razão é que a doença é muito rara. Embora o teste seja bastante confiável, existem muitas pessoas saudáveis na população. Como consequência, mesmo uma pequena taxa de falsos positivos pode produzir um número considerável de testes positivos entre pessoas que não estão doentes.

Uma maneira intuitiva de enxergar isso é imaginar uma população de 10 000 pessoas. Como a prevalência da doença é de 0,5%, esperamos que cerca de 50 pessoas estejam doentes. Dessas, aproximadamente

$$0{,}98 \cdot 50 = 49$$

testariam positivo.

Por outro lado, haveria cerca de 9 950 pessoas saudáveis. Como a taxa de falso positivo é de 2%, aproximadamente

$$0{,}02 \cdot 9\,950 = 199$$

pessoas saudáveis também testariam positivo.

Portanto, entre os resultados positivos, teríamos aproximadamente 49 pessoas realmente doentes e 199 pessoas saudáveis. Logo, a proporção de pessoas realmente doentes entre aquelas que testaram positivo seria

$$\frac{49}{49+199} = \frac{49}{248} \approx 0{,}1976.$$

Isso mostra um ponto importante sobre testes de triagem para doenças raras: mesmo testes com alta sensibilidade e alta especificidade podem produzir muitos falsos positivos quando a prevalência da doença é baixa.

◁

Como $E = (E \cap F) \cup (E \cap F^c)$, com união disjunta, obtemos a forma mais simples da regra da probabilidade total.

Teorema 2.3 — Probabilidade total

Se $0 < \mathbf{P}(F) < 1$, então, para todo evento E,

$$\mathbf{P}(E) = \mathbf{P}(E \mid F)\mathbf{P}(F) + \mathbf{P}(E \mid F^c)\mathbf{P}(F^c).$$

Mais geralmente, se $F_1, \ldots, F_n$ formam uma partição de Ω e $\mathbf{P}(F_i) > 0$ para todo i, então

$$\mathbf{P}(E) = \sum_{i=1}^{n} \mathbf{P}(E \mid F_i)\mathbf{P}(F_i).$$

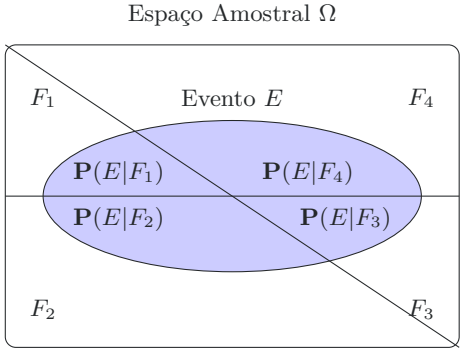


Figura 2.1: A partição de Ω separa o evento E nas partes $E \cap F_1, \ldots, E \cap F_4$.

De fato, E é a união disjunta dos eventos $E \cap F_i$; basta somar suas probabilidades e usar $\mathbf{P}(E \cap F_i) = \mathbf{P}(E \mid F_i)\mathbf{P}(F_i)$.

Exemplo 2.4.

Considere três recipientes, cada um contendo 100 esferas:

- O Recipiente 1 contém 75 esferas vermelhas e 25 azuis.
- O Recipiente 2 contém 60 esferas vermelhas e 40 azuis.
- O Recipiente 3 contém 45 esferas vermelhas e 55 azuis.

Um recipiente é selecionado aleatoriamente e, em seguida, uma esfera é escolhida do recipiente selecionado, também de forma aleatória. A tarefa é determinar a probabilidade de que a esfera escolhida seja vermelha.

Solução.

Seja R o evento de que a esfera escolhida seja vermelha, e B_i o evento de que o Recipiente i foi escolhido. As probabilidades condicionais são conhecidas:

$$\begin{aligned}\mathbf{P}\left(R \mid B_1\right) &= 0{,}75,\\ \mathbf{P}\left(R \mid B_2\right) &= 0{,}60,\\ \mathbf{P}\left(R \mid B_3\right) &= 0{,}45.\end{aligned}$$

A partição dos recipientes é dada por B_1, B_2, B_3. É importante observar que essa partição é válida, pois os eventos B_i são mutuamente exclusivos (apenas um pode ocorrer) e a união deles abrange todo o espaço amostral, ou seja, $\mathbf{P}\left(B_1 \cup B_2 \cup B_3\right) = 1$.

Utilizando a lei da probabilidade total, a probabilidade de escolher uma esfera vermelha pode ser expressa como:

$$\begin{aligned}\mathbf{P}(R) &= \mathbf{P}\left(R \mid B_1\right)\mathbf{P}\left(B_1\right) + \mathbf{P}\left(R \mid B_2\right)\mathbf{P}\left(B_2\right) + \mathbf{P}\left(R \mid B_3\right)\mathbf{P}\left(B_3\right)\\ &= (0{,}75)\cdot\frac{1}{3} + (0{,}60)\cdot\frac{1}{3} + (0{,}45)\cdot\frac{1}{3}\\ &= 0{,}60.\end{aligned}$$

Assim, a probabilidade de que a esfera escolhida seja vermelha é de 60%.

◁

Exemplo 2.5 — Monty Hall.

Um participante escolhe uma entre três portas. Atrás de uma delas há um carro; atrás das outras, cabras. O apresentador sabe onde está o carro, sempre abre uma das portas não escolhidas que contém uma cabra e sempre oferece a troca. Se puder escolher entre duas portas com cabras, decide ao acaso entre elas. É melhor manter a escolha inicial ou trocar?

◁

A priori, cada porta tem probabilidade 1/3 de esconder o carro. A regra do apresentador, porém, fornece informação sem redistribuir igualmente essa probabilidade entre as duas portas fechadas.

Acontece, porém, que o apresentador do programa fornece-lhe informações adicionais e por isso importa! Sem perda de generalidade, podemos assumir que você selecionou a porta A como sua primeira escolha. Existem três possibilidades igualmente prováveis: o carro está atrás da porta A, o carro está atrás da porta B, ou o carro está atrás da porta C. Vamos considerar cada um destas possibilidades separadamente.

Se o carro está atrás da porta A, em seguida, o apresentador pode abrir qualquer uma das outros duas portas. E, neste caso, a resposta correta (em retrospectiva) é ficar com a porta inicial.

Se o carro está atrás da porta B, em seguida, o apresentador vai abrir necessariamente a porta C. Neste caso, a resposta correta é trocar de porta.

Da mesma forma, se o carro está atrás da porta C, em seguida, o apresentador vai abrir a porta$B$, e novamente a resposta correta é trocar de porta.

Assim, em dois dos três casos igualmente prováveis, a estratégia correta é mudar de porta. Isto significa que são mais propensos a ganhar os que alternam de porta.

Podemos formalizar a solução do problema de Monty Hall pelo Teorema de Bayes.

Como antes, vamos supor sem perda de generalidade que a porta que você selecionou é denominada porta A. Além disso, vamos denominar a porta que o apresentador abre como B. Assim temos:

$$\mathbf{P}(\text{prêmio em A} \mid \text{apres. abre B})$$
$$= \frac{\mathbf{P}(\text{apres. abre B} \mid \text{prêmio em A})\mathbf{P}(\text{prêmio em A})}{\mathbf{P}(\text{apres. abre B})} = \frac{\frac{1}{2}\cdot\frac{1}{3}}{\frac{1}{2}} = \frac{1}{3},$$

$$\mathbf{P}(\text{prêmio em C} \mid \text{apres. abre B})$$
$$= \frac{\mathbf{P}(\text{apres. abre B} \mid \text{prêmio em C})\mathbf{P}(\text{prêmio em C})}{\mathbf{P}(\text{apres. abre B})} = \frac{1\cdot\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}.$$

Logo é vantajoso trocar de porta.

Exemplo 2.6.

Uma urna contém b bolas azuis e c bolas vermelhas. Uma bola é retirada ao acaso, sua cor é observada, e esta bola é devolvida para a urna juntamente com outras d bolas da mesma cor. Este procedimento é repetido indefinidamente. Qual é a probabilidade de que:

a. A segunda bola é tirada vermelha?

b. A primeira bola tirada é vermelha dado que a segunda bola tirada é vermelha?

◁

[a] Denotaremos por V_n o evento que a bola retirada na n-ésima etapa é vermelha. Então

$$\mathbf{P}(V_2) = \mathbf{P}(V_2|V_1)\mathbf{P}(V_1) + \mathbf{P}(V_2|V_1^{\mathsf{c}})\mathbf{P}(V_1^{\mathsf{c}}).$$

Assim dado que ocorreu V_1, ou seja a primeira vola retirada é vermelha, sabemos que a urna contém $c+d$ bolas vermelha para a segunda retirada e assim

$$\mathbf{P}(V_2|V_1) = \frac{c+d}{b+c+d}$$

De modo análogo, dado que ocorreu V_1^{c}, sabemos que a urna contém c bolas vermelha para a segunda retirada e assim

$$\mathbf{P}(V_2|V_1^{\mathsf{c}}) = \frac{c}{b+c+d}$$

E assim

$$\mathbf{P}(V_2) = \frac{c+d}{b+c+d}\frac{c}{b+c} + \frac{c}{b+c+d}\frac{b}{b+c} = \frac{c}{b+c} = \mathbf{P}(V_1).$$

[b]

$$\mathbf{P}(V_1 \mid V_2) = \frac{\mathbf{P}(V_1 \cap V_2)}{\mathbf{P}(V_2)} = \frac{\mathbf{P}(V_2 \mid V_1)\mathbf{P}(V_1)}{\mathbf{P}(V_2)} = \frac{c+d}{b+c+d}.$$

2.3 Independência

Agora podemos voltar à pergunta que motivou o capítulo. Quando a informação de que B ocorreu não muda em nada a probabilidade de A? Nesse caso, condicionar não produz informação nova sobre A:

$$\mathbf{P}(A \mid B) = \mathbf{P}(A).$$

Essa é a forma mais intuitiva de independência, embora ainda não seja a melhor definição formal.

Definição 2.2 — Eventos Independentes - Definição Provisória

Dado um espaço de probabilidade $(\Omega,\mathcal{F},\mathbf{P})$, dois eventos A e B com $\mathbf{P}(A) > 0$ e $\mathbf{P}(B) > 0$ são ditos **independentes** se $\mathbf{P}(A|B) = \mathbf{P}(A)$.

Essa definição, embora útil, é limitada a eventos com probabilidade positiva e apresenta uma aparente assimetria. A expressão $\mathbf{P}(A|B) = \mathbf{P}(A)$ não reflete a simetria presente na afirmação "A e B são independentes". De fato, com base apenas nessa expressão, poderia parecer mais razoável afirmar que A é independente de B, mas não necessariamente o contrário.

No entanto, se $\mathbf{P}(A|B) = \mathbf{P}(A) > 0$, então podemos deduzir que

$$\mathbf{P}(B|A) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(A)} = \frac{\mathbf{P}(A|B)\mathbf{P}(B)}{\mathbf{P}(A)} = \frac{\mathbf{P}(A)\mathbf{P}(B)}{\mathbf{P}(A)} = \mathbf{P}(B),$$

justificando assim a simetria da definição.

Mais importante ainda, para eventos com probabilidade positiva, temos que $\mathbf{P}(A|B) = \mathbf{P}(A)$ se, e somente se, $\frac{\mathbf{P}(A\cap B)}{\mathbf{P}(B)} = \mathbf{P}(A)$. Isso implica que

$$\mathbf{P}(A|B) = \mathbf{P}(A) \Leftrightarrow \mathbf{P}(A \cap B) = \mathbf{P}(A) \cdot \mathbf{P}(B).$$

Como a igualdade à direita não depende de $\mathbf{P}(A) > 0$ e $\mathbf{P}(B) > 0$, conseguimos uma definição mais geral e elegante de independência.

Definição 2.3 — Eventos Independentes

Dado um espaço de probabilidade $(\Omega,\mathcal{F},\mathbf{P})$, dois eventos A e B são ditos **independentes** se

$$\mathbf{P}(A \cap B) = \mathbf{P}(A) \cdot \mathbf{P}(B).$$

Exemplo 2.7.

Uma carta é selecionada aleatoriamente de um baralho comum de 52 cartas. Se E é o evento em que a carta selecionada é um ás e F é o evento em que a carta selecionada é vermelha, os eventos E e F são independentes?

◁

Os eventos E e F são independentes. Isso ocorre porque $\mathbf{P}(E \cap F) = 2/52$, enquanto $\mathbf{P}(E) = 4/52$ e $\mathbf{P}(F) = 1/2$.

Exemplo 2.8.

Um sistema formado por n componentes é denominado sistema em paralelo se ele funcionar quando pelo menos um dos componentes funcionar. Em um sistema como esse, é usual supor que o evento do i-ésimo componente funcionar é independente dos demais componentes. Qual a probabilidade do sistema funcionar?

◁

Seja A_i o evento em que o componente i funciona. Então

$$\begin{aligned} \mathbf{P}(\text{sistema funciona}) &= 1 - \mathbf{P}(\text{nenhum componente funciona}) && (2.1)\\ &= 1 - \mathbf{P}(\bigcap A_i^{\mathsf{c}}) && (2.2)\\ &= 1 - \prod_{i=1}^{n}(1 - p_i) && (2.3) \end{aligned}$$

Em resumo, as duas definições são equivalentes quando $\mathbf{P}(A), \mathbf{P}(B) > 0$, mas a segunda é preferível, pois abrange eventos para os quais $\mathbf{P}(A) = 0$ ou $\mathbf{P}(B) = 0$. Além disso, essa definição deixa claro que, se A é independente de B, então B também é independente de A.

Proposição 2.4

Um evento A é independente dele mesmo se, e somente se, $\mathbf{P}(A) = 0$ ou $\mathbf{P}(A) = 1$.

Demonstração.

$$\mathbf{P}(A) = \mathbf{P}(A \cap A) = \mathbf{P}(A)^2.$$

Resolvendo essa equação do segundo grau em $\mathbf{P}(A)$, concluímos que $\mathbf{P}(A) = 0$ ou $\mathbf{P}(A) = 1$. ■

Exemplo 2.9.

Considere o experimento de rolar dois dados justos. O espaço amostral é dado por $\Omega = \{(i,j) : i,j \in \{1,2,3,4,5,6\}\}$ e $\mathbf{P}$ é a medida de probabilidade uniforme em $(\Omega, 2^\Omega)$, tal que

$$\mathbf{P}(\{(i,j)\}) = \frac{1}{36},$$

para todo $(i,j) \in \Omega$.

Definimos os seguintes eventos:

$$A = \{\text{observamos 2 no primeiro dado}\} = \{(2,j) : j \in \{1,2,3,4,5,6\}\},$$
$$B = \{\text{observamos 4 no segundo dado}\} = \{(i,4) : i \in \{1,2,3,4,5,6\}\},$$
$$C = \{\text{a soma dos valores observados é 7}\} = \{(i,j) \in \Omega : i + j = 7\}.$$

Observe que $\mathbf{P}(A) = \mathbf{P}(B) = \mathbf{P}(C) = \frac{1}{6}$. Além disso,

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

e, analogamente, $\mathbf{P}(A \cap C) = \mathbf{P}(A)\mathbf{P}(C)$ e $\mathbf{P}(B \cap C) = \mathbf{P}(B)\mathbf{P}(C)$.

Isso mostra que os eventos A e B são independentes, assim como os eventos A e C e os eventos B e C.

◁

A definição de independência para mais de dois eventos é um pouco mais sutil. No exemplo anterior, descrevemos três eventos, cada par dos quais é independente, mas não é razoável afirmar que o trio A, B, C é independente. De fato, saber que o primeiro dado mostra 2 e o segundo 4 implica que a soma é 6, e portanto,

$$\mathbf{P}(C|A \cap B) = 0 \neq \mathbf{P}(C).$$

Isso demonstra que a independência entre pares, embora necessária, não é uma condição suficiente para a independência de múltiplos eventos.

Para explorar essa questão, considere o exercício a seguir:

Exercício 2.1

Construa um exemplo de espaço de probabilidade e três eventos A, B e C tais que $\mathbf{P}(A \cap B \cap C) = \mathbf{P}(A)\mathbf{P}(B)\mathbf{P}(C)$, mas os pares de eventos não são todos independentes.

A solução para esse problema é apresentada na seguinte definição.

Definição 2.4

Um conjunto finito de eventos A_i é dito **independente dois a dois** se e somente se cada par de eventos é independente, isto é, se e somente se para todos os pares distintos dos índices m, k,

$$\mathbf{P}(A_m \cap A_k) = \mathbf{P}(A_m)\mathbf{P}(A_k).$$

Um conjunto finito $A_1, \ldots, A_n$ é dito **independente** se, para todo subconjunto não vazio $J \subseteq \{1, \ldots, n\}$,

$$\mathbf{P}\left(\bigcap_{j \in J} A_j\right) = \prod_{j \in J} \mathbf{P}(A_j).$$

Definição 2.5

Experimentos aleatórios independentes, com apenas dois resultados possíveis — sucesso ou fracasso — e probabilidades de sucesso e fracasso constantes, são chamados ensaios de Bernoulli.

Exemplo 2.10 — Ensaios de Bernoulli.

Deve-se realizar uma sequência infinita de tentativas independentes. Cada tentativa tem probabilidade de sucesso p e probabilidade de fracasso $1 - p$. Qual é a probabilidade de que

1. pelo menos 1 sucesso ocorra nas primeiras n tentativas;
2. exatamente k sucessos ocorram nas primeiras n tentativas;
3. todas as tentativas resultem em sucessos?

◁

1. Para determinar a probabilidade de pelo menos 1 sucesso nas primeiras n tentativas, notamos que é mais simples calcular primeiro a probabilidade do evento complementar: de que nenhum sucesso ocorra nas primeiras n tentativas. Se E, representar o evento fracasso na i-ésima tentativa, então a probabilidade de não ocorrer sucessos é, por independência, $(1-p)^n$ Portanto, a resposta é $1-(1-p)^n$.

2. considere qualquer sequência na qual dentre os primeiros n resultados ocorram k sucessos e $n-k$ fracassos.

   Cada uma dessas sequências, pela hipótese de independência das tentativas, irá ocorrer com probabilidade

   $$p^k(1-p)^{n-k}.$$

   Como existem $\binom{n}{k}$ sequências como essa, a probabilidade desejada é

   $$\mathbf{P}(\text{exatamente k sucessos}) = \binom{n}{k} p^k (1-p)^{n-k}$$

   Para responder ao item 3, notamos, pela letra 1, que a probabilidade de que as primeiras n tentativas resultem em sucesso é dada por

   $$\mathbf{P}\left(\bigcap_{i=1}^{n} E_i^c\right) = p^n$$

   Assim, usando a propriedade da continuidade das probabilidades, vemos que a probabilidade desejada é dada por

   $$\begin{aligned}\mathbf{P}\left(\bigcap_{i=1}^{\infty} E_i^c\right) &= \mathbf{P}\left(\lim_{n\to\infty} \bigcap_{i=1}^{n} E_i^c\right) \\ &= \lim_{n\to\infty} \mathbf{P}\left(\bigcap_{i=1}^{n} E_i^c\right) \\ &= \lim_n p^n = \begin{cases} 0 \text{ se } p < 1 \\ 1 \text{ se } p = 1 \end{cases}\end{aligned}$$

Exemplo 2.11 — Ruína do jogador.
Em cada rodada de um jogo justo, um jogador ganha ou perde um real, com probabilidade 1/2 em cada caso. Seu capital inicial é o inteiro x, com $0 \leq x \leq m$, e ele joga até o capital atingir 0 ou m. Qual é a probabilidade de ruína?

◁

A probabilidade de ruína depende tanto do capital inicial x quanto do montante final desejado m.

Seja $p(x)$ a probabilidade de o jogador ir à ruína se ele começar com um capital de x dólares. Então, a probabilidade de ruína, uma vez que o jogador vence a primeira rodada, é $p(x+1)$, pois o capital do jogador passa a ser $x+1$ se ele vencer a primeira rodada. De forma similar, a probabilidade de ruína, dado que o jogador perde a primeira rodada, é $p(x-1)$, já que o capital do jogador se torna $x-1$ se ele perder a primeira rodada.

Em outras palavras, seja B_1 o evento de que o jogador vence a primeira rodada, B_2 o evento de que o jogador perde a primeira rodada, e R o evento de ruína.

Então temos:

$$\mathbf{P}(R|B_1) = p(x+1) \quad \mathbf{P}(R|B_2) = p(x-1)$$

Sabemos também que:

$$\mathbf{P}(B_1) = \frac{1}{2} \quad \text{e} \quad \mathbf{P}(B_2) = \frac{1}{2}$$

Pela fórmula da probabilidade total temos:

$$\mathbf{P}(R) = \mathbf{P}(R|B_1)\mathbf{P}(B_1) + \mathbf{P}(R|B_2)\mathbf{P}(B_2)$$

o que implica em:

$$p(x) = \frac{1}{2}\left(p(x-1) + p(x+1)\right), \qquad 1 \leq x \leq m-1,$$

e claramente

$$p(0) = 1 \qquad p(m) = 0.$$

A equação de recorrência anterior tem como solução $p(x) = c_1 + c_2 x$. As constantes podem ser determinadas pelas condições de fronteira.

$$c_1 = 1, \qquad c_1 + c_2 m = 0.$$

E logo

$$p(x) = 1 - \frac{x}{m} \quad 0 \leq x \leq m$$

Ao atravessar este capítulo

Condicionar significa atualizar o universo do cálculo. A regra do produto, a probabilidade total, Bayes e independência devem ser usados com os eventos condicionantes e suas probabilidades claramente identificados.

2.4 Exercícios

Exercício 2.2 — Teste diagnóstico

Uma condição afeta 1,5% da população. Um teste tem sensibilidade 94% e especificidade 97%. Calcule a probabilidade de uma pessoa ter a condição dado que o teste foi positivo. Explique por que o resultado não é 94%.

Exercício 2.3 — Urnas

Escolhe-se uma de três urnas com probabilidades 0,2, 0,5 e 0,3. Elas contêm, respectivamente, 3/5, 5/8 e 7/10 bolas azuis. Uma bola retirada é azul. Determine a probabilidade de ela ter vindo da segunda urna.

Exercício 2.4 — Independência

Se $\mathbf{P}(A) = 0{,}4$, $\mathbf{P}(B) = 0{,}5$ e $\mathbf{P}(A \cup B) = 0{,}7$, verifique se A e B são independentes. Determine ainda $\mathbf{P}(A^c \cap B)$ e $\mathbf{P}(A \mid A \cup B)$.

Exercício 2.5 — Ensaios de Bernoulli

Uma conexão transmite cada pacote corretamente com probabilidade 0,92, independentemente dos demais. Calcule a probabilidade de: (a) os primeiros seis pacotes chegarem corretamente; (b) ocorrer exatamente uma falha nos seis; (c) a primeira falha ocorrer no quinto pacote.

Exercício 2.6 — Ruína do jogador

Um jogador começa com 4 fichas e para ao atingir 0 ou 11. Em cada rodada ganha uma ficha com probabilidade 1/2 e perde uma com probabilidade 1/2. Use a recorrência do capítulo para calcular a probabilidade de atingir 11 antes de 0.

CAPÍTULO

3

Variáveis aleatórias

3.1 Variáveis aleatórias e funções de distribuição

Um experimento produz um resultado ω, mas quase nunca precisamos guardar toda a informação contida nele. Num conjunto de lançamentos de moeda, talvez interesse apenas o número de caras; num componente eletrônico, o tempo até a falha; num ponto do plano, a distância à origem. Em cada caso fazemos a mesma coisa: extraímos do resultado uma quantidade numérica.

Essa passagem é tão natural que vale formular a pergunta ao contrário: que propriedade uma função $X(\omega)$ precisa ter para que possamos fazer probabilidade com seus valores? Queremos, pelo menos, que perguntas como "$X \leq x$?" definam eventos. Por isso exigimos que

$$[X \leq x] = \{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$$

para todo $x \in \mathbb{R}$. Nos exemplos usuais essa condição será automática; o papel dela é garantir que expressões como $\mathbf{P}(X \leq x)$ façam sentido.

Definição 3.1

Uma **variável aleatória** X em um espaço de probabilidade $(\Omega, \mathcal{F}, \mathbf{P})$ é uma função real definida no espaço Ω tal que $[X \leq x]$ é evento para todo $x \in \mathbb{R}$; i.e., $X : \Omega \to \mathbb{R}$ é variável aleatória se $[X \leq x] \in \mathcal{F}$ para todo $x \in \mathbb{R}$.

É útil guardar a imagem $\Omega \xrightarrow{X} \mathbb{R}$: X é a regra; a aleatoriedade vem do resultado ω. Quando perguntamos se X caiu num conjunto B, voltamos ao espaço amostral pela imagem inversa,

$$[X \in B] = X^{-1}(B).$$

Exemplo 3.1 — Número de caras.

Lançamos uma moeda n vezes. Um resultado é uma sequência $\omega = (\omega_1, \ldots, \omega_n)$, com cada ω_i igual a cara ou coroa. Se X conta o número de caras, então

$$X(\omega) = \#\{i : \omega_i = \text{cara}\}.$$

A variável aleatória transforma uma sequência inteira em uma única contagem.

◁

Exemplo 3.2 — Uma transformação de um resultado contínuo.

Escolha um ponto ω uniformemente em [0,1] e defina

$$X(\omega) = \omega^2.$$

O experimento produz o ponto ω; a variável aleatória registra o seu quadrado.

◁

Exemplo 3.3 — Distância ao centro.

Escolha um ponto $\omega = (x,y)$ no disco unitário e defina

$$X(\omega) = \sqrt{x^2 + y^2}.$$

Dessa vez, um resultado bidimensional é resumido por uma única quantidade: a distância à origem.

◁

Se o próprio resultado já é um número real, podemos simplesmente tomar $X(\omega) = \omega$. Se o resultado é um ponto do plano, suas coordenadas podem ser registradas por duas variáveis X e Y; essa ideia será retomada no capítulo de vetores aleatórios.

Definição 3.2

A **função de distribuição** da variável aleatória X, representada por F_X ou simplesmente por F, é definida por

$$F_X(x) = \mathbf{P}(X \leq x), \quad x \in \mathbb{R}.$$

A função de distribuição também é chamada de **função de distribuição acumulada**. Ela guarda toda a distribuição de X. Por exemplo, se $a < b$,

$$\mathbf{P}(a < X \leq b) = F_X(b) - F_X(a).$$

Assim, depois de conhecer F_X, podemos calcular probabilidades de intervalos sem voltar ao espaço amostral original. Parte da complexidade de Ω ficou comprimida numa única função da reta real.

Exemplo 3.4.

Seja X = número de H em dois lançamentos independentes de uma moeda justa. Então temos

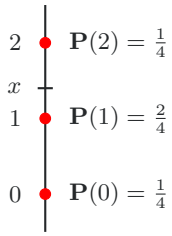


Figura 3.1: A função de distribuição acumulada no ponto x, $F_X(x)$, é a probabilidade de vermos um resultado menor que x. No caso ilustrado $F_X(x) = \frac{1}{4} + \frac{2}{4} = \frac{3}{4}$.

$$X = \begin{cases} 0, & \text{com probabilidade } \frac{1}{4} \\ 1, & \text{com probabilidade } \frac{1}{2} \\ 2, & \text{com probabilidade } \frac{1}{4} \end{cases}$$

Então, podemos criar a função de distribuição. Esta função apresentará descontinuidades do tipo salto nos pontos 0, 1 e 2.

$$F_X(x) = \mathbf{P}(X \leq x) = \begin{cases} 0, & x < 0 \\ \frac{1}{4}, & x \in [0,1) \\ \frac{1}{4} + \frac{1}{2}, & x \in [1,2) \\ \frac{1}{4} + \frac{1}{2} + \frac{1}{4}, & x \in [2, +\infty) \end{cases}$$

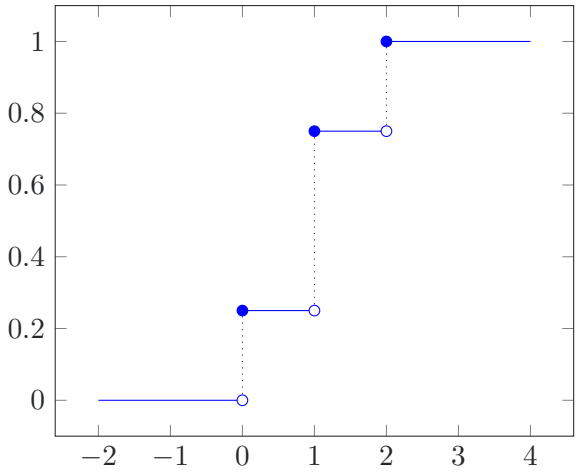


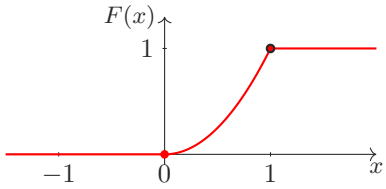
Figura 3.2: Distribuição do Exemplo 3.4

Exemplo 3.5.

Considere o experimento de sortear ao acaso um ponto no disco unitário $B = \{(a,b) : a^2 + b^2 \leq 1\}$ em $\mathbb{R}^2$, e defina X como a distância ao centro da bola.

Para modelar este problema podemos considerar que a probabilidade de sortearmos um ponto em uma região $A \in \mathcal{B}(B)$ é proporcional à área de A. Mais especificamente

$$\mathbf{P}(A) = \frac{\text{àrea } (A)}{\pi}.$$



Agora, note que $X(a,b) = \sqrt{a^2 + b^2}$ e portanto para $x \in [0,1]$

$$\mathbf{P}(X \leq x) = \mathbf{P}(\{(a,b) : a^2 + b^2 \leq x^2\}) = \frac{\pi\, x^2}{\pi} = x^2,$$

e portanto

$$F(x) = \begin{cases} 0, & x < 0 \\ x^2, & 0 \le x < 1 \\ 1, & x \ge 1 \end{cases}.$$

◁

Teorema 3.1 — Propriedades da Função de Distribuição

A função de distribuição $F(x)$ de uma variável aleatória X possui as seguintes propriedades:

1. F é não decrescente e $0 \le F(x) \le 1$.
2. $F(x) \to 0$ quando $x \to -\infty$ e $F(x) \to 1$ quando $x \to +\infty$.
3. F é contínua à direita, ou seja
$$\lim_{y \to x^+} F(y) = F(x).$$
4. F possui limites à esquerda em todos os pontos. Além disso,
$$F(x^-) := \lim_{y \to x^-} F(y) = \mathbf{P}(X < x).$$
Em particular,
$$\mathbf{P}(X = x) = F(x) - F(x^-).$$
5. F possui no máximo número enumerável de descontinuidades.

Uma função F contínua à direita e com limites à esquerda é chamada **càdlàg**, abreviação da expressão francesa *continue à droite, limites à gauche*.

Demonstração. 1. Se $x < y$, então $\{X \le x\} \subseteq \{X \le y\}$. A monotonicidade de $\mathbf{P}$ mostra que $F(x) \le F(y)$; os limites $0 \le F \le 1$ seguem dos axiomas.

2. Escolha sequências $x_n \downarrow -\infty$ e $y_n \uparrow +\infty$. Então $\{X \le x_n\} \downarrow \emptyset$ e $\{X \le y_n\} \uparrow \Omega$. Pela continuidade da probabilidade, $F(x_n) \to 0$ e $F(y_n) \to 1$. Como F é monótona, esses são os limites de F nos extremos da reta.

3. Se $x_n \downarrow x$, então $\{X \le x_n\} \downarrow \{X \le x\}$. Logo $F(x_n) \to F(x)$, o que prova a continuidade à direita.

4. Se $y_n \uparrow x$, então $\{X \leq y_n\} \uparrow \{X < x\}$. Portanto

$$F(x^-) = \lim_{n\to\infty} F(y_n) = \mathbf{P}(X < x).$$

Como $\{X \leq x\}$ é a união disjunta de $\{X < x\}$ e $\{X = x\}$, segue que $\mathbf{P}(X = x) = F(x) - F(x^-)$.

5. Para $m \geq 1$, seja $D_m = \{x : F(x) - F(x^-) \geq 1/m\}$. Como a soma das probabilidades $\mathbf{P}(X = x)$ não pode exceder 1, o conjunto D_m tem no máximo m pontos. Toda descontinuidade de F pertence a algum D_m; portanto o conjunto de descontinuidades, $\bigcup_{m\geq 1} D_m$, é enumerável.

■

Proposição 3.2

$\mathbf{P}(X \in (a, b]) = F(b) - F(a)$.

Definição 3.3

Uma função $F : \mathbb{R} \to [0, 1]$ tal que

1. F é não-decrescente;
2. F é contínua pela direita;
3. $\lim_{x\to-\infty} F(x) = 0$ e $\lim_{x\to+\infty} F(x) = 1$

é denominada **função de distribuição candidata**.

Mostraremos que toda candidata a função de distribuição de probabilidade é na realidade uma função de distribuição, i.e., construiremos uma variável aleatória que possua como função distribuição a função candidata dada. Isso nos permitirá construir uma sequência de variáveis com funções de distribuição pré-determinadas, sempre no mesmo espaço de probabilidade, permitindo a análise mais precisa do comportamento da sequência.

Proposição 3.3

Dada uma função candidata a função de distribuição de probabilidade F então existe uma variável aleatória X tal que a sua função de distribuição de probabilidade é F.

Uma construção útil justifica a afirmação. Se U é uniforme em (0,1), defina

$$X = \inf\{x \in \mathbb{R} : F(x) \geq U\}.$$

Pela monotonicidade e continuidade à direita de F, os eventos $\{X \leq x\}$ e $\{U \leq F(x)\}$ coincidem. Assim $\mathbf{P}(X \leq x) = \mathbf{P}(U \leq F(x)) = F(x)$.

3.2 Tipos de variáveis aleatórias

As variáveis aleatórias podem ser classificadas em dois tipos principais: discretas e contínuas, dependendo da natureza dos valores que elas podem assumir.

Definição 3.4

1. A variável aleatória X é dita **discreta** se toma um número finito ou enumerável de valores, i.e., se existe um conjunto finito ou enumerável $\{x_1, x_2, \ldots\} \subset \mathbb{R}$ tal que $X(\omega) \in \{x_1, x_2, \ldots\}$ para todo $\omega \in \Omega$. A função $p(x_i)$ definida por $p(x_i) = \mathbf{P}(X = x_i), i = 1,2,\ldots,$ é dita função de probabilidade de X.
2. A variável aleatória X é (absolutamente) **contínua** se existe uma função $f(x) \geq 0$ tal que

$$F_X(x) = \int_{-\infty}^{x} f(t)dt, \quad \text{para todo } x \in \mathbb{R}$$

Neste caso, dizemos que f é **função de densidade de probabilidade** de X ou simplesmente **densidade** de X.

Temos uma analogia entre densidade de probabilidade e densidade de massa, que pode ser entendida comparando os conceitos de distribuição de massa física em um objeto com a distribuição de probabilidade ao longo de

um eixo.

Densidade de Massa em Uma Dimensão Imagine uma barra fina e longa de material uniforme (por exemplo, um fio ou uma haste), onde a densidade de massa $\rho(x)$ indica como a massa está distribuída ao longo do comprimento x. A densidade de massa $\rho(x)$ é medida em unidades de massa por unidade de comprimento (por exemplo, kg/m). Para encontrar a massa total M em um intervalo $[a, b]$ ao longo do fio, integramos a densidade de massa sobre esse intervalo:

$$M = \int_a^b \rho(x)\, dx.$$

Densidade de Probabilidade em Uma Dimensão A densidade de probabilidade $p(x)$ funciona de maneira semelhante, mas, em vez de medir como a massa está distribuída, ela mede como a probabilidade está distribuída ao longo de um eixo. A densidade de probabilidade $p(x)$ indica a "concentração" da probabilidade em torno de um ponto x e é adimensional. A probabilidade total P de encontrar uma variável aleatória dentro de um intervalo $[a, b]$ é dada pela integral da densidade de probabilidade nesse intervalo:

$$P = \int_a^b p(x)\, dx.$$

Exemplo 3.6.

Sorteio de Pontos Segundo uma Distribuição de Probabilidade

Considere uma distribuição de probabilidade contínua descrita por uma densidade de probabilidade $p(x)$ definida no intervalo $[0, 1]$. Para ilustrar o conceito, vamos considerar uma distribuição linear simples dada por:

$$p(x) = 2x, \quad \text{para } 0 \leq x \leq 1.$$

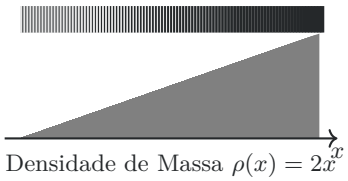


Figura 3.3: Fio com densidade de massa crescente $\rho(x) = 2x$.

Essa função $p(x)$ indica que a probabilidade de encontrar um valor próximo a x aumenta à medida que x cresce. Em outras palavras, valores de x mais

próximos de 1 são mais prováveis de serem escolhidos do que valores próximos de 0. Essa é uma situação análoga a ter uma barra de densidade de massa que se torna mais densa à medida que nos aproximamos de uma das extremidades. Podemos verificar que $p(x)$ é uma densidade de probabilidade válida, pois ela satisfaz a condição de normalização:

$$\int_0^1 2x\,dx = \Big[x^2\Big]_0^1 = 1.$$

Agora, vamos calcular a probabilidade de a variável estar dentro de um subintervalo específico $[a, b]$, onde $0 \leq a < b \leq 1$. Isso pode ser obtido integrando a densidade de probabilidade $p(x)$ nesse intervalo:

$$\mathbf{P}(a \leq x \leq b) = \int_a^b 2x\,dx = \Big[x^2\Big]_a^b = b^2 - a^2.$$

◁

Gráficos das Funções de Distribuição de Variáveis Aleatórias Discreta, Contínua e Mista Dependendo da natureza da variável aleatória — discreta, contínua ou mista — a forma da função de distribuição apresenta características distintas:

- Discreta: Para variáveis aleatórias discretas, a função de distribuição possui uma estrutura em degraus. Isso ocorre porque a probabilidade se acumula em valores específicos, resultando em saltos na função. Entre esses saltos, a função de distribuição permanece constante, indicando que a variável não assume valores intermediários.
- Contínua: Para variáveis aleatórias contínuas, a função de distribuição é uma função contínua, sem qualquer tipo de salto. Isso reflete o fato de que a probabilidade é distribuída de maneira contínua ao longo de um intervalo de valores. A continuidade da função indica que a variável pode assumir qualquer valor dentro de um intervalo onde a função de distribuição é crescente.função.
- Mista: A função de distribuição de uma variável aleatória mista combina as características das duas anteriores. Ela apresenta trechos contínuos não constantes, correspondentes a intervalos onde a variável pode assumir um conjunto infinito de valores, e saltos em pontos específicos, onde existe uma probabilidade acumulada discreta.

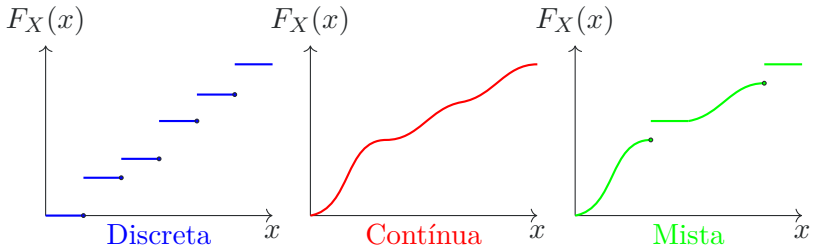


Figura 3.4: Funções de Distribuição de Variáveis Aleatórias Discreta, Contínua e Mista

Exemplo 3.7.

Determine para que valores da constante $c \in \mathbb{R}$, a função abaixo representa uma densidade de probabilidade.

$$f(x) = \begin{cases} c(1-x)^2, & \text{se } 0 \leq x < 1/2; \\ 1/(c+1), & \text{se } 1/2 \leq x \leq 1; \\ 0, & \text{caso contrário.} \end{cases}$$

Para que f seja uma densidade, precisamos de $f(x) \geq 0$ e $\int_{-\infty}^{\infty} f(x)\,dx = 1$. Se $c \geq 0$, a primeira condição já está satisfeita; resta impor a normalização:

$$\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{0} 0dx + \int_{0}^{1/2} c(1-x)^2 dx + \int_{1/2}^{1} \frac{1}{c+1} dx + \int_{1}^{\infty} 0dx = 1,$$

que resulta em

$$c\frac{-(1-x)^3}{3}\Big|_0^{1/2} + \frac{x}{c+1}\Big|_{1/2}^{1} = 1 \Rightarrow 7c^2 - 17c - 12 = 0.$$

A solução negativa dessa equação é descartada e obtemos $c = 3$.

◁

Proposição 3.4

1. Se X é discreta, então $[X \leq x] = \bigcup_{x_i \leq x} [X = x_i]$, logo

$$F_X(x) = \sum_{x_i \leq x} \mathbf{P}\left(X = x_i\right) = \sum_{x_i \leq x} p\left(x_i\right)$$

2. Se X é absolutamente contínua, então F_X, sendo uma integral indefinida de f, é contínua.

Uma função $f(x) \geq 0$ é densidade de alguma variável aleatória se, e somente se, $\int_{-\infty}^{\infty} f(x)dx = 1$, já que neste caso F definida por

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

é função de distribuição. Reciprocamente, se f é densidade então $\int_{-\infty}^{\infty} f(x)dx = 1$.

Como podemos verificar se X tem densidade? O critério seguinte cobre as funções de distribuição regulares encontradas neste curso.

Proposição 3.5

Uma variável aleatória X cuja função de distribuição F_X seja

a. contínua e

b. continuamente diferenciável por partes, em uma partição finita da reta, com derivada integrável,

possui densidade dada por

$$f(x) = \begin{cases} F_X'(x), & \text{se } F_X \text{ é diferenciável em } x, \\ 0, & \text{caso contrário} \end{cases}$$

O valor de f nos poucos pontos onde F_X não é diferenciável é irrelevante para a integral. Em cada trecho, o Teorema Fundamental do Cálculo fornece a variação de F_X; somando os trechos e usando $\lim_{x \to -\infty} F_X(x) = 0$, obtemos

$$F_X(a) = \int_{-\infty}^{a} f(x)\, dx.$$

Por exemplo, seja

$$F_X(x) = \begin{cases} 0, & x < 0 \\ x, & 0 \leq x \leq 1 \\ 1, & x > 1 \end{cases}$$

Então X tem densidade, pois F_X é contínua e

$$F'_X(x) = \begin{cases} 1, & x \in (0,1) \\ 0, & x \notin [0,1] \end{cases}$$

Logo a densidade de X é dada por

$$f(x) = F'_X(x) = \begin{cases} 1, & x \in (0,1), \\ 0, & x < 0 \text{ ou } x > 1. \end{cases}$$

O valor de f nos pontos 0 e 1 é arbitrário, pois qualquer que seja $f(0)$ e $f(1)$, a integral $\int_{-\infty}^{x} f(t)dt$ é ainda igual a $F_X(x)$. Costuma-se definir ou $f(0) = f(1) = 1$ ou $f(0) = f(1) = 0$.

Por outro lado, suponha que

$$F_X(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases}$$

Aqui X não tem densidade, pois F_X não é contínua. De fato X é uma variável aleatória discreta, e $\mathbf{P}(X = 0) = 1$.

É fácil construir um exemplo de variável aleatória que não é discreta nem absolutamente contínua, mas sim uma mistura dos dois tipos.

Exemplo 3.8 — Função de Distribuição Mista.

A variável aleatória X tem função de distribuição dada por:

$$F(x) = \begin{cases} 0, & \text{se } x < 0 \\ x/4, & \text{se } 0 \leq x < 1 \\ 1/2, & \text{se } 1 \leq x < 2; \\ 1, & \text{se } x \geq 2. \end{cases}$$

O trecho inclinado entre 0 e 1 corresponde à parte com densidade. Os saltos em 1 e 2, de tamanhos 1/4 e 1/2, correspondem às massas pontuais.

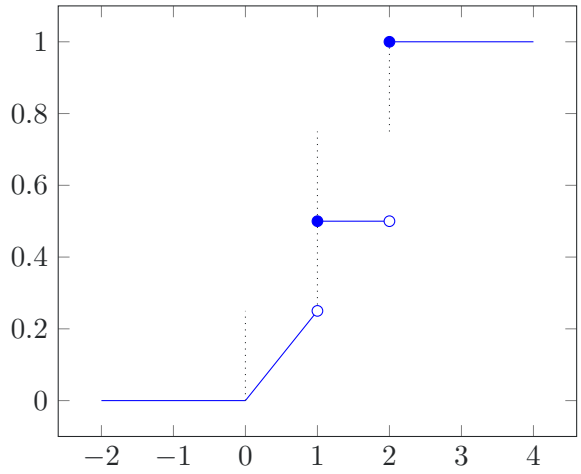


Figura 3.5: Função de distribuição mista do Exemplo 3.8.

◁

Teorema 3.6 — Decomposição mista sob hipóteses regulares

Seja F uma função de distribuição com saltos apenas nos pontos $x_1, \ldots, x_r$. Escreva

$$p_i = F(x_i) - F(x_i^-), \qquad H(x) = F(x) - \sum_{x_i \le x} p_i.$$

Suponha que H seja contínua, continuamente diferenciável por partes em uma partição finita da reta e que sua derivada $h = H'$ seja integrável. Então F é uma mistura de uma distribuição discreta e uma distribuição com densidade:

$$F = \alpha F_d + (1-\alpha) F_{ac}, \qquad \alpha = \sum_{i=1}^{r} p_i.$$

Quando $\alpha > 0$ e $\alpha < 1$, as duas funções de distribuição normalizadas são

$$F_d(x) = \frac{1}{\alpha} \sum_{x_i \le x} p_i, \qquad F_{ac}(x) = \frac{1}{1-\alpha} \int_{-\infty}^{x} h(t)\, dt.$$

Se $\alpha = 0$ ou $\alpha = 1$, permanece apenas a componente não nula.

Demonstração. A subtração da função-degrau remove precisamente os saltos de F, de modo que H é a parte contínua. Como H é continuamente diferenciável por partes e $h = H'$ é integrável, o Teorema Fundamental do Cálculo, aplicado em cada trecho, dá

$$H(x) = \int_{-\infty}^{x} h(t)\,dt.$$

Além disso, $h \geq 0$ nos pontos de diferenciabilidade, pois H é não decrescente. Fazendo $x \to \infty$, obtemos $\int_{-\infty}^{\infty} h(t)\,dt = 1-\alpha$. Logo as normalizações exibidas no enunciado são funções de distribuição, e a identidade decorre diretamente da definição de H. ∎

A discussão acima dá um método de decompor F em suas partes discreta e absolutamente contínua. Consideremos um exemplo.

Exemplo 3.9.

Seja

$$F_Y(x) = \begin{cases} 0, & x < 0 \\ x, & 0 \leq x < \frac{1}{2} \\ 1, & x \geq \frac{1}{2} \end{cases}$$

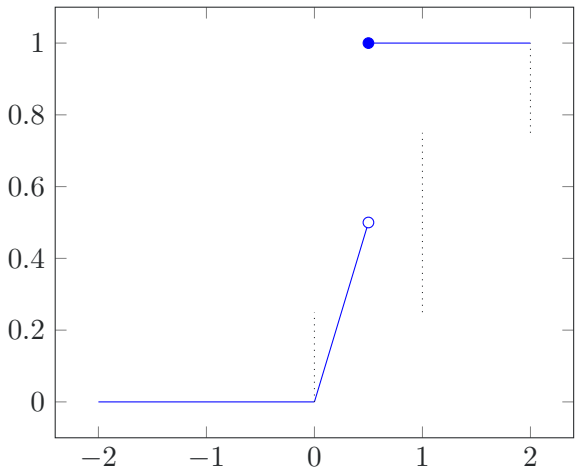


Figura 3.6: Distribuição do Exemplo 3.9

F_Y tem apenas um salto, em $x = 1/2$, e $p_1 =$ salto no ponto $1/2 = 1/2$. Logo

$$F_d(x) = \begin{cases} 0, & x < \frac{1}{2} \\ \frac{1}{2}, & x \geq \frac{1}{2} \end{cases}$$

Diferenciando, temos

$$f(x) = F_Y'(x) = \begin{cases} 0, & x < 0 \quad \text{ou} \quad x > \frac{1}{2} \\ 1, & 0 < x < \frac{1}{2} \end{cases}$$

$$f(x) = 0 \quad \text{se} \quad x = 0 \quad \text{ou} \quad 1/2 \text{ (por definição).}$$

Logo

$$F_{ac}(x) = \int_{-\infty}^{x} f(t)dt = \begin{cases} 0, & x \leq 0 \\ x, & 0 < x \leq \frac{1}{2} \\ \frac{1}{2}, & x > \frac{1}{2} \end{cases}$$

Como $F_d + F_{ac} = F_Y$, metade da probabilidade está concentrada no ponto $1/2$ e a outra metade tem densidade constante no intervalo [0,1/2]. Normalizando as componentes, temos $F_Y = \frac{1}{2}F_d^* + \frac{1}{2}F_{ac}^*$.

◁

3.3 Distribuição de uma variável aleatória

Até aqui descrevemos X por eventos do tipo $[X \leq x]$. Mas, uma vez que a variável foi definida, podemos perguntar mais: qual é a probabilidade de X cair num intervalo, numa união de intervalos ou em outro conjunto usual da reta? A condição da definição é suficiente para garantir que essas perguntas também estejam bem definidas. Em particular:

Proposição 3.7

Se X é variável aleatória em $(\Omega, \mathcal{F}, \mathbf{P})$, então o evento

$$[X \in B] := \{\omega \in \Omega : X(\omega) \in B\}$$

é evento para todo boreliano B, i.e.,

$$[X \in B] \in \mathcal{F}, \text{ para todo } B \in \mathcal{B} = \sigma\text{-álgebra de Borel.}$$

Proposição 3.8

Seja $\mathbf{P}_X(B) = \mathbf{P}(X \in B)$, para B boreliano, então $\mathbf{P}_X$ é uma probabilidade em $\mathcal{B}$, isto é, $\mathbf{P}_X(B)$ satisfaz::

Axioma 1. $\mathbf{P}_X(B) = \mathbf{P}(X \in B) \geq 0$.

Axioma 1. $\mathbf{P}_X(\mathbb{R}) = \mathbf{P}(X \in \mathbb{R}) = 1$.

Axioma 3. Se $B_1, B_2, \ldots \in \mathcal{B}$ são disjuntos, então

$$\mathbf{P}_X\left(\cup B_n\right) = \mathbf{P}\left(X \in \cup B_n\right) = \mathbf{P}\left(\cup\left[X \in B_n\right]\right) = \\ = \sum_n \mathbf{P}\left(X \in B_n\right) = \sum_n \mathbf{P}_X\left(B_n\right).$$

Definição 3.5

A probabilidade $\mathbf{P}_X$, definida na σ-álgebra de Borel por $\mathbf{P}_X(B) = \mathbf{P}(X \in B)$, é denominada **distribuição** de X.

Nos casos discreto e contínuo, podemos descrever a distribuição por meio da função de probabilidade ou densidade:

Proposição 3.9

1. Se a variável aleatória X é discreta e toma valores somente no conjunto $\{x_1, x_2, \ldots\}$, então

$$\mathbf{P}_X(B) = \sum_{x_i \in B} \mathbf{P}\left(X = x_i\right) = \sum_{x_i \in B} p\left(x_i\right), \quad \text{para todo } B \in \mathcal{B}.$$

2. Se X é absolutamente contínua com densidade $f(x)$, então

$$\mathbf{P}_X(B) = \int_B f(x)dx, \quad \text{para todo } B \in \mathcal{B}.$$

A distribuição de X é determinada por qualquer das seguintes funções:

1. A função de distribuição F_X.
2. A densidade $f(x)$, se X é absolutamente contínua.
3. A função de probabilidade $p(x_i)$, no caso discreto.

Para conhecer a distribuição de X, tanto faz conhecer qualquer das suas representações. Costuma-se escolher a representação mais conveniente para descrever a distribuição de uma dada variável aleatória. No caso contínuo, esta é geralmente a densidade.

3.4 Esperança

Conhecer toda a distribuição é muita informação. E se quisermos apenas um valor que indique onde a variável tende a se concentrar? A primeira resposta é a esperança: uma média ponderada pelos pesos probabilísticos.

Definição 3.6

Definimos a **esperança** ou **valor esperado** de uma variável aleatória como:

- $\mathbf{E}[X] = \sum_x xp(x)$
- $\mathbf{E}(X) = \int_{-\infty}^{+\infty} xf(x)\,\mathrm{d}x$

A esperança $\mathbf{E}[X]$ é a média ponderada dos possíveis valores. Outro modo de ver a esperança é como o centro de gravidade de um conjunto de partículas cuja x-ésima partícula esta na posição x e tem massa $p(x)$.

Exemplo 3.10.

Determine o valor esperado quando rolamos um dado honesto. denominado

$$\mathbf{E}[X] = \frac{1}{6} + \frac{2}{6} + \frac{3}{6} + \frac{4}{6} + \frac{5}{6} + \frac{6}{6}$$

◁

Exemplo 3.11.

Seja X uma variável aleatória que pode receber os valores $-2, -1,0,1$ e 3 com respectivas probabilidades

$$\mathbf{P}[X=-2]=0{,}2 \quad \mathbf{P}[X=-1]=0{,}4 \quad \mathbf{P}[X=0]=0{,}2,$$
$$\mathbf{P}[X=1]=0{,}1 \quad \mathbf{P}[X=3]=0{,}1.$$

Calcule $\mathbf{E}[X^2]$ e $\mathbf{E}[X]^2$

◁

Teorema 3.10

Seja X uma variável aleatória, e $g:\mathbb{R}\to\mathbb{R}$. Então a esperança de $g(X)$ é dada por

$$\mathbf{E}[g(X)]=\sum_x g(x)p(x)$$

sempre que a última soma estiver bem definida.

Demonstração.

$$\begin{aligned}
\mathbf{E}[g(X)] &= \sum_y y\mathbf{P}[g(X)=y]=\sum_y y\sum_{x:g(x)=y}\mathbf{P}[X=x] && (3.1)\\
&= \sum_y\sum_{x:g(x)=y} y\mathbf{P}[X=x]=\sum_y\sum_{x:g(x)=y} g(x)\mathbf{P}[X=x] && (3.2)\\
&= \sum_x g(x)\mathbf{P}[X=x] && (3.3)
\end{aligned}$$

■

Corolário 3.11

Se a e b são constantes então

$$\mathbf{E}[aX+b]=a\mathbf{E}[X]+b$$

Para duas variáveis aleatórias X e Y no mesmo espaço amostral, podemos definir novas variáveis aleatórias $X+Y$, $X\cdot Y$, etc. como

$$(X+Y)(\omega)=X(\omega)+Y(\omega),\ \text{e}\ (XY)(\omega)=X(\omega)Y(\omega).$$

Teorema 3.12

$$\mathbf{E}[X+Y] = \mathbf{E}[X] + \mathbf{E}[Y]$$

Teorema 3.13

Se as variáveis aleatórias X e Y são independentes e $\mathbf{E}[X]$ e $\mathbf{E}[Y]$ são finitos, então $\mathbf{E}[XY]$ está bem definida e satisfaz

$$\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y].$$

Definição 3.7

Definimos o n-ésimo momento da variável aleatória X como

$$\mathbf{E}[X^n] := \sum x^n p(x)$$

No caso contínuo temos:

$$\mathrm{Var}(X) = \int_{-\infty}^{+\infty} (x - \mathbf{E}(X))^2 f(x)\,\mathrm{d}x = \mathbf{E}(X^2) - \mathbf{E}(X)^2.$$

3.5 Variância

A esperança localiza um centro, mas não diz quão espalhados estão os valores em torno dele. Duas distribuições podem ter a mesma média e comportamentos muito diferentes. Como medir essa dispersão?

Uma possibilidade seria usar $|X - \mathbf{E}[X]|$. Por razões algébricas, porém, o quadrado $(X - \mathbf{E}[X])^2$ é mais conveniente e conduz à variância. Nesta seção suporemos finitos os momentos necessários.

Definição 3.8

Definimos a variância de uma variável aleatória X, de esperança finita μ como

$$\mathrm{Var}[X] = \mathbf{E}[(x-\mu)^2]$$

O desvio padrão $\sigma(X)$ de X é definido como a raiz quadrada da variância,

$$\sigma(X) = \sqrt{\mathrm{Var}(X)}.$$

Proposição 3.14

$$\mathrm{Var}[X] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2$$

Demonstração. Escreveremos $\mathbf{E}(X) = \mu$. Então temos:

$$\begin{aligned} \mathrm{Var}[X] &= \mathbf{E}(X^2 - 2\mu X + \mu^2) && (3.4)\\ &= \mathbf{E}(X^2) - 2\mu\mathbf{E}(X) + \mu^2 && (3.5)\\ &= \mathbf{E}(X^2) - 2\mu^2 + \mu^2 = \mathbf{E}(X^2) - \mu^2. && (3.6) \end{aligned}$$

■

Exemplo 3.12.

Determine a variância de um dado honesto. denominado

$$\mathbf{E}[X^2] = 1\frac{1}{6} + 4\frac{1}{6} + 9\frac{1}{6} + 16\frac{1}{6} + 25\frac{1}{6} + 36\frac{1}{6} = \frac{91}{6}$$

Como $\mathbf{E}[X] = \dfrac{7}{2}$ temos que

$$\mathrm{Var}(X) = \frac{35}{12}$$

◁

Exemplo 3.13.

Considere o experimento de jogar 5 moedas honestas. Se H representar o número de caras que aparecerem. Calcule a esperança e a variância de H.

Lembramos que

$$\mathbf{P}(X = 0) = \frac{1}{2}^5 = \frac{1}{32}$$

$$\mathbf{P}(X = 1) = \binom{5}{1}\frac{1}{2}^5 = \frac{5}{32}$$

$$\mathbf{P}(X = 2) = \binom{5}{2}\frac{1}{2}^5 = \frac{10}{32}$$

$$\mathbf{P}(X=3)=\binom{5}{3}\frac{1}{2}^5=\frac{10}{32}$$

$$\mathbf{P}(X=4)=\binom{5}{4}\frac{1}{2}^5=\frac{5}{32}$$

$$\mathbf{P}(X=5)=\binom{5}{5}\frac{1}{2}^5=\frac{1}{32}$$

Logo a esperança é

$$\mathbf{E}[H]=1\frac{5}{32}+2\frac{10}{32}+3\frac{10}{32}+4\frac{5}{32}+5\frac{1}{32}=\frac{80}{32}=\frac{5}{2}$$

Para o calculo da variância começaremos calculando

$$\mathbf{E}[H^2]=1\frac{5}{32}+4\frac{10}{32}+9\frac{10}{32}+16\frac{5}{32}+25\frac{1}{32}=\frac{15}{2}$$

Logo a variância é

$$\mathrm{Var}(X)=\frac{15}{2}-\frac{25}{4}=\frac{5}{4}$$

◁

Teorema 3.15

Dadas X e Y variáveis aleatoriamente e $a, b \in \mathbb{R}$ então

a. $\mathrm{Var}(aX+b)=a^2\,\mathrm{Var}(X)$.

b. $\mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)+2(\mathbf{E}(XY)-\mathbf{E}(X)\mathbf{E}(Y))$. Em particular, se X e Y são independentes, então

$$\mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)$$

Ao atravessar este capítulo

A função de distribuição $F_X(x) = \mathbf{P}(X \leq x)$ unifica variáveis discretas, contínuas e mistas. Saltos fornecem massas; em trechos regulares, derivadas fornecem densidades.

3.6 Exercícios

Exercício 3.1 — Validação de uma função de distribuição

Determine os valores de a para os quais

$$F(x) = \begin{cases} 0, & x < 0, \\ ax, & 0 \leq x < 2, \\ 1, & x \geq 2 \end{cases}$$

é uma função de distribuição. Para o valor admissível, calcule $\mathbf{P}(0{,}4 < X \leq 1{,}6)$.

Exercício 3.2 — Saltos e massas

Uma função de distribuição tem saltos de tamanhos 0,18, 0,27 e 0,15 nos pontos -1, 0 e 2, respectivamente, e cresce linearmente de 0,60 para 1 no intervalo [2,5]. Determine todas as massas pontuais e a densidade no trecho contínuo.

Exercício 3.3 — Quantis

Se X tem função de distribuição $F(x) = 1 - e^{-2x}$, $x \geq 0$, encontre a mediana e o quantil de ordem 0,9. Calcule também $\mathbf{P}(X > q_{0,9} + 1 \mid X > q_{0,9})$.

Exercício 3.4 — Transformada inversa

Se $U \sim U(0,1)$, mostre diretamente que $X = 3\sqrt{U}$ possui função de distribuição $F_X(x) = x^2/9$ em [0,3]. Obtenha sua densidade.

Exercício 3.5 — Distribuição mista regular

Uma variável X assume o valor 0 com probabilidade 0,3; com a probabilidade restante, tem densidade $2x$ em $(0,1)$. Escreva F_X, calcule $\mathbf{P}(X \leq 1/2)$ e identifique explicitamente as partes discreta e contínua.

CAPÍTULO

4

Variáveis aleatórias discretas

Uma variável discreta assume valores que podem ser listados, mas isso ainda não diz qual modelo usar. Contar defeitos em vinte peças, contar quantas peças precisamos testar até o primeiro defeito e contar defeitos em vinte metros de cabo são três perguntas de contagem — e levam a distribuições diferentes.

O que decide o modelo é a estrutura do experimento. Há um número fixo de tentativas? As retiradas são com ou sem reposição? Estamos esperando o primeiro sucesso ou contando ocorrências numa região? Ao longo do capítulo, cada nova distribuição aparecerá como resposta a uma dessas perguntas.

4.1 Distribuição Uniforme Discreta

Comecemos pelo caso mais simples. Se uma variável aleatória X assume os valores distintos $x_1, \ldots, x_n$, todos com a mesma probabilidade, dizemos que X tem distribuição uniforme discreta. Sua função de probabilidade é

$$\mathbf{P}(X = x_i) = \frac{1}{n}, \qquad i = 1, \ldots, n.$$

Assim, a probabilidade de um evento é a razão entre a quantidade de valores favoráveis e o total de valores possíveis.

A média e a variância são

$$\mathbf{E}[X] = \frac{x_1 + \cdots + x_n}{n}$$

e

$$\mathrm{Var}(X) = \frac{(x_1 - \mu)^2 + \cdots + (x_n - \mu)^2}{n}, \qquad \mu = \mathbf{E}[X].$$

No caso particular em que X assume os valores $1,2,\ldots,n$, essas expressões se reduzem a

$$\mathbf{E}[X] = \frac{n+1}{2} \quad \text{e} \quad \mathrm{Var}(X) = \frac{n^2 - 1}{12}.$$

Exemplo 4.1 — Uma roleta equilibrada.

Uma roleta tem oito setores de mesma área, numerados de 1 a 8. Seja X o número indicado após um giro. Como os oito resultados são igualmente prováveis, X é uniforme em $\{1,\ldots,8\}$. Portanto,

$$\mathbf{P}(X \geq 6) = \frac{3}{8}.$$

Além disso,

$$\mathbf{E}[X] = \frac{8+1}{2} = 4{,}5 \quad \text{e} \quad \mathrm{Var}(X) = \frac{8^2 - 1}{12} = 5{,}25.$$

A média 4,5 não precisa ser um resultado possível: ela representa o valor em torno do qual a média de muitos giros tende a se estabilizar.

◁

Na distribuição uniforme, a simetria decide os pesos. E se os resultados não forem equiprováveis, mas cada tentativa ainda puder ser classificada apenas como sucesso ou fracasso? Essa mudança leva ao modelo binomial.

4.2 Distribuição Binomial

Considere $n \geq 1$ tentativas independentes, cada uma com dois resultados possíveis: sucesso, com probabilidade $p \in [0,1]$, e fracasso, com probabilidade $1 - p$. Se X conta o número de sucessos nas n tentativas, então

$$\mathbf{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0,1,\ldots,n.$$

Dizemos, nesse caso, que $X \sim \mathrm{Bin}(n,p)$.

A fórmula reúne duas ideias. Uma sequência específica com k sucessos e $n-k$ fracassos tem probabilidade $p^k(1-p)^{n-k}$; existem $\binom{n}{k}$ posições possíveis para os k sucessos. A independência e a constância de p são essenciais: se uma tentativa altera as probabilidades das seguintes, o modelo binomial pode deixar de ser adequado.

Se Y_i vale 1 quando há sucesso na i-ésima tentativa e 0 em caso contrário, então

$$X = Y_1 + \cdots + Y_n.$$

Essa representação permite obter

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

Em particular, np é o número médio de sucessos em muitas repetições do experimento completo.

Exemplo 4.2 — Inspeção de componentes.

Vinte componentes produzidos sucessivamente são inspecionados. Admita que os componentes sejam independentes e que cada um tenha probabilidade 0,07 de ser defeituoso. O lote é aceito quando há no máximo um defeituoso. Se X é o número de defeituosos, então $X \sim \mathrm{Bin}(20{,}0{,}07)$ e

$$\begin{aligned}\mathbf{P}(X \le 1) &= \mathbf{P}(X=0) + \mathbf{P}(X=1)\\ &= 0{,}93^{20} + 20(0{,}07)(0{,}93)^{19}\\ &\approx 0{,}5869.\end{aligned}$$

Assim, sob o modelo adotado, aproximadamente 58,69% dos lotes seriam aceitos.

◁

Exemplo 4.3 — Erros de transmissão.

Uma palavra de 100 bits é transmitida por um canal no qual cada bit, de modo independente, tem probabilidade 0,01 de ser recebido com erro. Um sistema corrige até dois erros. Logo, se $X \sim \mathrm{Bin}(100{,}0{,}01)$, a probabilidade de correção é

$$\mathbf{P}(X \le 2) = \sum_{k=0}^{2} \binom{100}{k} (0{,}01)^k (0{,}99)^{100-k} \approx 0{,}9206.$$

◁

O exemplo anterior calcula o desempenho para uma confiabilidade individual já conhecida. Em projetos de redundância, a pergunta costuma ser inversa: a confiabilidade desejada do sistema determina a exigência sobre cada componente.

Exemplo 4.4 — Confiabilidade de um sistema 4-de-5.

Um sistema possui cinco módulos independentes e permanece operacional desde que pelo menos quatro estejam funcionando. Cada módulo funciona durante a missão com probabilidade p. Se X é o número de módulos operacionais, então

$$X \sim \mathrm{Bin}(5,p)$$

e a confiabilidade do sistema é

$$R(p) = \mathbf{P}(X \geq 4) = \binom{5}{4} p^4(1-p) + p^5 = 5p^4 - 4p^5.$$

Suponha que a especificação do projeto exija confiabilidade de pelo menos 99%. Não estamos mais calculando uma probabilidade a partir de p: precisamos determinar quão confiável deve ser cada módulo. Como $R(p)$ é crescente em $[0,1]$, resolvemos numericamente

$$5p^4 - 4p^5 = 0{,}99,$$

obtendo

$$p \approx 0{,}9673.$$

Assim, cada módulo precisa funcionar com probabilidade de aproximadamente 96,73% para que a arquitetura 4-de-5 atinja a meta de 99%. O exemplo ilustra uma diferença importante entre modelagem e simples cálculo: frequentemente a probabilidade desejada é uma restrição de projeto, e o parâmetro da distribuição é a incógnita.

◁

Na binomial, fixamos o número de tentativas e deixamos o número de sucessos ser aleatório. Podemos inverter a pergunta: e se o número de sucessos de interesse for um — o primeiro — e o que variar for o tempo de espera até ele?

4.3 Distribuição Geométrica

Considere uma sequência de tentativas independentes, todas com probabilidade $0 < p \leq 1$ de sucesso. Seja X o número da tentativa em que ocorre

o primeiro sucesso. Para que $X = k$, as $k-1$ primeiras tentativas devem fracassar e a k-ésima deve ter sucesso. Portanto,

$$\mathbf{P}(X = k) = (1-p)^{k-1}p, \qquad k = 1,2,3,\ldots. \tag{4.1}$$

Dizemos que $X \sim \text{Geom}(p)$. A soma das probabilidades é

$$\sum_{k=1}^{\infty}(1-p)^{k-1}p = p\sum_{j=0}^{\infty}(1-p)^{j} = 1,$$

como deve ocorrer para qualquer função de probabilidade.

Às vezes é mais rápido calcular a cauda. O evento $X > k$ significa que as primeiras k tentativas foram fracassos, de modo que

$$\mathbf{P}(X > k) = (1-p)^{k}.$$

Consequentemente, para inteiros $k \geq 1$,

$$F_X(k) = \mathbf{P}(X \leq k) = 1 - (1-p)^{k}.$$

A média e a variância são

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

Exemplo 4.5 — Teste até encontrar um defeito.
Em uma linha de produção, cada item tem probabilidade 0,20 de ser defeituoso, independentemente dos demais. Os itens são testados até que o primeiro defeito seja encontrado. Se X é o número de itens testados, então $X \sim \text{Geom}(0{,}20)$. A probabilidade de encontrar o primeiro defeito até o terceiro teste é

$$\mathbf{P}(X \leq 3) = 1 - (0{,}80)^3 = 0{,}488.$$

Já a probabilidade de serem necessários mais de cinco testes é

$$\mathbf{P}(X > 5) = (0{,}80)^5 = 0{,}32768.$$

Em média, serão testados $\mathbf{E}[X] = 1/0{,}20 = 5$ itens até o primeiro defeito.

◁

A propriedade sem memória

O fato de as tentativas serem independentes produz uma propriedade especial. Se as primeiras k tentativas fracassaram, a espera adicional até o primeiro sucesso se comporta como uma nova espera, iniciada naquele instante.

Teorema 4.1 — Propriedade sem memória
Se $X \sim \mathrm{Geom}(p)$, com $0 < p < 1$, então, para inteiros $k,n \geq 0$,

$$\mathbf{P}(X > k+n \mid X > k) = \mathbf{P}(X > n).$$

Demonstração. Como $\{X > k+n\} \subseteq \{X > k\}$, temos

$$\begin{aligned}\mathbf{P}(X > k+n \mid X > k) &= \frac{\mathbf{P}(X > k+n)}{\mathbf{P}(X > k)} \\ &= \frac{(1-p)^{k+n}}{(1-p)^k} = (1-p)^n = \mathbf{P}(X > n).\end{aligned}$$

■

Até aqui, cada nova tentativa começava nas mesmas condições. Mas retirar um objeto sem reposição muda o experimento seguinte. A probabilidade de sucesso já não permanece constante. Que distribuição surge quando a população é finita e a amostra altera sua composição? A resposta é a hipergeométrica.

4.4 Distribuição Hipergeométrica

Uma população finita contém $N \geq 1$ elementos, dos quais K são classificados como sucessos e $N-K$ como fracassos, com $0 \leq K \leq N$. Escolhem-se n elementos ao acaso, sem reposição, com $0 \leq n \leq N$, e X conta os sucessos encontrados. Para obter $X = i$, devemos escolher i elementos entre os K sucessos e $n-i$ entre os $N-K$ fracassos. Assim,

$$\mathbf{P}(X = i) = \frac{\binom{K}{i}\binom{N-K}{n-i}}{\binom{N}{n}}.$$

Dizemos que $X \sim \mathrm{HGeom}(N,K,n)$. O suporte efetivo é

$$\max\{0, n-(N-K)\} \leq i \leq \min\{n,K\}.$$

A fórmula também pode ser usada fora desse intervalo se adotarmos $\binom{r}{s} = 0$ quando $s < 0$ ou $s > r$.

Se $N > 1$, a média e a variância são

$$\mathbf{E}[X] = n\frac{K}{N},$$
$$\operatorname{Var}(X) = n\frac{K}{N}\left(1 - \frac{K}{N}\right)\frac{N-n}{N-1}.$$

O último fator, chamado correção para população finita, expressa a redução da variabilidade causada pelas retiradas sem reposição. Quando $N = 1$, X é constante e, portanto, $\operatorname{Var}(X) = 0$.

Exemplo 4.6 — Amostragem sem reposição.

Uma caixa contém 12 componentes, dos quais 4 são defeituosos. Três componentes são escolhidos ao acaso, sem reposição. Se X é o número de defeituosos escolhidos, então $X \sim \mathrm{HGeom}(12,4,3)$. A probabilidade de encontrar exatamente um defeituoso é

$$\mathbf{P}(X = 1) = \frac{\binom{4}{1}\binom{8}{2}}{\binom{12}{3}} = \frac{28}{55} \approx 0{,}5091.$$

A probabilidade de encontrar no máximo um defeituoso é

$$\mathbf{P}(X \le 1) = \frac{\binom{4}{0}\binom{8}{3} + \binom{4}{1}\binom{8}{2}}{\binom{12}{3}} = \frac{42}{55} \approx 0{,}7636.$$

Além disso, $\mathbf{E}[X] = 3(4/12) = 1$. Se cada componente fosse reposto antes da retirada seguinte, a proporção $4/12$ permaneceria constante e o modelo seria binomial, não hipergeométrico.

◁

A hipergeométrica também aparece de uma forma menos evidente: ao condicionar duas contagens binomiais pelo total de sucessos. O cálculo a seguir mostra que o parâmetro p desaparece completamente depois do condicionamento.

Exemplo 4.7 — Uma binomial condicionada torna-se hipergeométrica.

Sejam

$$X \sim \mathrm{Bin}(m,p), \qquad Y \sim \mathrm{Bin}(n,p),$$

independentes. Suponha que conhecemos apenas o número total de sucessos, $X + Y = r$. Qual é a distribuição de X sob essa informação?

Para

$$\max\{0, r-n\} \le k \le \min\{m, r\},$$

temos

$$\mathbf{P}(X = k \mid X + Y = r) = \frac{\mathbf{P}(X = k, Y = r - k)}{\mathbf{P}(X + Y = r)}.$$

O numerador é

$$\binom{m}{k}\binom{n}{r-k}p^r(1-p)^{m+n-r}.$$

Para o denominador, somamos sobre todas as maneiras de distribuir os r sucessos entre os dois grupos:

$$\mathbf{P}(X + Y = r) = p^r(1-p)^{m+n-r}\sum_j \binom{m}{j}\binom{n}{r-j}.$$

A identidade de Vandermonde

$$\sum_j \binom{m}{j}\binom{n}{r-j} = \binom{m+n}{r}$$

segue, por exemplo, contando de duas maneiras a escolha de r elementos em dois grupos de tamanhos m e n. Portanto,

$$\boxed{\mathbf{P}(X = k \mid X + Y = r) = \frac{\binom{m}{k}\binom{n}{r-k}}{\binom{m+n}{r}}}.$$

Essa é exatamente a função de probabilidade de uma distribuição hipergeométrica. O aspecto mais interessante é que p desaparece: uma vez fixado o total de sucessos, importa apenas como os r sucessos se distribuem entre os dois grupos.

◁

Binomial e hipergeométrica começam fixando quantas observações faremos. Há outra pergunta igualmente comum: fixamos um intervalo de tempo, uma área ou um comprimento e perguntamos quantas ocorrências aparecerão ali. Essa troca de ponto de vista conduz à distribuição de Poisson.

4.5 Distribuição de Poisson

Seja X o número de ocorrências de certo tipo em um intervalo de tempo, comprimento, área ou volume. Quando essas ocorrências surgem de maneira

aproximadamente independente e com taxa média constante, um modelo natural é

$$\mathbf{P}(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}, \qquad k = 0{,}1{,}2{,}\ldots,$$

onde $\lambda > 0$. Nesse caso, escrevemos $X \sim \text{Poisson}(\lambda)$.

O parâmetro λ é, ao mesmo tempo, a média e a variância:

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

Se a taxa média é r ocorrências por unidade e observamos uma região de tamanho t, então usamos $\lambda = rt$. Essa conversão de escala deve ser feita antes de calcular qualquer probabilidade.

Exemplo 4.8 — Partículas em uma superfície.

O número de partículas contaminantes em um disco rígido é modelado por uma distribuição de Poisson. A taxa média é 0,1 partícula por centímetro quadrado, e a área examinada tem $100\,\text{cm}^2$. Para essa área, $\lambda = 0{,}1 \times 100 = 10$. Se X é o número de partículas observadas, então

$$\mathbf{P}(X = 12) = e^{-10}\frac{10^{12}}{12!} \approx 0{,}0948.$$

Observe que a unidade de λ deve coincidir com a área efetivamente examinada.

◁

O modelo de Poisson é coerente com três características: contagens em regiões disjuntas são independentes; em uma região muito pequena, a probabilidade de uma ocorrência é aproximadamente proporcional ao tamanho da região; e a probabilidade de duas ou mais ocorrências na mesma região muito pequena é desprezível em comparação com seu tamanho. Essas hipóteses são aproximações de modelagem, que devem ser avaliadas no contexto do problema.

Entre as aplicações usuais estão o número de chamadas recebidas por uma central, de clientes que chegam a uma fila, de falhas em um comprimento de cabo, de erros de impressão por página e de partículas emitidas em um intervalo de tempo.

Duas operações básicas preservam a família de Poisson: separar aleatoriamente as ocorrências e somar fontes independentes. Elas são úteis em filas, redes, telecomunicações e processos de contagem.

Exemplo 4.9 — Afinamento de uma contagem de Poisson.

Durante certo intervalo, o número total de requisições recebidas por um servidor é

$$N \sim \text{Poisson}(\lambda).$$

Cada requisição, independentemente das demais, é encaminhada para uma fila prioritária com probabilidade p. Seja X o número de requisições prioritárias. Condicionalmente a $N = n$,

$$X \mid N = n \sim \text{Bin}(n,p).$$

Assim, para $k \geq 0$,

$$\begin{aligned}
\mathbf{P}(X = k) &= \sum_{n=k}^{\infty} \binom{n}{k} p^k (1-p)^{n-k} e^{-\lambda} \frac{\lambda^n}{n!} \\
&= e^{-\lambda} \frac{(\lambda p)^k}{k!} \sum_{j=0}^{\infty} \frac{[\lambda(1-p)]^j}{j!} \\
&= e^{-\lambda p} \frac{(\lambda p)^k}{k!}.
\end{aligned}$$

Logo,

$$X \sim \text{Poisson}(\lambda p).$$

Há mais estrutura escondida nessa conta. Se $Y = N - X$ é o número de requisições não prioritárias, então

$$\mathbf{P}(X = k, Y = \ell) = e^{-\lambda} \frac{(\lambda p)^k}{k!} \frac{[\lambda(1-p)]^\ell}{\ell!}.$$

Como $e^{-\lambda} = e^{-\lambda p} e^{-\lambda(1-p)}$, essa expressão se fatora nas marginais. Portanto,

$$X \sim \text{Poisson}(\lambda p), \qquad Y \sim \text{Poisson}(\lambda(1-p)),$$

e X e Y são independentes. Selecionar independentemente ocorrências de uma Poisson produz duas Poissons independentes: esse procedimento é chamado de *afinamento*.

◁

Exemplo 4.10 — Superposição de fontes de Poisson.
Duas fontes independentes produzem contagens

$$X \sim \text{Poisson}(\lambda), \qquad Y \sim \text{Poisson}(\mu).$$

Se observamos apenas o total $S = X + Y$, então, para $n \geq 0$,

$$\begin{aligned}
\mathbf{P}(S = n) &= \sum_{k=0}^{n} \mathbf{P}(X = k)\mathbf{P}(Y = n - k) \\
&= e^{-(\lambda+\mu)} \sum_{k=0}^{n} \frac{\lambda^k}{k!} \frac{\mu^{n-k}}{(n-k)!} \\
&= e^{-(\lambda+\mu)} \frac{1}{n!} \sum_{k=0}^{n} \binom{n}{k} \lambda^k \mu^{n-k} \\
&= e^{-(\lambda+\mu)} \frac{(\lambda+\mu)^n}{n!}.
\end{aligned}$$

Portanto,

$$X + Y \sim \text{Poisson}(\lambda + \mu).$$

Em aplicações, isso significa que fluxos independentes de ocorrências podem ser agregados simplesmente somando suas taxas.

◁

Essas propriedades explicam por que a Poisson continua aparecendo depois que um fluxo é dividido em classes ou vários fluxos independentes são observados em conjunto.

Aproximação da binomial pela distribuição de Poisson

A distribuição de Poisson também surge como limite da binomial quando há muitas tentativas, a probabilidade de sucesso em cada uma é pequena e o número médio de sucessos permanece de tamanho moderado.

Teorema 4.2 — Teorema limite de Poisson
Se $p_n \to 0$ e $np_n \to \lambda > 0$, então, para cada inteiro fixo $k \geq 0$,

$$\binom{n}{k} p_n^k (1 - p_n)^{n-k} \longrightarrow e^{-\lambda} \frac{\lambda^k}{k!}.$$

Demonstração. Para $n \geq k$,

$$\binom{n}{k} p_n^k (1-p_n)^{n-k} = \frac{1}{k!} \left[\prod_{j=0}^{k-1} \left(1 - \frac{j}{n}\right) \right] (np_n)^k (1-p_n)^n (1-p_n)^{-k}.$$

Além do fator constante $1/k!$, os quatro fatores convergem, respectivamente, para 1, λ^k, $e^{-\lambda}$ e 1. Para o terceiro fator, basta observar que $n \log(1-p_n) \to -\lambda$. ■

O teorema anterior vale para sequências gerais p_n. No regime canônico $p_n = \lambda/n$, todos os fatores podem ser vistos explicitamente.

Exemplo 4.11 — Eventos raros: da binomial à Poisson.

Considere

$$X_n \sim \text{Bin}\left(n, \frac{\lambda}{n}\right),$$

com $\lambda > 0$ fixo. O número de tentativas cresce, mas a probabilidade de sucesso em cada tentativa diminui de modo que o número médio de sucessos permanece igual a λ. Para k fixo,

$$\mathbf{P}(X_n = k) = \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}.$$

Reescrevendo,

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

Quando $n \to \infty$, o primeiro e o último fatores tendem a 1, enquanto

$$\left(1 - \frac{\lambda}{n}\right)^n \longrightarrow e^{-\lambda}.$$

Assim,

$$\boxed{\mathbf{P}(X_n = k) \longrightarrow e^{-\lambda} \frac{\lambda^k}{k!}}.$$

A distribuição de Poisson surge, portanto, como o limite natural de muitas tentativas independentes com pequena probabilidade individual de sucesso. Essa é a justificativa matemática da expressão "modelo de eventos raros".

◁

O exemplo seguinte usa essa aproximação em uma situação numérica finita.

Exemplo 4.12 — Falhas raras em transmissões.

Em 200 transmissões independentes, cada uma falha com probabilidade 0,015. Se X é o número de falhas, então $X \sim \text{Bin}(200{,}0{,}015)$ e $np = 3$. A aproximação de Poisson fornece

$$\mathbf{P}(X \leq 2) \approx e^{-3}\left(1 + 3 + \frac{3^2}{2}\right) \approx 0{,}4232.$$

O cálculo binomial exato continua sendo a referência quando estiver disponível; a aproximação é útil sobretudo quando ele se torna trabalhoso.

◁

Exemplo 4.13 — Terremotos de grande intensidade.

Entre 2000 e 2009 ocorreram 144 terremotos de magnitude pelo menos 7,0 no mundo. Admitindo uma taxa constante e usando a distribuição de Poisson como modelo, a taxa anual estimada é $\lambda = 144/10 = 14{,}4$. Logo, a probabilidade de ocorrerem exatamente três terremotos dessa intensidade no próximo ano é

$$\mathbf{P}(X = 3) = e^{-14{,}4}\frac{14{,}4^3}{3!} \approx 0{,}000277.$$

O valor muito pequeno é compatível com o fato de 3 estar bastante abaixo da média anual 14,4.

◁

Os exemplos acima mostram que as distribuições discretas não são modelos isolados: condicionamento, superposição, afinamento e passagem ao limite criam pontes entre elas. Quando o resultado pode assumir qualquer valor em um intervalo, passamos às variáveis contínuas. Antes disso, encerramos o capítulo vendo como a lei de uma variável discreta se transforma quando definimos uma nova quantidade a partir dela.

4.6 Funções de variáveis aleatórias discretas

Se X é discreta, $Y = g(X)$ também é: seus valores possíveis são as imagens dos valores de X. Como valores diferentes de X podem produzir o mesmo valor de Y, suas probabilidades devem ser agrupadas. Se $p_X(x_i) = \mathbf{P}(X = x_i)$, então

$$p_Y(y_j) = \sum_{x_i : g(x_i) = y_j} p_X(x_i).$$

Exemplo 4.14.

Considere a função de probabilidade conjunta de X e Y apresentada na Tabela 4.1.

$y\backslash x$	-1	0	1
0	1/16	1/16	1/16
1	1/16	1/16	2/16
2	2/16	1/16	6/16

Tabela 4.1: Função de probabilidade conjunta de X e Y

Seja $Z = XY$. Calcularemos:

1. a função de probabilidade de Z;
2. a esperança de Z.

Solução. A variável aleatória Z pode assumir cinco valores distintos: $-2, -1, 0, 1$ e 2. Cada ponto da tabela corresponde a um valor de Z: o ponto $(-1, 2)$ corresponde a $z = -2$, $(-1, 1)$ a $z = -1$, $(1, 1)$ a $z = 1$, $(1, 2)$ a $z = 2$, e todos os outros pontos produzem $z = 0$. A partir da tabela, obtemos:

z	-2	-1	0	1	2
$p_Z(z)$	$\frac{2}{16}$	$\frac{1}{16}$	$\frac{5}{16}$	$\frac{2}{16}$	$\frac{6}{16}$

Portanto, a esperança de Z é dada por

$$\mathbf{E}[Z] = (-2) \cdot \frac{2}{16} + (-1) \cdot \frac{1}{16} + 0 + (1) \cdot \frac{2}{16} + (2) \cdot \frac{6}{16} = \frac{9}{16}.$$

◁

Exemplo 4.15.

Sejam N e M duas variáveis aleatórias independentes, ambas com distribuição de Poisson com parâmetro $\lambda = 1$. Desejamos calcular a função de probabilidade de Z, onde

$$Z = g(N, M) := \begin{cases} 1 & \text{se } N = M, \\ 0 & \text{se } N \neq M \end{cases}.$$

A função de probabilidade conjunta do vetor aleatório (N, M) é dada por

$$p_{N,M}(n, m) = \frac{e^{-2}}{n!\, m!} \quad \text{para } n, m = 0, 1, \ldots$$

Neste caso, Z tem distribuição de Bernoulli com parâmetro p, onde

$$p := \mathbf{P}(N = M) = \sum_{n=0}^{\infty} \frac{e^{-2}}{(n!)^2}.$$

◁

Ao atravessar este capítulo

Modelos discretos são escolhidos pela estrutura do experimento: número fixo de ensaios, espera pelo primeiro sucesso, amostragem sem reposição ou contagem de ocorrências raras.

4.7 Exercícios

Exercício 4.1 — Binomial

Em 18 componentes, cada um apresenta defeito com probabilidade 0,08, independentemente. Calcule a probabilidade de haver: (a) nenhum defeito; (b) exatamente dois; (c) ao menos um. Determine média e variância da contagem.

Exercício 4.2 — Geométrica

Uma tentativa tem probabilidade de sucesso 0,35, e as tentativas são independentes. Seja X o número da tentativa do primeiro sucesso. Calcule $\mathbf{P}(X > 5)$, $\mathbf{P}(X = 5)$ e $\mathbf{P}(X > 9 \mid X > 4)$.

Exercício 4.3 — Hipergeométrica

Uma caixa contém 9 peças conformes e 5 defeituosas. Retiram-se 4 peças sem reposição. Encontre a função de probabilidade do número de defeituosas retiradas e calcule a probabilidade de aparecer no máximo uma.

Exercício 4.4 — Poisson

Chamadas chegam a uma central à taxa média de 3,6 por minuto. Sob o modelo de Poisson, calcule a probabilidade de: (a) nenhuma chamada em 30 segundos; (b) pelo menos quatro em um minuto; (c) exatamente sete em dois minutos.

Exercício 4.5 — Escolha do modelo

Para cada situação, indique uma distribuição adequada e seus parâmetros: (a) número de ases em sete cartas retiradas de um baralho; (b) número de clientes que chegam em dez minutos; (c) número de lançamentos até a primeira coroa; (d) número de respostas corretas em doze questões independentes com probabilidade comum de acerto.

CAPÍTULO

5

Variáveis aleatórias contínuas

No capítulo anterior, probabilidades eram somas de massas colocadas em pontos. Agora o resultado é uma medida: tempo, comprimento, massa, temperatura. Se um valor isolado tem probabilidade zero, onde fica a probabilidade? Ela se espalha por intervalos e é descrita por uma densidade:

$$\mathbf{P}(a \leq X \leq b) = \int_a^b f(x)\,dx.$$

Por isso, para variáveis contínuas, incluir ou não os extremos não altera a probabilidade.

Aqui convém separar duas ideias: $f_X(x)$ é uma altura, não a probabilidade do ponto x. Probabilidades são áreas. Assim, uma densidade pode ser maior que 1: se $X \sim \text{Unif}(0,1/2)$, então $f_X = 2$ no intervalo, mas

$$\mathbf{P}(0 \leq X \leq 1/4) = 2 \cdot \frac{1}{4} = \frac{1}{2}.$$

Também aqui a escolha da distribuição vem da pergunta. Todos os pontos de um intervalo devem ser tratados simetricamente? Estamos esperando uma ocorrência com taxa constante? Observamos uma soma de muitos pequenos efeitos? Ou o tempo até várias ocorrências sucessivas? Uniforme, exponencial, normal e gama respondem, respectivamente, a essas situações.

5.1 Distribuição Uniforme Contínua

Se todos os subintervalos de mesmo comprimento devem ter a mesma probabilidade, a densidade precisa ser constante. Dizemos que X tem distribuição uniforme no intervalo $[a,b]$, com $a < b$, quando

$$f_X(x) = \begin{cases} \dfrac{1}{b-a}, & a \le x \le b, \\ 0, & \text{caso contrário.} \end{cases}$$

Escrevemos $X \sim \mathrm{Unif}(a,b)$. Se $a \le y < z \le b$, então

$$\mathbf{P}(y \le X \le z) = \int_y^z \frac{1}{b-a}\,\mathrm{d}x = \frac{z-y}{b-a}.$$

Assim, a probabilidade é a razão entre o comprimento do intervalo favorável e o comprimento total.

A função de distribuição é

$$F_X(x) = \begin{cases} 0, & x < a, \\ \dfrac{x-a}{b-a}, & a \le x \le b, \\ 1, & x > b, \end{cases}$$

e a média e a variância são

$$\mathbf{E}[X] = \frac{a+b}{2} \quad \text{e} \quad \mathrm{Var}(X) = \frac{(b-a)^2}{12}.$$

Exemplo 5.1 — Tempo de espera por um ônibus.
Um ônibus passa a cada 14 minutos, e uma pessoa chega ao ponto em um instante que pode ser considerado uniforme entre duas passagens. Se X é o tempo de espera, em minutos, então $X \sim \mathrm{Unif}(0{,}14)$. A probabilidade de esperar entre 3 e 8 minutos é

$$\mathbf{P}(3 \le X \le 8) = \frac{8-3}{14} = \frac{5}{14} \approx 0{,}3571.$$

Já a probabilidade de esperar mais de 10 minutos é

$$\mathbf{P}(X > 10) = \frac{14-10}{14} = \frac{2}{7} \approx 0{,}2857.$$

O tempo médio de espera é $\mathbf{E}[X] = 7$ minutos.

◁

A uniforme supõe um horizonte limitado. Mas muitos tempos de espera não vêm com um limite máximo natural. Se, além disso, a taxa de ocorrência é constante, a distribuição exponencial aparece como o modelo mais simples.

5.2 Distribuição Exponencial

Suponha que ocorrências se deem ao longo do tempo com taxa média constante $\lambda > 0$, como chegadas a uma central ou falhas de certo tipo. O tempo X até a próxima ocorrência é frequentemente modelado pela distribuição exponencial, cuja densidade é

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

Nesse caso, escrevemos $X \sim \mathrm{Exp}(\lambda)$. Para $t \geq 0$,

$$F_X(t) = \int_0^t \lambda e^{-\lambda x}\,\mathrm{d}x = 1 - e^{-\lambda t},$$

e, portanto,

$$\mathbf{P}(X > t) = e^{-\lambda t}.$$

A média e a variância são

$$\mathbf{E}[X] = \frac{1}{\lambda} \quad \text{e} \quad \mathrm{Var}(X) = \frac{1}{\lambda^2}.$$

Logo, uma taxa maior corresponde a uma espera média menor.

Exemplo 5.2 — Espera até a próxima chamada.

As chamadas chegam a uma central a uma taxa média de 0,20 por minuto. Admitindo um modelo exponencial para o tempo X até a próxima chamada, $X \sim \mathrm{Exp}(0{,}20)$. A probabilidade de a espera ultrapassar 8 minutos é

$$\mathbf{P}(X > 8) = e^{-0{,}20 \cdot 8} = e^{-1{,}6} \approx 0{,}2019.$$

Por outro lado,

$$\mathbf{P}(X \leq 3) = 1 - e^{-0{,}20 \cdot 3} = 1 - e^{-0{,}6} \approx 0{,}4512.$$

O tempo médio de espera é $1/0{,}20 = 5$ minutos.

◁

A distribuição exponencial aparece também em confiabilidade, como modelo para a duração de componentes com taxa de falha constante, e em sistemas de filas, como modelo para tempos entre chegadas. A hipótese de taxa constante deve ser avaliada com cuidado: componentes sujeitos a envelhecimento, por exemplo, podem não se ajustar bem a esse modelo.

Proposição 5.1 — Propriedade sem memória
Se $X \sim \text{Exp}(\lambda)$, então, para $s,t \geq 0$,

$$\mathbf{P}(X > s+t \mid X > s) = \mathbf{P}(X > t).$$

Demonstração. Usando a expressão da cauda,

$$\mathbf{P}(X > s+t \mid X > s) = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = \mathbf{P}(X > t).$$

■

No exemplo das chamadas, saber que já se passaram 8 minutos sem uma nova chamada não altera, sob o modelo exponencial, a distribuição da espera adicional. Essa propriedade é a versão contínua da propriedade sem memória da distribuição geométrica.

A propriedade sem memória não esgota a estrutura da exponencial. Mínimos de tempos exponenciais modelam riscos concorrentes; quantis convertem metas de confiabilidade em prazos; e a própria distribuição pode ser simulada a partir de uma uniforme.

Exemplo 5.3 — Riscos competitivos.
Uma máquina pode parar por dois mecanismos independentes. O tempo até uma falha do tipo 1 é

$$T_1 \sim \text{Exp}(0{,}02),$$

e o tempo até uma falha do tipo 2 é

$$T_2 \sim \text{Exp}(0{,}03),$$

com taxas medidas por dia. O tempo efetivo até a parada é

$$T = \min(T_1, T_2).$$

Para $t \geq 0$,

$$\begin{aligned}\mathbf{P}(T > t) &= \mathbf{P}(T_1 > t, T_2 > t) \\ &= e^{-0{,}02t} e^{-0{,}03t} = e^{-0{,}05t}.\end{aligned}$$

Portanto,

$$T \sim \text{Exp}(0{,}05), \qquad \mathbf{E}[T] = 20 \text{ dias}.$$

Podemos também perguntar qual mecanismo provavelmente causará a primeira falha. Temos

$$\begin{aligned}\mathbf{P}(T_1 < T_2) &= \int_0^\infty f_{T_1}(t)\mathbf{P}(T_2 > t)\,dt \\ &= \int_0^\infty 0{,}02e^{-0{,}02t}e^{-0{,}03t}\,dt = \frac{0{,}02}{0{,}05} = 0{,}4.\end{aligned}$$

Assim, embora o sistema pare em média após vinte dias, apenas 40% das paradas são atribuídas ao primeiro mecanismo. Em geral, para riscos exponenciais independentes com taxas $\lambda_1, \ldots, \lambda_m$, o tempo até a primeira ocorrência é exponencial com taxa $\lambda_1 + \cdots + \lambda_m$, e a probabilidade de o mecanismo i ser o primeiro é

$$\frac{\lambda_i}{\lambda_1 + \cdots + \lambda_m}.$$

◁

Exemplo 5.4 — Garantia definida por um quantil.

O tempo de vida X, em horas, de um componente é modelado por uma exponencial com média 20 000 horas. Em vez de perguntar a probabilidade de falha antes de um prazo dado, o fabricante deseja escolher um período de garantia w para que apenas 5% dos componentes falhem antes dele.

Como $\lambda = 1/20\,000$, devemos resolver

$$\mathbf{P}(X \le w) = 0{,}05.$$

Logo,

$$1 - e^{-w/20\,000} = 0{,}05,$$

e portanto

$$\boxed{w = -20\,000 \log(0{,}95) \approx 1\,026 \text{ horas.}}$$

O problema é o inverso dos exemplos usuais: a probabilidade desejada é dada e o valor da variável precisa ser determinado. Essa é a interpretação dos *quantis* de uma distribuição.

◁

Exemplo 5.5 — Da uniforme à exponencial: simulação por transformação inversa.

Geradores de números pseudoaleatórios produzem, em sua forma mais básica, valores aproximadamente uniformes em (0,1). Como transformar esses valores em tempos exponenciais?

Seja $U \sim \mathrm{Unif}(0,1)$ e defina

$$X = -\frac{1}{\lambda} \log U.$$

Para $x \geq 0$,

$$\begin{aligned} \mathbf{P}(X \leq x) &= \mathbf{P}\left(-\frac{1}{\lambda} \log U \leq x\right) \\ &= \mathbf{P}(U \geq e^{-\lambda x}) \\ &= 1 - e^{-\lambda x}. \end{aligned}$$

Logo,

$$X \sim \mathrm{Exp}(\lambda).$$

Essa identidade não serve apenas para reconhecer uma distribuição: ela fornece um algoritmo. Gerando U uniforme e calculando $-\log U/\lambda$, obtemos uma observação exponencial. O mesmo princípio, aplicado à inversa da função de distribuição, é uma das técnicas fundamentais de simulação probabilística.

◁

Esses exemplos mostram três leituras diferentes da mesma cauda exponencial: competição entre mecanismos, escolha de um nível de garantia e geração computacional de tempos aleatórios.

5.3 Distribuição Normal

A exponencial é adequada para esperas positivas e assimétricas. Mas e quando os desvios podem ocorrer para os dois lados de um valor central? Erros de medição e grandezas formadas pela soma de muitos pequenos efeitos sugerem uma curva simétrica: a distribuição normal.

Dizemos que X tem distribuição normal com média μ e variância $\sigma^2 > 0$, e escrevemos $X \sim \mathcal{N}(\mu,\sigma^2)$, quando

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \qquad x \in \mathbb{R}.$$

O parâmetro μ determina o centro da curva, e σ determina sua dispersão. Em particular,

$$\mathbf{E}[X] = \mu \quad \text{e} \quad \operatorname{Var}(X) = \sigma^2.$$

Padronização

A integral da densidade normal não possui antiderivada elementar. Todos os cálculos podem, porém, ser reduzidos à normal padrão. Se $X \sim \mathcal{N}(\mu,\sigma^2)$, então

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0,1).$$

Denotamos a função de distribuição de Z por

$$\Phi(z) = \mathbf{P}(Z \le z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\,\mathrm{d}t.$$

Consequentemente,

$$\mathbf{P}(X \le x) = \Phi\left(\frac{x - \mu}{\sigma}\right).$$

Para um intervalo, padronizamos os dois extremos. A simetria da curva fornece $\Phi(-z) = 1 - \Phi(z)$.

Exemplo 5.6 — Diâmetro de peças usinadas.

O diâmetro X, em milímetros, de uma peça é modelado por $\mathcal{N}(20{,}0, 12^2)$. Uma peça atende à especificação se seu diâmetro está entre 19,8 e 20,2 mm. Então

$$\begin{aligned}\mathbf{P}(19{,}8 \le X \le 20{,}2) &= \mathbf{P}\left(-1{,}6667 \le Z \le 1{,}6667\right) \\ &= 2\Phi(1{,}6667) - 1 \\ &\approx 0{,}9044.\end{aligned}$$

Assim, aproximadamente 90,44% das peças atendem à especificação, de acordo com o modelo.

◁

A padronização também aparece quando a probabilidade desejada é uma restrição de projeto. No exemplo seguinte, o nível de falso alarme determina o limiar do sensor e, em seguida, esse mesmo limiar determina a probabilidade de detecção.

Exemplo 5.7 — Limiar de detecção com ruído gaussiano.
Um sensor produz a medida

$$Y = \theta + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0{,}2^2).$$

Na ausência de sinal, $\theta = 0$. Um alarme é disparado quando $Y > c$. O limiar deve ser escolhido de modo que a probabilidade de falso alarme seja apenas 1%:

$$\mathbf{P}(Y > c \mid \theta = 0) = 0{,}01.$$

Se $z_{0{,}99} \approx 2{,}326$ é o percentil 99% da normal padrão, então

$$\frac{c}{2} = z_{0{,}99}, \qquad c \approx 4{,}653.$$

Agora suponha que um sinal de intensidade $\theta = 6$ esteja presente. A probabilidade de detecção é

$$\begin{aligned}\mathbf{P}(Y > c \mid \theta = 6) &= \mathbf{P}\left(Z > \frac{c-6}{2}\right) \\ &\approx \mathbf{P}(Z > -0{,}674) = \Phi(0{,}674) \approx 0{,}75.\end{aligned}$$

O mesmo limiar que controla falsos alarmes determina, portanto, a capacidade de detectar um sinal real. Esse é o embrião probabilístico dos problemas de detecção e testes de hipóteses.

◁

Aproximação da binomial pela normal

Uma variável binomial é uma soma de variáveis que valem 0 ou 1. Quando o número de parcelas é grande e nenhuma das duas respostas é excessivamente rara, a forma de sua distribuição se aproxima da curva normal.

Teorema 5.2 — Aproximação da binomial pela normal
Fixe $0 < p < 1$ e, para cada n, seja $X_n \sim \mathrm{Bin}(n,p)$. Então

$$\lim_{n\to\infty} \mathbf{P}\left(\frac{X_n - np}{\sqrt{np(1-p)}} \le z\right) = \Phi(z), \qquad z \in \mathbb{R}.$$

Assim, para n grande, X_n pode ser aproximada por uma variável normal com média np e variância $np(1-p)$.

Esse resultado será retomado como consequência do Teorema Central do Limite no Capítulo 9.

Como regra prática, a aproximação costuma ser adequada quando $np \geq 5$ e $n(1-p) \geq 5$; quanto maiores forem essas duas quantidades, melhor tende a ser o resultado. Essa regra é apenas um diagnóstico inicial, não uma garantia de erro máximo.

Há ainda uma diferença de natureza entre os modelos: a binomial assume valores inteiros, enquanto a normal é contínua. Para compensá-la, usa-se a correção de continuidade. Se $Y \sim \mathcal{N}(np, np(1-p))$, aproximamos, por exemplo,

$$\begin{aligned}
\mathbf{P}(X \leq a) &\approx \mathbf{P}(Y \leq a + 0{,}5),\\
\mathbf{P}(X \geq a) &\approx \mathbf{P}(Y \geq a - 0{,}5),\\
\mathbf{P}(X = a) &\approx \mathbf{P}(a - 0{,}5 \leq Y \leq a + 0{,}5).
\end{aligned}$$

Exemplo 5.8 — Controle de qualidade.

Em um processo de produção, 1000 itens são inspecionados e cada item tem, independentemente, probabilidade 0,1 de ser defeituoso. Se X é o número de defeituosos, então $X \sim \mathrm{Bin}(1000, 0{,}1)$, com

$$\mu = np = 100 \quad \text{e} \quad \sigma^2 = np(1-p) = 90.$$

Como $np = 100$ e $n(1-p) = 900$, a aproximação normal é apropriada. Aplicando a correção de continuidade,

$$\begin{aligned}
\mathbf{P}(90 \leq X \leq 110) &\approx \mathbf{P}(89{,}5 \leq Y \leq 110{,}5)\\
&= \mathbf{P}(-1{,}1068 \leq Z \leq 1{,}1068)\\
&\approx 0{,}7316,
\end{aligned}$$

onde $Y \sim \mathcal{N}(100, 90)$. Portanto, a probabilidade procurada é aproximadamente 73,16%.

◁

Exponencial e normal têm formas bastante rígidas. E se a variável for positiva, mas uma espera exponencial simples não for suficiente — por exemplo, porque queremos o tempo até a terceira ou a quarta ocorrência? A família gama amplia o modelo exponencial e permite diferentes graus de assimetria.

5.4 Distribuição Gama

A distribuição gama é usada para modelar grandezas positivas. Quando o parâmetro de forma é maior que 1, sua densidade começa em zero, cresce

até um máximo e depois decresce, o que é adequado para situações em que valores muito pequenos e muito grandes são pouco frequentes. Ela também descreve o tempo acumulado até várias ocorrências sucessivas de um processo com taxa constante.

Dizemos que X tem distribuição gama com parâmetro de forma $\alpha > 0$ e parâmetro de taxa $\beta > 0$ quando

$$f_X(x) = \begin{cases} \dfrac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, & x > 0, \\ 0, & x \le 0, \end{cases}$$

onde

$$\Gamma(\alpha) = \int_0^\infty u^{\alpha-1} e^{-u}\, \mathrm{d}u.$$

Escrevemos $X \sim \text{Gama}(\alpha,\beta)$. Adotaremos em todo o livro a parametrização forma–taxa; nessa convenção,

$$\mathbf{E}[X] = \frac{\alpha}{\beta} \quad \text{e} \quad \text{Var}(X) = \frac{\alpha}{\beta^2}.$$

Quando $\alpha = 1$, a densidade gama se reduz à densidade exponencial de taxa β.

Se $\alpha = m$ é um inteiro positivo, a distribuição pode ser interpretada como o tempo até a m-ésima ocorrência de um processo de Poisson de taxa β. Nesse caso, para $t \ge 0$, a probabilidade de ainda não terem ocorrido m eventos até o instante t é

$$\mathbf{P}(X > t) = e^{-\beta t} \sum_{j=0}^{m-1} \frac{(\beta t)^j}{j!}.$$

A relação com a Poisson merece uma derivação completa, pois ela explica de onde vem a forma da densidade gama quando o parâmetro de forma é inteiro.

Exemplo 5.9 — Da contagem de Poisson ao tempo da k-ésima ocorrência. Suponha que $N(t)$ conte o número de ocorrências até o instante t e que

$$N(t) \sim \text{Poisson}(\lambda t), \qquad t \ge 0.$$

Seja T_k o instante da k-ésima ocorrência. Os eventos

$$\{T_k \le t\} \quad \text{e} \quad \{N(t) \ge k\}$$

são exatamente os mesmos: a k-ésima ocorrência já aconteceu até t se, e somente se, pelo menos k ocorrências foram contadas nesse intervalo. Portanto,

$$\begin{aligned} F_{T_k}(t) &= \mathbf{P}(T_k \leq t) \\ &= \mathbf{P}(N(t) \geq k) \\ &= 1 - \sum_{j=0}^{k-1} e^{-\lambda t} \frac{(\lambda t)^j}{j!}, \qquad t \geq 0. \end{aligned}$$

Derivando,

$$f_{T_k}(t) = \frac{\lambda^k}{(k-1)!} t^{k-1} e^{-\lambda t}, \qquad t > 0.$$

Logo,

$$\boxed{T_k \sim \text{Gama}(k,\lambda)}.$$

Para $k = 1$, recuperamos a distribuição exponencial. Assim, exponencial, gama e Poisson não são três modelos independentes: a Poisson descreve o número de ocorrências, a exponencial o tempo até a primeira e a gama o tempo até a k-ésima.

◁

O caso seguinte especializa essa identidade para $k = 3$ e a usa numericamente.

Exemplo 5.10 — Tempo até a terceira ocorrência.

Eventos chegam a um sistema à taxa média de 2 por hora. Seja X o tempo, em horas, até a terceira ocorrência. Sob o modelo de Poisson, $X \sim \text{Gama}(3,2)$. Logo,

$$\mathbf{E}[X] = \frac{3}{2} = 1{,}5 \quad \text{e} \quad \text{Var}(X) = \frac{3}{2^2} = 0{,}75.$$

A probabilidade de a terceira ocorrência demorar mais de duas horas é

$$\begin{aligned} \mathbf{P}(X > 2) &= e^{-4}\left(1 + 4 + \frac{4^2}{2!}\right) \\ &= 13e^{-4} \approx 0{,}2381. \end{aligned}$$

Equivalentemente, até duas horas ocorreram no máximo dois eventos.

◁

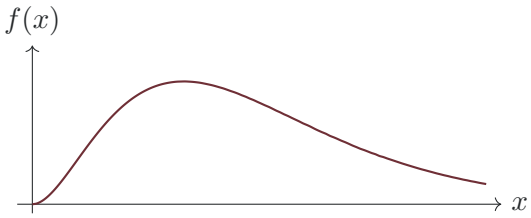


Figura 5.1: Densidade gama com parâmetros $\alpha = 3$ e $\beta = 1$.

Uniforme, exponencial, normal e gama não formam uma lista de fórmulas intercambiáveis: cada uma traduz uma estrutura diferente do fenômeno. Os exemplos técnicos mostraram também que quantis, processos de contagem, mecanismos concorrentes e transformações podem revelar a distribuição antes mesmo de qualquer integração. Encerramos o capítulo vendo como a distribuição muda quando aplicamos uma função a uma variável contínua.

Ao atravessar este capítulo

Em modelos contínuos, probabilidades são áreas sob densidades. Uniforme, exponencial, normal e gama são reconhecidas por suporte, forma, parâmetros e propriedades estruturais.

5.5 Exercícios

Exercício 5.1 — Normalização

Considere $f(x) = c(1 - x^2)$, para $-1 < x < 1$, e $f(x) = 0$ fora desse intervalo. Determine c, F_X e $\mathbf{P}(|X| < 1/2)$.

Exercício 5.2 — Uniforme

O atraso de um ônibus é uniforme entre 0 e 14 minutos. Calcule a probabilidade de o atraso: (a) superar 9 minutos; (b) ficar entre 3 e 8 minutos; (c) superar 11 minutos dado que já superou 6.

Exercício 5.3 — Exponencial

O tempo de vida de um componente é exponencial com média 450 horas. Determine a probabilidade de ele durar mais de 600 horas e a probabilidade de durar mais 200 horas sabendo que já funcionou por 500.

Exercício 5.4 — Normal

O diâmetro de uma peça é normal com média 20,0 mm e desvio padrão 0,12 mm. Peças entre 19,8 e 20,2 mm são aceitas. Calcule a proporção esperada de peças aceitas e o percentil 95 do diâmetro.

Exercício 5.5 — Gama

Se $X \sim \text{Gama}(3{,}2)$, na parametrização forma–taxa, determine $\mathbf{E}[X]$, $\text{Var}(X)$ e escreva a integral que fornece $\mathbf{P}(X > 2)$. Para o parâmetro de forma inteiro, simplifique essa probabilidade.

Exercício 5.6 — Transformação monotônica

Se $X \sim U(-2{,}3)$ e $Y = (X+2)^2$, determine F_Y, f_Y e o suporte de Y.

Exercício 5.7 — Convolução

Se X e Y são exponenciais independentes com taxas 2 e 5, respectivamente, encontre a densidade de $Z = X + Y$. Verifique que ela integra para 1.

CAPÍTULO

6

Vetores aleatórios

6.1 Vetores aleatórios e funções de distribuição

Até aqui, uma variável aleatória foi suficiente para registrar a quantidade de interesse: o tempo de vida de uma bateria, o número de chamadas recebidas ou a distância percorrida. Muitos experimentos, porém, produzem várias quantidades simultaneamente. Se escolhemos ao acaso um ponto no disco unitário, por exemplo, suas duas coordenadas precisam ser estudadas em conjunto. Se ω é o resultado do experimento, o ponto observado é

$$(X(\omega),Y(\omega)) \in \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}.$$

O que se perde se estudarmos apenas X e apenas Y? Justamente a maneira como variam juntas. Duas coordenadas podem ter marginais simples e, ainda assim, apresentar uma dependência forte. É essa informação conjunta que precisamos preservar ao relacionar tempo de espera e duração do atendimento, volume de vendas e lucro ou duas medições do mesmo indivíduo.

Um vetor aleatório é uma coleção ordenada de variáveis aleatórias observadas no mesmo experimento. Neste capítulo, começaremos pela distribuição conjunta, passaremos aos casos discreto e contínuo e, em seguida, estudaremos independência, somas e distribuições condicionais. Em cada etapa, a ideia central será a mesma: calcular probabilidades em regiões do espaço dos valores possíveis.

Definição 6.1

1. Um vetor $X = (X_1, \ldots, X_n)$, cujas componentes são variáveis aleatórias definidas no mesmo espaço de probabilidade, é denominado **vetor aleatório**.
2. A **função de distribuição conjunta** de $X = (X_1, \ldots, X_n)$ é

$$F_X(x_1, \ldots, x_n) = \mathbf{P}(X_1 \leq x_1, \ldots, X_n \leq x_n), \qquad (x_1, \ldots, x_n) \in \mathbb{R}^n.$$

Exemplo 6.1.

Em uma central de reservas, seja X_1 o tempo de espera, em minutos, até o início do atendimento e seja X_2 a duração do atendimento. Suponha que a função de distribuição conjunta seja

$$F_{X_1,X_2}(x_1,x_2) = \begin{cases} 0, & x_1 < 0 \text{ ou } x_2 < 0, \\ 1 - e^{-x_1} - e^{-2x_2} + e^{-(x_1+2x_2)}, & x_1 \geq 0,\ x_2 \geq 0. \end{cases}$$

Por exemplo, a probabilidade de uma chamada esperar no máximo um minuto e durar no máximo dois minutos é $F_{X_1,X_2}(1,2)$. Assim, uma única função permite avaliar simultaneamente as duas etapas do serviço.

◁

Proposição 6.1

A função de distribuição F de um vetor aleatório (X,Y) possui as seguintes propriedades:

F1. $F(x,y)$ é não-decrescente em cada uma das variáveis. Por exemplo, é não-decrescente em x : se $x_1 \leq x_2$, então

$$F(x_1, y) \leq F(x_2, y).$$

F2. $F(x,y)$ é contínua à direita em cada uma das variáveis. Ou seja, para quaisquer x, y:

$$\lim_{u \to x^+} F(u, y) = F(x, y)$$

e

$$\lim_{v \to y^+} F(x,v) = F(x,y)$$

F3. Para cada y fixado, $\lim_{x\to-\infty} F(x,y) = 0$, e, para cada x fixado, $\lim_{y\to-\infty} F(x,y) = 0$. Além disso,

$$\lim_{x\to\infty}\lim_{y\to\infty} F(x,y) = 1.$$

F4. Todo retângulo semiaberto tem incremento não negativo: se $a_1 < a_2$ e $b_1 < b_2$, então

$$F(a_2,b_2) - F(a_1,b_2) - F(a_2,b_1) + F(a_1,b_1) \geq 0.$$

Demonstração. As propriedades **F1** e **F2** seguem, como no caso unidimensional, da inclusão de eventos e da continuidade da probabilidade. Para **F3**, se $x_m \downarrow -\infty$, então $[X \leq x_m, Y \leq y] \downarrow \varnothing$; portanto, $F(x_m,y) \to 0$. O mesmo argumento vale na segunda coordenada. Se $x_m \uparrow \infty$ e $y_m \uparrow \infty$, então $[X \leq x_m, Y \leq y_m] \uparrow \Omega$, e daí $F(x_m,y_m) \to 1$.

Por fim, o incremento em **F4** é exatamente

$$\mathbf{P}(a_1 < X \leq a_2,\ b_1 < Y \leq b_2),$$

e, portanto, é não negativo. ∎

Para recuperar a distribuição de uma única componente, fazemos as demais coordenadas tenderem a $+\infty$. No caso bidimensional,

$$F_X(a) = \lim_{b\to\infty} F(a,b) := F(a,\infty), \qquad F_Y(b) = \lim_{a\to\infty} F(a,b) := F(\infty,b).$$

Essas são as **distribuições marginais**. Elas descrevem cada coordenada separadamente, mas não determinam, em geral, como X e Y se relacionam.

A distribuição conjunta, por sua vez, determina probabilidades de regiões. Por exemplo,

$$\begin{aligned}
\mathbf{P}(X > a, Y > b) &= 1 - \mathbf{P}\left((X > a, Y > b)^c\right) \\
&= 1 - \mathbf{P}\left((X > a)^c \cup (Y > b)^c\right) \\
&= 1 - \mathbf{P}((X \leq a) \cup (Y \leq b)) \\
&= 1 - [\mathbf{P}(X \leq a) + \mathbf{P}(Y \leq b) - \mathbf{P}(X \leq a, Y \leq b)] \\
&= 1 - F_X(a) - F_Y(b) + F(a,b)
\end{aligned}$$

A equação anterior é um caso especial da equação a seguir:

$$\mathbf{P}(a_1 < X \leq a_2, b_1 < Y \leq b_2) = F(a_2, b_2) + F(a_1, b_1) - F(a_1, b_2) - F(a_2, b_1) \tag{6.1}$$

sempre que $a_1 < a_2, b_1 < b_2$. Geometricamente, a fórmula soma o retângulo até (a_2,b_2), retira as duas faixas excedentes e devolve a região retirada duas vezes. Esse princípio de inclusão e exclusão reaparecerá tanto nas tabelas do caso discreto quanto nas integrais do caso contínuo.

Há duas maneiras básicas de descrever essa informação conjunta. Em vetores discretos, distribuímos massa entre pontos; em vetores contínuos, distribuímos probabilidade por meio de uma densidade sobre regiões. Veremos os dois casos separadamente.

6.2 Vetores aleatórios discretos

No caso discreto, a distribuição conjunta vira uma tabela: cada célula contém $p(x,y) = \mathbf{P}(X = x, Y = y)$, e as marginais aparecem somando linhas ou colunas. Assim, uma função de probabilidade conjunta deve satisfazer

$$p(x,y) \geq 0, \qquad \sum_x \sum_y p(x,y) = 1,$$

com soma sobre o suporte, finito ou enumerável. Reciprocamente, qualquer tabela não negativa com soma total igual a 1 define uma distribuição conjunta.

A função de distribuição é recuperada acumulando as massas:

$$F(a,b) = \mathbf{P}(X \leq a, Y \leq b) = \sum_{x \leq a} \sum_{y \leq b} p(x,y).$$

Proposição 6.2

O vetor $(X_1, \ldots, X_n)$ é discreto se, e somente se, as variáveis $X_1, \ldots X_n$ são discretas.

Demonstração. Se o vetor assume valores em um conjunto enumerável, cada coordenada assume valores na projeção desse conjunto e, portanto, é discreta. Reciprocamente, se cada X_i assume valores em um conjunto enumerável C_i, então o vetor assume valores no produto $C_1 \times \cdots \times C_n$, que também é enumerável. ■

Para obter a **função de probabilidade marginal** de X, basta somar $p(x,y)$ sobre todos os valores possíveis de y, resultando em:

$$\begin{aligned} p_X(x) &= \mathbf{P}(X = x) \\ &= \sum_{y:p(x,y)>0} p(x,y). \end{aligned}$$

Analogamente, a função de probabilidade marginal de Y é dada pela soma de $p(x,y)$ sobre todos os valores possíveis de x:

$$p_Y(y) = \sum_{x:p(x,y)>0} p(x,y).$$

Exemplo 6.2.

Uma caixa contém cinco lâmpadas, duas delas defeituosas. As lâmpadas são testadas em ordem aleatória, sem reposição. Seja X a posição em que aparece a primeira defeituosa e Y a posição da segunda. Então $1 \leq X < Y \leq 5$.

Há apenas $\binom{5}{2} = 10$ pares possíveis de posições para as duas lâmpadas defeituosas, todos equiprováveis. Por isso, cada par (x,y) com $1 \leq x < y \leq 5$ tem probabilidade $1/10$. A tabela conjunta é:

X	1	2	3	4
$Y = 2$	1/10	0	0	0
$Y = 3$	1/10	1/10	0	0
$Y = 4$	1/10	1/10	1/10	0
$Y = 5$	1/10	1/10	1/10	1/10

Probabilidades de eventos são somas de células. Por exemplo, $\mathbf{P}(X > 2) = 3/10$. As marginais surgem somando as colunas e as linhas. Para X:

X	1	2	3	4
$\mathbf{P}(X = x)$	$\frac{4}{10}$	$\frac{3}{10}$	$\frac{2}{10}$	$\frac{1}{10}$

De modo análogo, para Y:

Y	2	3	4	5
$\mathbf{P}(Y = y)$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{4}{10}$

◁

Exemplo 6.3.

Duas canetas esferográficas são selecionadas aleatoriamente de uma caixa que contém 3 canetas azuis, 2 canetas vermelhas e 3 canetas verdes. Se X é o número de canetas azuis selecionadas e Y é o número de canetas vermelhas selecionadas, encontre

1. a função de probabilidade conjunta $p(x, y)$,
2. $\mathbf{P}[(X, Y) \in A]$, onde A é a região $\{(x, y)|x + y \leq 1\}$.

Solução. Os pares possíveis de valores (x, y) são $(0,0), (0,1), (1,0), (1,1), (0,2)$ e $(2,0)$.

1. Por exemplo, $p(0,1)$ corresponde à escolha de uma caneta vermelha e uma verde. Entre as $\binom{8}{2} = 28$ duplas equiprováveis, há $\binom{2}{1}\binom{3}{1} = 6$ desse tipo; logo $p(0,1) = 3/14$. O mesmo raciocínio fornece as demais células da tabela, cuja soma total é 1.
2. A probabilidade de que (X, Y) pertença à região A é

$$\begin{aligned}\mathbf{P}[(X, Y) \in A] &= \mathbf{P}(X + Y \leq 1) = p(0,0) + p(0,1) + p(1,0) \\ &= \frac{3}{28} + \frac{3}{14} + \frac{9}{28} = \frac{9}{14}.\end{aligned}$$

$p(x,y)$	$x = 0$	$x = 1$	$x = 2$	Totais
$y = 0$	$\frac{3}{28}$	$\frac{9}{28}$	$\frac{3}{28}$	$\frac{15}{28}$
$y = 1$	$\frac{3}{14}$	$\frac{3}{14}$	0	$\frac{3}{7}$
$y = 2$	$\frac{1}{28}$	0	0	$\frac{1}{28}$
Totais	$\frac{5}{14}$	$\frac{15}{28}$	$\frac{3}{28}$	1

Tabela 6.1: Distribuição de Probabilidade Conjunta para o Exemplo 6.3.

◁

Exemplo 6.4.

Suponha que realizamos sucessivos lançamentos de uma moeda com probabilidade p de cair cara. Defina as variáveis aleatórias X_1 como o número do lançamento em que ocorre a primeira cara, e X_2 como o número do lançamento em que ocorre a segunda cara. O suporte da distribuição conjunta de (X_1, X_2) não é finito, pois X_1 e X_2 podem assumir valores arbitrariamente grandes, desde que $1 \leq X_1 \leq X_2 - 1$.

Para $x_2 \geq x_1+1$ e $x_1 \geq 1$, podemos recorrer às propriedades da distribuição geométrica para determinar que

$$\mathbf{P}(X_1 = x_1, X_2 = x_2) \;=\; \left[p\,(1-p)^{\,x_1-1}\right]\left[p\,(1-p)^{\,x_2-x_1-1}\right].$$

O primeiro termo representa a probabilidade de ser necessário x_1 lançamentos para obter a primeira cara, e o segundo termo corresponde à probabilidade de que sejam necessários $x_2 - x_1$ lançamentos adicionais para aparecer a segunda cara. Essas probabilidades multiplicam-se em razão da independência dos lançamentos. Simplificando, obtemos

$$\mathbf{P}(X_1 = x_1, X_2 = x_2) = p^2(1-p)^{x_2-2}, \quad \text{para } 1 \leq x_1 < x_2.$$

A normalização também pode ser verificada diretamente. Escrevendo $q = 1-p$,

$$\sum_{x_2=2}^{\infty}\sum_{x_1=1}^{x_2-1} p^2 q^{x_2-2} = p^2\sum_{k=0}^{\infty}(k+1)q^k = \frac{p^2}{(1-q)^2} = 1.$$

◁

Exemplo 6.5.

Retomando o Exemplo 6.4, já identificamos que a distribuição marginal de X_1 deve ser geométrica com parâmetro p. Para confirmar essa afirmação de outra forma, podemos recorrer à fórmula da distribuição marginal. No caso em questão, a função de probabilidade marginal de X_1 é dada por:

$$\begin{aligned}\mathbf{P}(X_1 = x_1) &= \sum_{x_2=x_1+1}^{\infty} \mathbf{P}(X_1 = x_1, X_2 = x_2)\\ &= \sum_{x_2=x_1+1}^{\infty} p^2(1-p)^{x_2-2}\\ &= p^2(1-p)^{x_1-1}\sum_{k=0}^{\infty}(1-p)^k.\end{aligned}$$

Essa soma é uma série geométrica de razão $(1-p)$ e primeiro termo $p^2(1-p)^{x_1-1}$, cuja soma é dada por:

$$\begin{aligned}\mathbf{P}(X_1 = x_1) &= p^2(1-p)^{x_1-1}\frac{1}{p}\\ &= p(1-p)^{x_1-1}, \quad \text{para } x_1 \geq 1.\end{aligned}$$

Isso mostra que X_1 segue uma distribuição geométrica com parâmetro p.

A função de probabilidade marginal de X_2 é obtida somando sobre x_1:

$$\mathbf{P}(X_2 = x_2) = \sum_{x_1=1}^{x_2-1} \mathbf{P}(X_1 = x_1, X_2 = x_2).$$

Substituindo a função de probabilidade conjunta:

$$\mathbf{P}(X_2 = x_2) = \sum_{x_1=1}^{x_2-1} p^2(1-p)^{x_2-2}.$$

O termo $p^2(1-p)^{x_2-2}$ não depende de x_1, então a soma simplesmente conta o número de termos, que é $x_2 - 1$:

$$\mathbf{P}(X_2 = x_2) = (x_2 - 1)p^2(1-p)^{x_2-2}, \quad \text{para } x_2 \geq 2.$$

◁

Modelo multinomial A binomial conta quantas vezes ocorre um resultado em ensaios com duas categorias. Com m categorias, as contagens devem ser estudadas em conjunto. Se, em cada um de n ensaios independentes, a categoria i ocorre com probabilidade p_i, $i = 1, \ldots, m$, e $\sum_i p_i = 1$, então o vetor $X = (X_1, \ldots, X_m)$ das contagens tem distribuição multinomial:

Definição 6.2 — Modelo multinomial

Para inteiros $k_i \geq 0$ com $k_1 + \cdots + k_m = n$,

$$\mathbf{P}(X_1 = k_1, \ldots, X_m = k_m) = \frac{n!}{k_1! \cdots k_m!} p_1^{k_1} \cdots p_m^{k_m}.$$

Fora desse suporte, a probabilidade é zero. O coeficiente multinomial conta as sequências de n resultados que produzem as mesmas contagens.

Exemplo 6.6.
Em dez lançamentos independentes de um dado equilibrado, seja X_i o número de ocorrências da face i. Então $(X_1,\ldots,X_6)$ é multinomial e

$$\mathbf{P}(X_1 = k_1,\ldots,X_6 = k_6) = \frac{10!}{k_1!\cdots k_6!}\left(\frac{1}{6}\right)^{10}, \qquad k_1 + \cdots + k_6 = 10.$$

◁

6.3 Vetores aleatórios contínuos

No caso contínuo, as somas sobre pontos são substituídas por integrais sobre regiões. Por isso, antes de integrar, convém sempre identificar duas coisas: o suporte da densidade e a parte desse suporte que representa o evento de interesse.

No caso bidimensional, dizemos que (X,Y) é contínuo quando existe uma função não negativa f, com integral total igual a 1, tal que, para toda região C considerada,

$$\mathbf{P}((X,Y) \in C) = \iint_{(x,y)\in C} f(x,y)dxdy \tag{6.2}$$

Se A e B são quaisquer conjuntos de números reais, então, para o conjunto

$$C = \{(x,y) : x \in A, y \in B\},$$

obtemos da Equação 6.2 que

$$\mathbf{P}(X \in A, Y \in B) = \int_B \int_A f(x,y)dxdy.$$

Interpretação da densidade Da função de distribuição conjunta,

$$F(a,b) = \mathbf{P}(X \in (-\infty,a], Y \in (-\infty,b]) = \int_{-\infty}^{b}\int_{-\infty}^{a} f(x,y)dxdy \tag{6.3}$$

obtemos, nos pontos em que as derivadas existem,

$$f(a,b) = \frac{\partial^2}{\partial a \partial b}F(a,b).$$

O valor $f(a,b)$ não é uma probabilidade pontual. Ele indica como a probabilidade se concentra nas proximidades de (a,b). Para pequenos $\Delta a, \Delta b > 0$,

$$\mathbf{P}(a < X < a+\Delta a, b < Y < b+\Delta b) = \int_b^{b+\Delta b}\int_a^{a+\Delta a} f(x,y)dxdy \tag{6.4}$$

$$\approx f(a,b)\Delta a\Delta b \tag{6.5}$$

quando f é contínua em (a,b). A aproximação corresponde ao volume de um paralelepípedo de base $\Delta a\Delta b$ e altura $f(a,b)$.

Marginais Se X e Y são variáveis aleatórias conjuntamente contínuas, então elas também são contínuas individualmente. As funções de densidade de probabilidade marginais de X e Y podem ser determinadas da seguinte forma:

$$\begin{aligned}\mathbf{P}(X \in A) &= \mathbf{P}(X \in A, Y \in (-\infty,\infty)) \\ &= \int_A \int_{-\infty}^{\infty} f(x,y)dydx \\ &= \int_A f_X(x)dx\end{aligned}$$

onde

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy.$$

Similarmente, a função densidade de probabilidade de Y é dada por

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx.$$

Exemplo 6.7 — Distribuição uniforme bidimensional.

A distribuição uniforme no quadrado unitário é o exemplo bidimensional mais simples. Se (X,Y) é uniforme em $(0,1)^2$, a densidade é constante no quadrado e zero fora dele. Como a área do suporte é 1, a constante também deve ser 1:

$$f_{XY}(x,y) = \begin{cases} 1 & \text{se } 0 < x < 1, 0 < y < 1 \\ 0 & \text{caso contrário} \end{cases}$$

A Figura 6.3 mostra o suporte e a superfície da densidade. Como a altura é 1, probabilidades são simplesmente áreas. Assim, a região $Y > 1/2$ ocupa metade do quadrado e tem probabilidade $1/2$.

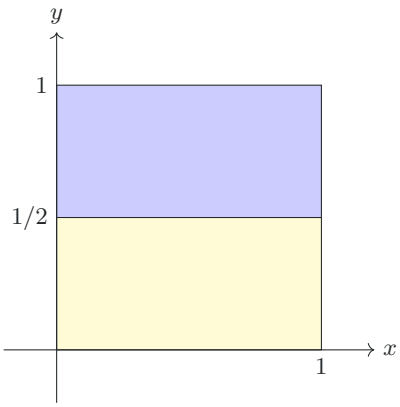


Figura 6.1: Região 1

Para calcular $\mathbf{P}(X + 2Y > 2)$, identificamos no quadrado a região acima da reta $x + 2y = 2$, mostrada na Figura 6.2.

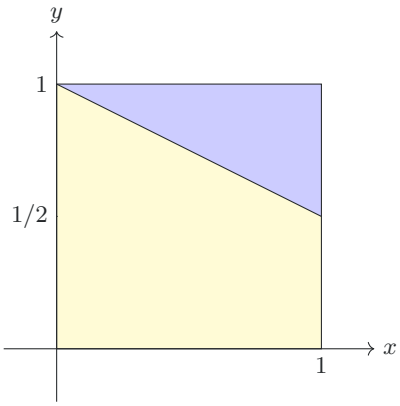


Figura 6.2: Região 2

O triângulo sombreado tem área 1/4; portanto $\mathbf{P}(X + 2Y > 2) = 1/4$.

◁

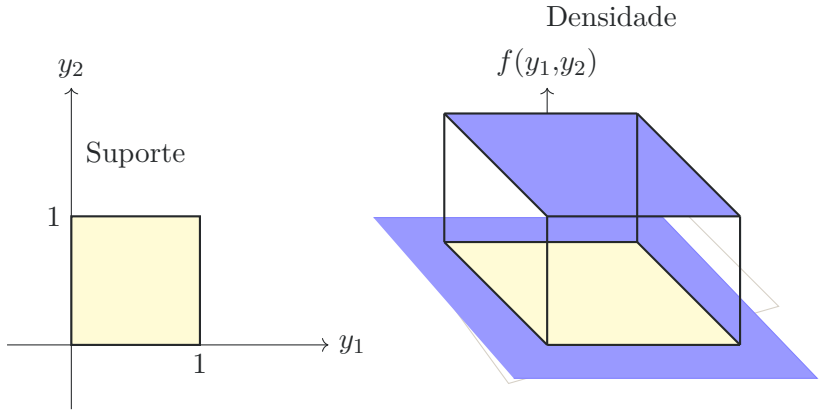


Figura 6.3: Função densidade para a distribuição uniforme.

Exemplo 6.8.

A função densidade conjunta de X e Y é dada por

$$f_{XY}(x,y) = \begin{cases} 2e^{-x}e^{-2y} & 0 < x < \infty, 0 < y < \infty \\ 0 & \text{caso contrário} \end{cases}$$

Calcule

1. $\mathbf{P}(X > 1, Y < 1)$
2. $\mathbf{P}(X < Y)$
3. $\mathbf{P}(X < a)$.

Solução.

1

$$\begin{aligned} \mathbf{P}(X > 1, Y < 1) &= \int_0^1 \int_1^\infty 2e^{-x}e^{-2y}dxdy \\ &= \int_0^1 2e^{-2y}\left(-e^{-x}|_1^\infty\right)dy \\ &= e^{-1}\int_0^1 2e^{-2y}dy \\ &= e^{-1}\left(1 - e^{-2}\right) \end{aligned}$$

2

$$\begin{aligned}
\mathbf{P}(X < Y) &= \iint_{(x,y):x<y} 2e^{-x}e^{-2y}dxdy \\
&= \int_0^\infty \int_0^y 2e^{-x}e^{-2y}dxdy \\
&= \int_0^\infty 2e^{-2y}\left(1 - e^{-y}\right)dy \\
&= \int_0^\infty 2e^{-2y}dy - \int_0^\infty 2e^{-3y}dy \\
&= 1 - \frac{2}{3} \\
&= \frac{1}{3}
\end{aligned}$$

3

$$\begin{aligned}
\mathbf{P}(X < a) &= \int_0^a \int_0^\infty 2e^{-2y}e^{-x}dydx \\
&= \int_0^a e^{-x}dx \\
&= 1 - e^{-a}
\end{aligned}$$

◁

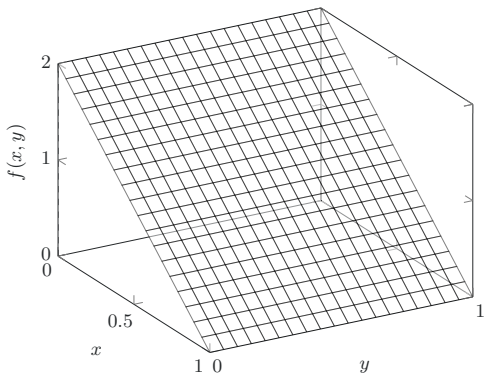


Figura 6.4: Função densidade de probabilidade para o Exemplo 6.9

Exemplo 6.9.

Em um processo de fabricação de chips de computadores, pode haver dois tipos de impurezas que afetam a qualidade do produto final. Para um certo lote de chips, seja X a proporção total de impurezas encontradas no lote, e seja Y a proporção das impurezas do tipo 1 em relação ao total de impurezas presentes. Suponha que, após analisar diversos lotes, a distribuição conjunta de X e Y possa ser adequadamente modelada pela função de densidade:

$$f_{X,Y}(x,y) = \begin{cases} 2(1-x), & \text{para } 0 \le x \le 1, 0 \le y \le 1, \\ 0, & \text{caso contrário.} \end{cases}$$

Solução.

Desejamos calcular a probabilidade de que a proporção total de impurezas X seja inferior a 0,5 e que a proporção de impurezas do tipo 1, Y, esteja entre 0,4 e 0,7.

$$\begin{aligned} \mathbf{P}(0 \le X \le 0.5,\ 0.4 \le Y \le 0.7) &= \int_{0.4}^{0.7} \int_{0}^{0.5} 2(1-x)\, dx\, dy \\ &= \int_{0.4}^{0.7} \Big[-(1-x)^2\Big]_0^{0.5}\, dy \\ &= \int_{0.4}^{0.7} (0.75)\, dy \\ &= 0.75 \int_{0.4}^{0.7} dy \\ &= 0.75(0.7 - 0.4) \\ &= 0.75 \cdot 0.3 = 0.225. \end{aligned}$$

Portanto, a probabilidade de que X seja menor que 0,5 e Y esteja entre 0,4 e 0,7 é 0,225.

◁

Exemplo 6.10.

Considere um círculo de raio R e suponha que um ponto em seu interior seja escolhido aleatoriamente e uniformemente distribuído no interior do círculo. Se o centro do círculo está na origem do sistema de coordenadas, e X e Y correspondem às coordenadas do ponto escolhido, então, como (X, Y) tem a mesma probabilidade de estar na vizinhança de qualquer ponto no círculo, a

função densidade conjunta de X e Y é dada pela uniforme:

$$f(x,y) = \begin{cases} K \text{ se } x^2 + y^2 \leq R^2 \\ 0 \text{ se } x^2 + y^2 > R^2 \end{cases}$$

para algum valor de K.

1. Determine K.
2. Encontre as funções densidade marginais de X e Y.
3. Calcule a probabilidade de que D, a distância da origem ao ponto selecionado, seja menor ou igual a a.
4. Determine $\mathbf{E}[D]$, a esperança da distância D.

Solução.

1. **Determinação de c:**

   Para que $f(x,y)$ seja uma função densidade de probabilidade, devemos ter:

   $$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dy\,dx = 1.$$

   Assim, obtemos:

   $$K\iint_{x^2+y^2\leq R^2} dy\,dx = 1.$$

   Podemos calcular $\iint_{x^2+y^2\leq R^2} dy\,dx$ usando coordenadas polares, ou mais diretamente, reconhecendo que essa integral representa a área do círculo de raio R, que é πR^2. Portanto,

   $$K = \frac{1}{\pi R^2}.$$

2. **Funções densidade marginais de X e Y:**

   Para encontrar a densidade marginal de X, temos:

   $$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f(x,y)\,dy \\ &= \frac{1}{\pi R^2}\int_{x^2+y^2\leq R^2} dy \\ &= \frac{1}{\pi R^2}\int_{-\sqrt{R^2-x^2}}^{\sqrt{R^2-x^2}} dy \\ &= \frac{2}{\pi R^2}\sqrt{R^2-x^2}, \quad \text{para } x^2 \leq R^2. \end{aligned}$$

Para $x^2 > R^2$, temos $f_X(x) = 0$. Por simetria, a densidade marginal de Y é dada por:

$$\begin{aligned} f_Y(y) &= \frac{2}{\pi R^2}\sqrt{R^2 - y^2}, \quad \text{para } y^2 \leq R^2, \\ &= 0, \quad \text{para } y^2 > R^2. \end{aligned}$$

3. **Probabilidade de que $D \leq a$, onde $D = \sqrt{X^2 + Y^2}$:**

   Para $0 \leq a \leq R$, a função de distribuição de D é:

$$\begin{aligned} F_D(a) &= \mathbf{P}(\sqrt{X^2 + Y^2} \leq a) \\ &= \mathbf{P}(X^2 + Y^2 \leq a^2) \\ &= \iint_{x^2+y^2 \leq a^2} f(x, y)\, dy\, dx \\ &= \frac{1}{\pi R^2} \iint_{x^2+y^2 \leq a^2} dy\, dx \\ &= \frac{\pi a^2}{\pi R^2} \\ &= \frac{a^2}{R^2}, \end{aligned}$$

   Para $a < 0$, $F_D(a) = 0$; para $a \geq R$, $F_D(a) = 1$.

4. **Esperança de D, $\mathbf{E}[D]$:**

   Do item anterior, temos que a densidade de D é:

$$f_D(a) = \frac{2a}{R^2}, \quad 0 \leq a \leq R.$$

   Portanto,

$$\mathbf{E}[D] = \frac{2}{R^2}\int_0^R a^2\, da = \frac{2R}{3}.$$

◁

6.4 Independência

Conhecer as marginais não basta, em geral, para reconstruir a distribuição conjunta. Quando bastaria? Precisamente quando observar uma coordenada não traz informação sobre a outra. A independência formaliza essa situação e permite recuperar a distribuição conjunta multiplicando as marginais. É

uma condição forte: ausência de correlação, estudada depois, não basta. Formalmente, X e Y são independentes se, para quaisquer conjuntos borelianos $A,B \subseteq \mathbb{R}$,

$$\mathbf{P}(X \in A, Y \in B) = \mathbf{P}(X \in A)\mathbf{P}(Y \in B)$$

Em outras palavras, os eventos $[X \in A]$ e $[Y \in B]$ devem ser independentes para todas essas escolhas.

Definição 6.3

As variáveis aleatórias $X_1, \ldots, X_n$ são **independentes** se

$$\mathbf{P}\left(X_1 \in B_1, \ldots, X_n \in B_n\right) = \prod_{i=1}^{n} \mathbf{P}\left(X_i \in B_i\right). \tag{6.6}$$

A igualdade deve valer para todos os conjuntos borelianos $B_1, \ldots, B_n \subseteq \mathbb{R}$.

Para qualquer família de variáveis aleatórias independentes, qualquer subfamília é também formada por variáveis independentes. Por exemplo, se X, Y e Z são independentes, então X e Y também o são.

Os critérios mais úteis são fatorizações da distribuição conjunta. Eles permitem reconhecer independência sem voltar à definição para todos os conjuntos.

Proposição 6.3 — Critérios para independência

1. As variáveis $X_1, \ldots, X_n$ são independentes se, e somente se,

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} F_{X_i}(x_i), \qquad (x_1, \ldots, x_n) \in \mathbb{R}^n.$$

2. Se $X_1, \ldots, X_n$ são discretas, elas são independentes se, e somente se,

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_{X_i}(x_i) \tag{6.7}$$

para todo $(x_1, \ldots, x_n)$.

3. Suponha que X e Y tenham densidades conjunta e marginais contínuas. Então são independentes se, e somente se,

$$f_{X,Y}(x,y) = f_X(x)f_Y(y), \qquad (x,y) \in \mathbb{R}^2.$$

Demonstração. Para 1, independência aplicada aos eventos $[X_i \leq x_i]$ fornece a fatorização da função de distribuição. Reciprocamente, se a função de distribuição fatora, diferenças sucessivas mostram que as probabilidades de retângulos semiabertos fatoram; como esses retângulos determinam a distribuição conjunta, segue a independência.

Para 2, no caso bidimensional, a independência aplicada aos eventos $[X = x]$ e $[Y = y]$ dá $p(x,y) = p_X(x)p_Y(y)$. Na direção inversa, para quaisquer conjuntos A e B,

$$\begin{aligned}\mathbf{P}(X \in A, Y \in B) &= \sum_{x\in A}\sum_{y\in B} p_X(x)p_Y(y)\\ &= \mathbf{P}(X \in A)\mathbf{P}(Y \in B).\end{aligned}$$

O argumento em dimensão maior é o mesmo.

Para 3, se X e Y são independentes, então $F_{X,Y}(x,y) = F_X(x)F_Y(y)$; derivando, obtemos $f_{X,Y}(x,y) = f_X(x)f_Y(y)$. Reciprocamente, se as densidades fatoram,

$$F_{X,Y}(x,y) = \int_{-\infty}^{x} f_X(u)\,du \int_{-\infty}^{y} f_Y(v)\,dv = F_X(x)F_Y(y),$$

e 1 conclui a prova. ∎

Exemplo 6.11.

Suponha que X e Y sejam variáveis aleatórias com a função de distribuição conjunta dada por

$$F_{X,Y}(x,y) = \begin{cases} 1 - e^{-x} - e^{-y} + e^{-x-y} & \text{se } x, y \geq 0,\\ 0 & \text{caso contrário.}\end{cases}$$

Para encontrar a função de distribuição marginal de X, podemos considerar:

$$F_X(x) = \lim_{y\to\infty} F_{X,Y}(x,y) = \begin{cases} 1 - e^{-x} & \text{se } x \geq 0,\\ 0 & \text{caso contrário.}\end{cases}$$

Isso indica que X segue uma distribuição exponencial com parâmetro 1. Um cálculo análogo mostra que Y também segue uma distribuição exponencial com o mesmo parâmetro.

Podemos verificar a independência das variáveis X e Y observando que:

$$F_{X,Y}(x,y) = F_X(x)F_Y(y) \quad \text{para todos } x,y \in \mathbb{R},$$

o que confirma que X e Y são variáveis independentes.

◁

Exemplo 6.12.

Sejam $n+m$ experimentos independentes, cada um com probabilidade de sucesso p. Se X conta o número de sucessos nas primeiras n tentativas e Y nas últimas m, então X e Y são variáveis aleatórias independentes. Isso ocorre porque o resultado das primeiras n tentativas não influencia o número de sucessos nas m tentativas restantes. De fato, para x e y inteiros,

$$\begin{aligned}\mathbf{P}(X=x,Y=y) &= \binom{n}{x} p^x(1-p)^{n-x}\binom{m}{y} p^y(1-p)^{m-y} \quad 0 \le x \le n,\\ &= \mathbf{P}(X=x)\mathbf{P}(Y=y)\end{aligned}$$

Por outro lado, o número de sucessos nas primeiras n tentativas, X, e o número total de sucessos, Z são dependentes.

◁

Exemplo 6.13 — Afinamento de Poisson e independência.

No capítulo anterior vimos que, se $N \sim \text{Poisson}(\lambda)$ e cada ocorrência é classificada independentemente como detectada com probabilidade p, o número X de detectadas é $\text{Poisson}(\lambda p)$. Seja $Y = N - X$ o número de não detectadas. Além das marginais, o afinamento produz independência.

De fato, para $i,j \ge 0$,

$$\begin{aligned}\mathbf{P}(X=i,Y=j) &= \mathbf{P}(N=i+j)\binom{i+j}{i}p^i(1-p)^j\\ &= \frac{e^{-\lambda p}(\lambda p)^i}{i!}\,\frac{e^{-\lambda(1-p)}[\lambda(1-p)]^j}{j!}.\end{aligned}$$

A última expressão é o produto das massas de $\text{Poisson}(\lambda p)$ e $\text{Poisson}(\lambda(1-p))$. Portanto, X e Y são independentes.

◁

Exemplo 6.14.

Suponha que X e Y sejam variáveis aleatórias com a seguinte função de distribuição conjunta:

$$F_{X,Y}(x,y) = \begin{cases} 1 - e^{-x} - e^{-y} + e^{-x-y}, & \text{se } x, y \geq 0, \\ 0, & \text{caso contrário.} \end{cases}$$

Para determinar a função de distribuição marginal de X, podemos calcular:

$$F_X(x) = \lim_{y \to \infty} F_{X,Y}(x,y) = \begin{cases} 1 - e^{-x}, & \text{se } x \geq 0, \\ 0, & \text{caso contrário.} \end{cases}$$

Assim, podemos concluir que X segue uma distribuição exponencial com parâmetro 1. Um cálculo semelhante revela que Y também segue uma distribuição exponencial com o mesmo parâmetro.

Além disso, podemos verificar que:

$$F_{X,Y}(x,y) = F_X(x)F_Y(y) \quad \text{para todos } x, y \in \mathbb{R},$$

o que implica que X e Y são variáveis independentes.

◁

Exemplo 6.15 — Variáveis Aleatórias Dependentes.

Para ver a dependência geometricamente, considere X e Y como definidos no Exemplo 6.10. Para $-R < x < R$ e $-R < y < R$, temos:

$$f_X(x)f_Y(y) = \frac{4\sqrt{R^2 - x^2}\sqrt{R^2 - y^2}}{\pi^2 R^4},$$

o que não coincide com a densidade conjunta dessas variáveis e assim concluímos que X e Y são variáveis aleatórias dependentes.

Isso está de acordo com nossa intuição de dependência, pois, quando X está próximo de R, Y deve estar próximo de zero. Assim, obter informações sobre X nos fornece informações sobre Y.

◁

Exemplo 6.16.

Dois amigos planejam se encontrar na biblioteca dentro do intervalo de uma hora. Seus horários de chegada são independentes e distribuídos uniformemente ao longo do período de uma hora. Cada um concorda em esperar por 15 minutos, ou até o final da hora. Se o outro não aparecer nesse tempo, a pessoa parte. Qual é a probabilidade de os dois amigos se encontrarem?

Solução. Se X_1 denota o horário de chegada do primeiro amigo no intervalo (0,1), o período de uma hora, e se X_2 denota o horário de chegada do segundo amigo, então (X_1, X_2) pode ser modelado como tendo uma distribuição uniforme bidimensional sobre o quadrado unitário; isto é,

$$f(x_1, x_2) = \begin{cases} 1, & 0 \leq x_1 \leq 1, 0 \leq x_2 \leq 1 \\ 0, & \text{caso contrário.} \end{cases}$$

O evento de os dois amigos se encontrarem depende do tempo Y entre suas chegadas, onde

$$Y = |X_1 - X_2|.$$

Podemos resolver o problema encontrando primeiro a distribuição de Y. Como

$$\begin{aligned} Y \leq y &\Rightarrow |X_1 - X_2| \leq y \\ &\Rightarrow -y \leq X_1 - X_2 \leq y. \end{aligned}$$

A Figura 6.5 mostra que a região quadrada sobre a qual (X_1, X_2) tem probabilidade positiva e a região definida por $Y \leq y$. A probabilidade de $Y \leq y$ pode ser encontrada integrando a função de densidade conjunta de (X_1, X_2) sobre a região de seis lados mostrada no centro da Figura 6.5. Isso pode ser simplificado integrando sobre os triângulos (A_1 e A_2) e subtraindo de 1, como veremos agora.

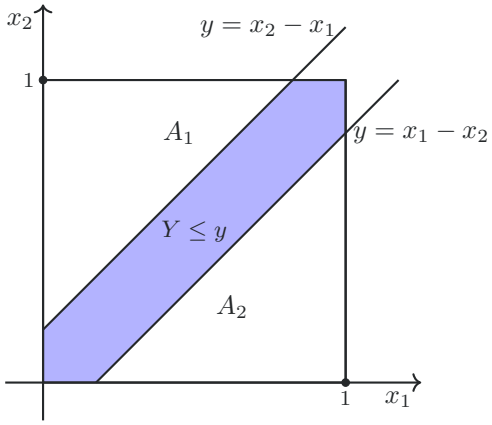


Figura 6.5: Tempo de Espera

Temos

$$\begin{aligned} F_Y(y) = \mathbf{P}(Y \leq y) &= \int\int_{|x_1 - x_2| \leq y} f(x_1, x_2)\, \mathrm{d}x_1\, \mathrm{d}x_2 \\ &= 1 - \int_{A_1}\int f(x_1, x_2)\, \mathrm{d}x_1\, \mathrm{d}x_2 - \int_{A_2}\int f(x_1, x_2)\, \mathrm{d}x_1\, \mathrm{d}x_2 \\ &= 1 - \int_y^1 \int_0^{x_2 - y} (1) \mathrm{d}x_1\, \mathrm{d}x_2 - \int_0^{1-y}\int_{x_2+y}^1 (1) \mathrm{d}x_1\, \mathrm{d}x_2 \end{aligned}$$

$$\begin{aligned} &= 1 - \int_y^1 (x_2 - y)\, \mathrm{d}x_2 - \int_0^{1-y} (1 - y - x_2)\, \mathrm{d}x_2 \\ &= 1 - \left[\frac{1}{2}x_2^2 - x_2 y\right]_y^1 - \left[x_2 - x_2 y - \frac{1}{2}x_2^2\right]_0^{1-y} \\ &= 1 - \frac{1}{2}(1-y)^2 - \frac{1}{2}(1-y)^2 \\ &= 1 - (1-y)^2, \quad 0 \leq y \leq 1. \end{aligned}$$

O problema solicita a probabilidade de um encontro se cada amigo esperar até 15 minutos. Como 15 minutos é 1/4 de hora, isso pode ser encontrado

avaliando

$$\begin{aligned}\mathbf{P}\left(Y \le \frac{1}{4}\right) &= F_Y\left(\frac{1}{4}\right) \\ &= 1 - \left(1 - \frac{1}{4}\right)^2 \\ &= 1 - \left(\frac{3}{4}\right)^2 \\ &= \frac{7}{16} \\ &= 0.4375.\end{aligned}$$

Há menos de 50% de chance de que os dois amigos se encontrem sob essa regra.

◁

Exemplo 6.17 — Agulha de Buffon.

Uma agulha de comprimento $2a$ é lançada ao acaso sobre um plano coberto por retas paralelas separadas por distância $2b$, com $a < b$. Seja X a distância do centro da agulha à reta mais próxima e Θ o ângulo entre a agulha e a direção perpendicular às retas. Pelo modelo geométrico,

$$X \sim \mathrm{Unif}(0,b), \qquad \Theta \sim \mathrm{Unif}(0,\pi/2),$$

e tomamos X e Θ independentes. Assim, no retângulo $0 < x < b$, $0 < \theta < \pi/2$, a densidade conjunta é constante:

$$f_{X,\Theta}(x,\theta) = \frac{2}{\pi b}.$$

A agulha cruza uma reta exatamente quando $X < a\cos\Theta$. Portanto,

$$\begin{aligned}p &= \mathbf{P}(X < a\cos\Theta) \\ &= \frac{2}{\pi b}\int_0^{\pi/2}\int_0^{a\cos\theta} dx\,d\theta = \frac{2a}{\pi b}.\end{aligned}$$

Se, em n lançamentos, ocorrerem n_ℓ cruzamentos, a frequência relativa n_ℓ/n sugere a aproximação

$$\pi \approx \frac{2an}{bn_\ell}.$$

◁

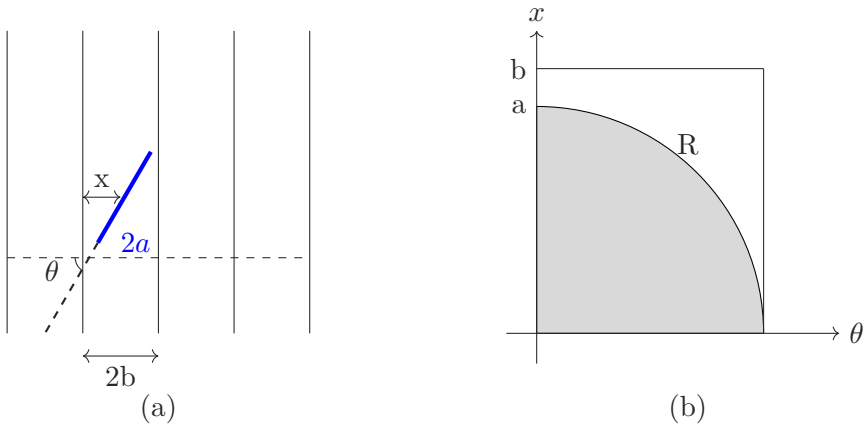


Figura 6.6: Agulha de Buffon

Exemplo 6.18 — Normal bivariada.

Dizemos que (X,Y) tem distribuição normal bivariada com parâmetros μ_1,μ_2, $\sigma_1,\sigma_2 > 0$ e $-1 < \rho < 1$ quando

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-}{}\right.\right.$$

A forma da densidade parece complicada, mas completar o quadrado revela sua estrutura. Escrevendo $u = (x-\mu_1)/\sigma_1$ e $v = (y-\mu_2)/\sigma_2$,

$$\frac{u^2 - 2\rho uv + v^2}{1-\rho^2} = u^2 + \frac{(v-\rho u)^2}{1-\rho^2}.$$

Por isso, para cada x, a densidade pode ser vista como o produto da densidade $N(\mu_1,\sigma_1^2)$ em x por uma densidade normal em y com média

$$\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1)$$

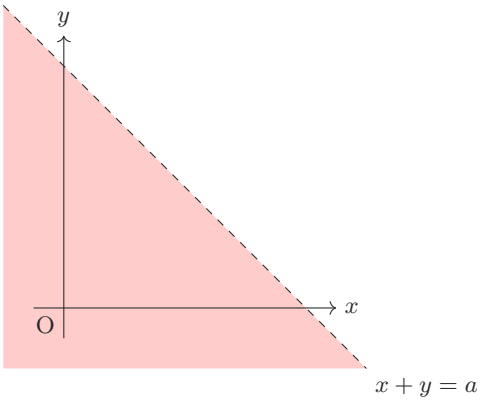
e variância $(1-\rho^2)\sigma_2^2$. Integrando em y, obtemos $X \sim N(\mu_1,\sigma_1^2)$; por simetria, $Y \sim N(\mu_2,\sigma_2^2)$.

Quando $\rho = 0$, a densidade conjunta se reduz ao produto das duas marginais, e X e Y são independentes. Para $\rho \neq 0$, a fatorização falha. Mais adiante, ao estudar correlação, veremos que o parâmetro ρ coincide com o coeficiente de correlação desse vetor.

◁

6.5 Soma de variáveis aleatórias independentes

Somas aparecem ao acumular tempos, custos, contagens ou erros. Mas conhecer as distribuições de X e Y basta para conhecer a de $X+Y$? Em geral, não: a dependência também importa. Quando X e Y são independentes, porém, a soma pode ser construída diretamente a partir das marginais, e a geometria é clara: integramos ao longo das retas $x + y = z$. Suponha inicialmente que (X,Y) tenha densidade conjunta $f_{X,Y}$. Então:



$$
\begin{aligned}
F_{X+Y}(z) &= \mathbf{P}(X+Y \le z) && (6.8)\\
&= \iint_{x+y\le z} f_{X,Y}(x,y)dxdy && (6.9)\\
&= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{z-x} f_{X,Y}(x,y)dy\right]dx \quad \text{subst. } y = u-x. && (6.10)\\
&= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{z} f_{X,Y}(x,u-x)du\right]dx && (6.11)\\
& && (6.12)
\end{aligned}
$$

Para obter a densidade tomamos a derivada:

$$\frac{dF_{X+Y}(z)}{dz} = \frac{d}{dz}\left\{\int_{-\infty}^{z}\left[\int_{-\infty}^{\infty} f_{X,Y}(x,u-x)\,dx\right]du\right\},$$

assim, concluímos que

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_{XY}(x,z-x)dx. \tag{6.13}$$

Se as variáveis X e Y são independentes podemos reescrever 6.13 como

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(z-t) f_Y(t) dt = \int_{-\infty}^{\infty} f_X(t) f_Y(z-t) dt \tag{6.14}$$

Com isso está provada a seguinte proposição.

Proposição 6.4

Sejam X e Y variáveis aleatórias.

1. Se X e Y têm densidade conjunta $f_{X,Y}(x,y)$, então

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_{X,Y}(z-t,t) dt = \int_{-\infty}^{\infty} f_{X,Y}(t,z-t) dt$$

2. Se X e Y são independentes e possuem densidades f_X e f_Y, então (por 1 e a 3.) $X+Y$ tem densidade

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(z-t) f_Y(t) dt = \int_{-\infty}^{\infty} f_X(t) f_Y(z-t) dt$$

O mesmo raciocínio fornece, para a diferença,

$$f_{X-Y}(w) = \int_{-\infty}^{\infty} f_{X,Y}(w+y,y)\, dy.$$

Se X e Y forem independentes, essa expressão é a convolução de f_X com a densidade de $-Y$.

Definição 6.4

Se f_1 e f_2 são densidades de variáveis aleatórias, sua **convolução** $f_1 * f_2$ é definida por

$$f_1 * f_2(x) = \int_{-\infty}^{\infty} f_1(x-t) f_2(t) dt$$

Portanto, pela proposição, se X e Y são independentes e absolutamente contínuas, então $f_X * f_Y$ é densidade da soma $X+Y$.

Soma de duas variáveis aleatórias uniformes independentes

A convolução já produz um exemplo geométrico importante: a soma de duas uniformes não é uniforme.

Proposição 6.5 — Soma de duas variáveis aleatórias uniformes independentes

Se X e Y são variáveis aleatórias independentes, ambas uniformemente distribuídas em (0,1), então a densidade de $X+Y$ é

$$f_{X+Y}(a) = \begin{cases} a & 0 \leq a \leq 1 \\ 2-a & 1 < a < 2 \\ 0 & \text{caso contrário} \end{cases}$$

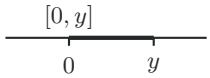
Demonstração. A densidade de $X+Y$ toma valores diferentes de 0 em (0,2). Da Equação 6.14, como

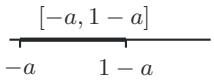
$$f_X(a) = f_Y(a) = \begin{cases} 1 & 0 < a < 1 \\ 0 & \text{caso contrário} \end{cases}$$

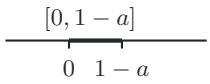
obtemos

$$f_{X+Y}(a) = \int_0^1 f_X(a-y)dy$$

O intervalo de integração depende de a:







Para $0 \leq a \leq 1$, temos que

$$f_{X+Y}(a) = \int_0^a dy = a.$$

Para $1 < a < 2$, obtemos

$$f_{X+Y}(a) = \int_{a-1}^1 dy = 2-a.$$

E assim

$$f_{X+Y}(a) = \begin{cases} a & 0 \leq a \leq 1 \\ 2-a & 1 < a < 2 \\ 0 & \text{caso contrário} \end{cases}$$

Devido ao formato característico de sua função densidade de probabilidade, a variável aleatória $U = X+Y$ é conhecida por ter uma distribuição triangular. Esta forma distintiva pode ser visualizada na Figura 6.7.

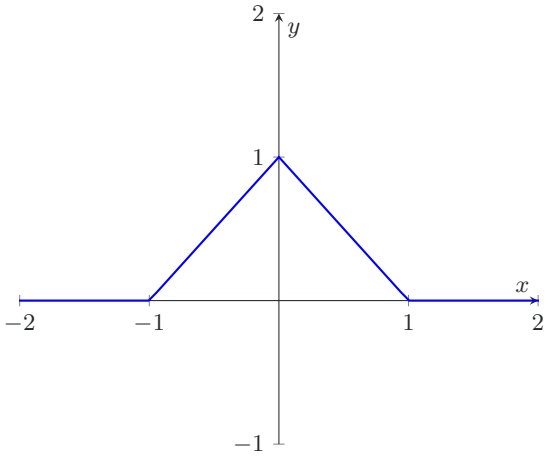


Figura 6.7: Soma de Uniformes

Variáveis aleatórias normais

Também podemos usar a Equação 6.14 para demonstrar o seguinte resultado importante sobre as variáveis aleatórias normais.

Proposição 6.6

Se $X_i, i = 1, \ldots, n$, são variáveis aleatórias independentes normalmente distribuídas com respectivos parâmetros $\mu_i, \sigma_i^2, i = 1, \ldots, n$, então $\sum_{i=1}^{n} X_i$ é normalmente distribuída com parâmetros $\sum_{i=1}^{n} \mu_i$ e $\sum_{i=1}^{n} \sigma_i^2$.

Demonstração. Para começar, suponha que X e Y sejam variáveis aleatórias normais independentes, onde X possui média 0 e variância σ^2, enquanto Y possui média 0 e variância 1. Nosso objetivo é determinar a densidade de $X + Y$ a partir da Equação 6.14.

Primeiramente, observemos que

$$\begin{aligned} f_X(a-y)\, f_Y(y) &= \frac{1}{\sqrt{2\pi}\,\sigma} \exp\Big(-\tfrac{(a-y)^2}{2\sigma^2}\Big)\ \frac{1}{\sqrt{2\pi}} \exp\Big(-\tfrac{y^2}{2}\Big) \\ &= \frac{1}{2\pi\,\sigma} \exp\Big(-\tfrac{a^2}{2\sigma^2}\Big)\ \exp\Big[-c\Big(y^2 - 2\,y\,\tfrac{a}{1+\sigma^2}\Big)\Big], \end{aligned}$$

onde

$$c \;=\; \frac{1}{2\,\sigma^2} + \frac{1}{2} \;=\; \frac{1+\sigma^2}{2\,\sigma^2}.$$

A partir da Equação 6.14 e completando quadrados, deduzimos que a densidade de $X + Y$ em a é:

$$\begin{aligned} f_{X+Y}(a) &= \frac{1}{2\pi\,\sigma}\ \exp\Big(-\tfrac{a^2}{2\sigma^2}\Big)\ \exp\Big(\tfrac{a^2}{2\,\sigma^2\,(1+\sigma^2)}\Big) \int_{-\infty}^{\infty} \exp\Big[-c\Big(y - \tfrac{a}{1+\sigma^2}\Big)^2\Big]\, dy \\ &= \frac{1}{2\pi\,\sigma}\ \exp\Big(-\tfrac{a^2}{2\,(1+\sigma^2)}\Big) \int_{-\infty}^{\infty} \exp(-c\,x^2)\, dx \\ &= C\, \exp\Big(-\tfrac{a^2}{2\,(1+\sigma^2)}\Big), \end{aligned}$$

onde C não depende de a. Conclui-se, portanto, que $X + Y$ é normal com média 0 e variância $1 + \sigma^2$.

Para passar ao caso geral, suponha que X_1 e X_2 sejam normais independentes com médias μ_1 e μ_2 e variâncias σ_1^2 e σ_2^2, respectivamente. Então,

$$X_1 + X_2 = \sigma_2\Big(\tfrac{X_1-\mu_1}{\sigma_2} + \tfrac{X_2-\mu_2}{\sigma_2}\Big) + (\mu_1 + \mu_2).$$

A variável $\frac{X_1-\mu_1}{\sigma_2}$ é normal com média 0 e variância $\frac{\sigma_1^2}{\sigma_2^2}$, enquanto $\frac{X_2-\mu_2}{\sigma_2}$ é normal com média 0 e variância 1. Com base no resultado anterior, a soma

$$\frac{X_1 - \mu_1}{\sigma_2} \;+\; \frac{X_2 - \mu_2}{\sigma_2}$$

é normal com média 0 e variância $1 + \frac{\sigma_1^2}{\sigma_2^2}$. Dessa forma, $X_1 + X_2$ é normal com média $\mu_1 + \mu_2$ e variância

$$\sigma_2^2\Big(1 + \tfrac{\sigma_1^2}{\sigma_2^2}\Big) = \sigma_1^2 + \sigma_2^2.$$

Assim, comprova-se a Proposição 6.6 para $n = 2$. O caso geral segue por indução.

■

Exemplo 6.19.

Um time de futebol jogará uma temporada com 40 partidas. Vinte e seis dessas partidas serão contra times da divisão A, e 14 contra times da divisão B. Suponha que o time ganhe cada partida disputada contra um adversário da divisão A com probabilidade 0,4, e ganhe cada partida disputada contra um adversário da divisão B com probabilidade 0,8. Suponha também que os resultados das diferentes partidas sejam independentes. Obtenha um valor aproximado para a probabilidade de que

1. o time vença 25 partidas ou mais;
2. o time vença mais partidas contra times da divisão A do que contra times da divisão B.

Solução. 1 Sejam X_A e X_B variáveis aleatórias que representam, respectivamente, o número de partidas que o time vence contra equipes da divisão A e B. X_A e X_B são binomiais independentes com

$$\begin{aligned} \mathbf{E}[X_A] &= 26(0{,}4) = 10{,}4 \quad \text{Var}(X_A) = 26(0{,}4)(0{,}6) = 6{,}24 \\ \mathbf{E}[X_B] &= 14(0{,}8) = 11{,}2 \quad \text{Var}(X_B) = 14(0{,}8)(0{,}2) = 2{,}24 \end{aligned}$$

Pela aproximação normal para a distribuição binomial, X_A e X_B têm aproximadamente a mesma distribuição que teriam variáveis aleatórias normais independentes com os valores esperados e as variâncias acima. Portanto, pela Proposição 6.6, $X_A + X_B$ terá aproximadamente uma distribuição normal com média $\mu = 21{,}6$ e variância $\sigma^2 = 8{,}48$. Assim, fazendo com que Z represente uma variável aleatória normal padrão, temos:[1]

$$\begin{aligned} \mathbf{P}(X_A + X_B \geq 25) &= \mathbf{P}(X_A + X_B \geq 24{,}5) \quad \text{(correção de continuidade)} \\ &= \mathbf{P}\left(\frac{X_A + X_B - \mu}{\sigma} \geq \frac{24{,}5 - \mu}{\sigma}\right) \\ &\approx \mathbf{P}\left(Z \geq \frac{24{,}5 - 21{,}6}{\sqrt{8{,}48}}\right) \\ &\approx \mathbf{P}(Z \geq 0{,}9947) \\ &\approx 1 - \Phi(0{,}9947) \\ &\approx 0{,}1600 \end{aligned}$$

onde Φ é a função de distribuição da normal padrão.

[1] Leia sobre a correção de continuidade na Seção 5.3

2 Notamos que $X_A - X_B$ tem aproximadamente uma distribuição normal com média $\mu = -0{,}8$ e variância $\sigma^2 = 8{,}48$. Com isso,

$$\begin{aligned}
\mathbf{P}(X_A - X_B \geq 1) &= \mathbf{P}(X_A - X_B \geq 0{,}5) \quad \text{(correção de continuidade)} \\
&= \mathbf{P}\left(\frac{X_A - X_B - \mu}{\sigma} \geq \frac{0{,}5 - \mu}{\sigma}\right) \\
&\approx \mathbf{P}\left(Z \geq \frac{0{,}5 + 0{,}8}{\sqrt{8{,}48}}\right) \\
&\approx \mathbf{P}(Z \geq 0{,}4472) \\
&\approx 1 - \Phi(0{,}4472) \\
&\approx 0{,}3274
\end{aligned}$$

Assim, há aproximadamente 16,00% de chances de que o time ganhe pelo menos 25 partidas, e aproximadamente 32,74% de chances de que o time ganhe mais partidas contra times da divisão A do que contra times da divisão B.

◁

Soma de variáveis aleatórias gama

Recordemos a definição da variável aleatória gama de parâmetros (t,λ), cuja função densidade de probabilidade tem a forma:

$$f(y) = \frac{\lambda e^{-\lambda y}(\lambda y)^{t-1}}{\Gamma(t)}, \quad 0 < y < \infty$$

onde $\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}dx$ é a função gama. Esta função é uma extensão do fatorial para os reais positivos; para um inteiro positivo n, temos $\Gamma(n+1) = n!$.

Proposição 6.7

Se X e Y são variáveis aleatórias gama independentes com parâmetros (s,λ) e (t,λ) respectivamente, então $X + Y$ é uma variável aleatória gama com parâmetros $(s+t,\lambda)$.

Demonstração. Utilizando a fórmula de convolução (Equação 6.14), obtemos:

$$\begin{aligned} f_{X+Y}(a) &= \frac{1}{\Gamma(s)\Gamma(t)} \int_0^a \lambda e^{-\lambda(a-y)} [\lambda(a-y)]^{s-1} \lambda e^{-\lambda y} (\lambda y)^{t-1} dy \\ &= K e^{-\lambda a} \int_0^a (a-y)^{s-1} y^{t-1} dy \\ &= K e^{-\lambda a} a^{s+t-1} \int_0^1 (1-x)^{s-1} x^{t-1} dx \quad \text{(fazendo } x = \frac{y}{a}\text{)} \\ &= C e^{-\lambda a} a^{s+t-1} \end{aligned}$$

onde K e C são constantes que não dependem de a.

Como $f_{X+Y}(a)$ é uma função densidade de probabilidade, sua integral em todo o domínio deve ser igual a 1, o que nos permite determinar o valor de C. Consequentemente, temos:

$$f_{X+Y}(a) = \frac{\lambda e^{-\lambda a} (\lambda a)^{s+t-1}}{\Gamma(s+t)}$$

Isto conclui a demonstração, mostrando que $X + Y$ segue uma distribuição gama com parâmetros $(s + t, \lambda)$. ■

Exemplo 6.20 — Duas lâmpadas em sequência.

Duas lâmpadas têm vidas independentes exponenciais com média 5 anos, isto é, taxa $1/5$. Se a segunda é instalada quando a primeira queima, o tempo total $T = X_1 + X_2$ satisfaz

$$T \sim \text{Gama}\left(2, \frac{1}{5}\right)$$

na parametrização forma–taxa adotada neste livro. Portanto,

$$\mathbf{P}(T \geq 15) = e^{-3}(1+3) = 4e^{-3} \approx 0{,}199.$$

◁

Pode-se demonstrar por indução, utilizando a Proposição 6.7, que para variáveis aleatórias gama X_i, $i = 1, \ldots, n$, com parâmetros respectivos (t_i, λ), $i = 1, \ldots, n$, a soma $\sum_{i=1}^n X_i$ segue uma distribuição gama com parâmetros $(\sum_{i=1}^n t_i, \lambda)$. A prova formal deste resultado fica como exercício para o leitor.

Exemplo 6.21.

Considere $X_1, X_2, \ldots, X_n$ como n variáveis aleatórias exponenciais independentes, cada uma com parâmetro λ. Sabendo que uma variável aleatória exponencial com parâmetro λ é equivalente a uma variável aleatória gama com parâmetros $(1, \lambda)$, podemos aplicar a Proposição 6.7. Consequentemente, $X_1 + X_2 + \cdots + X_n$ segue uma distribuição gama com parâmetros (n, λ).

◁

Somas de variáveis aleatórias de Poisson independentes

Proposição 6.8

Se X e Y são variáveis aleatórias de Poisson independentes com respectivos parâmetros λ_1 e λ_2, então $X+Y$ tem uma distribuição de Poisson com parâmetro $\lambda_1 + \lambda_2$.

Demonstração. Como o evento $[X + Y = n]$ pode ser expresso como a união dos eventos disjuntos $[X = k, Y = n - k]$, para $0 \leq k \leq n$, temos:

$$\begin{aligned}
\mathbf{P}(X + Y = n) &= \sum_{k=0}^{n} \mathbf{P}(X = k, Y = n - k) \\
&= \sum_{k=0}^{n} \mathbf{P}(X = k)\mathbf{P}(Y = n - k) \\
&= \sum_{k=0}^{n} e^{-\lambda_1} \frac{\lambda_1^k}{k!} e^{-\lambda_2} \frac{\lambda_2^{n-k}}{(n-k)!} \\
&= e^{-(\lambda_1+\lambda_2)} \sum_{k=0}^{n} \frac{\lambda_1^k \lambda_2^{n-k}}{k!(n-k)!} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{n!} \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} \lambda_1^k \lambda_2^{n-k} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{n!} (\lambda_1 + \lambda_2)^n \qquad \text{(binômio de Newton)}
\end{aligned}$$

Logo, $X + Y$ segue uma distribuição de Poisson com parâmetro $\lambda_1 + \lambda_2$. ■

Somas de variáveis aleatórias binomiais independentes

Proposição 6.9

Sejam X e Y variáveis aleatórias binomiais independentes com respectivos parâmetros (n, p) e (m, p). Então $X+Y$ é binomial com parâmetros $(n+m, p)$.

Demonstração. A partir da interpretação de uma variável aleatória binomial, e sem necessidade de cálculos, podemos concluir diretamente que $X + Y$ é binomial com parâmetros $(n + m, p)$. Isso se deve ao fato de que X representa o número de sucessos em n tentativas independentes, cada uma com probabilidade de sucesso p, e Y representa o número de sucessos em m tentativas independentes, também com probabilidade de sucesso p. Assim, pela independência de X e Y, $X + Y$ corresponde ao número de sucessos em um total de $n+m$ tentativas independentes, cada uma com probabilidade p de sucesso. Portanto, $X + Y$ é uma variável aleatória binomial com parâmetros $(n + m, p)$.

Para verificar essa conclusão analiticamente, note que

$$\begin{aligned}\mathbf{P}(X+Y=k) &= \sum_{i=0}^{n} \mathbf{P}(X=i, Y=k-i) \\ &= \sum_{i=0}^{n} \mathbf{P}(X=i)\mathbf{P}(Y=k-i) \\ &= \sum_{i=0}^{n} \binom{n}{i} p^i q^{n-i} \binom{m}{k-i} p^{k-i} q^{m-k+i}\end{aligned}$$

onde $q = 1 - p$ e adotamos $\binom{r}{j} = 0$ quando $j < 0$ ou $j > r$. Assim,

$$\mathbf{P}(X+Y=k) = p^k q^{n+m-k} \sum_{i=0}^{n} \binom{n}{i} \binom{m}{k-i}$$

e a conclusão segue da identidade combinatória:

$$\binom{n+m}{k} = \sum_{i=0}^{n} \binom{n}{i} \binom{m}{k-i}$$

■

6.6 Distribuição condicional

A distribuição conjunta guarda toda a informação sobre X e Y. Podemos usá-la para responder a uma pergunta que já apareceu no capítulo de probabilidade condicional: o que acontece com X quando descobrimos algo sobre Y? A distribuição condicional é a versão dessa atualização para variáveis aleatórias.

A ideia é sempre a mesma. No caso discreto, fixamos um evento $Y = y$ com probabilidade positiva e renormalizamos. No caso contínuo, $\mathbf{P}(Y = y) = 0$, portanto não dividimos por essa probabilidade. Condicionamos primeiro numa faixa $y < Y \leq y + h$ e deixamos $h \downarrow 0$. Sob as hipóteses usuais,

$$\mathbf{P}(X \in A \mid y < Y \leq y + h) \longrightarrow \int_A f_{X|Y}(x \mid y)\, dx.$$

É esse limite local que dá sentido ao condicionamento contínuo. As fórmulas finais ficam paralelas:

$$p_{X|Y}(x \mid y) = \frac{p_{XY}(x,y)}{p_Y(y)}, \qquad f_{X|Y}(x \mid y) = \frac{f_{XY}(x,y)}{f_Y(y)}.$$

Para cada valor admissível de y, elas definem uma distribuição em x.

Antes de fixar um valor, vale registrar a versão para uma faixa. Se $\mathbf{P}(y_1 < Y \leq y_2) > 0$, definimos

$$F_{X|\{y_1<Y\leq y_2\}}(x) := \mathbf{P}(X \leq x \mid y_1 < Y \leq y_2) \tag{6.15}$$

$$= \frac{F_{X,Y}(x,y_2) - F_{X,Y}(x,y_1)}{F_Y(y_2) - F_Y(y_1)}. \tag{6.16}$$

6.6.1 Variáveis aleatórias discretas

Se Y é discreta e $\mathbf{P}(Y = y) > 0$, a definição usual de probabilidade condicional pode ser aplicada diretamente. Para qualquer X,

$$F_{X|Y}(x \mid y) = \mathbf{P}(X \leq x \mid Y = y) = \frac{\mathbf{P}(X \leq x, Y = y)}{\mathbf{P}(Y = y)}.$$

A lei da probabilidade total recupera a distribuição marginal:

$$F_X(x) = \sum_y \mathbf{P}(Y = y) F_{X|Y}(x \mid y).$$

Se X também é discreta, podemos trabalhar diretamente com as massas pontuais.

Definição 6.5

Se X e Y são ambas variáveis aleatórias discretas

- A **função de probabilidade condicional** de X dado $Y = y$ é definida, para todo y com $p_Y(y) > 0$, como

$$\begin{aligned} p_{X|Y}(x \mid y) &:= \mathbf{P}(X = x \mid Y = y) \\ &= \frac{\mathbf{P}(X = x, Y = y)}{\mathbf{P}(Y = y)} \\ &= \frac{p(x, y)}{p_Y(y)} \end{aligned}$$

  para todos os valores de y tais que $p_Y(y) > 0$.

- a **função de distribuição condicional** de X dado $Y = y$ é definida, para todo y com $p_Y(y) > 0$, como

$$\begin{aligned} F_{X|Y}(x \mid y) &:= \mathbf{P}(X \leq x \mid Y = y) \\ &= \sum_{a \leq x} p_{X|Y}(a \mid y) \end{aligned} \quad .$$

Em outras palavras, essas definições são análogas ao caso incondicional de X, mas agora tudo é condicionado no evento $Y = y$. Se X for independente de Y, então as funções de probabilidade e de distribuição *condicionais* coincidem com as respectivas funções *incondicionais*. Isso ocorre pois, nesse caso,

$$\begin{aligned} p_{X|Y}(x \mid y) &= \mathbf{P}(X = x \mid Y = y) \\ &= \frac{\mathbf{P}(X = x, Y = y)}{\mathbf{P}(Y = y)} \\ &= \frac{\mathbf{P}(X = x)\mathbf{P}(Y = y)}{\mathbf{P}(Y = y)} \\ &= \mathbf{P}(X = x) \end{aligned}$$

Exemplo 6.22.

Se X e Y são variáveis aleatórias de Poisson independentes com respectivos parâmetros λ_1 e λ_2, calcule a distribuição condicional de X dado que $X + Y = n$.

Solução.

A função de probabilidade condicional de X dado $X+Y=n$ é

$$\begin{aligned} p_{X|X+Y}(k \mid n) &= \mathbf{P}(X=k \mid X+Y=n) \\ &= \frac{\mathbf{P}(X=k, X+Y=n)}{\mathbf{P}(X+Y=n)} \\ &= \frac{\mathbf{P}(X=k, Y=n-k)}{\mathbf{P}(X+Y=n)} \\ &= \frac{\mathbf{P}(X=k)\mathbf{P}(Y=n-k)}{\mathbf{P}(X+Y=n)} \end{aligned}$$

onde a última igualdade é consequência da hipótese de independência de X e Y. Na Proposição 6.8 mostramos que $X+Y$ tem uma distribuição de Poisson com parâmetro $\lambda_1+\lambda_2$, e assim temos

$$\begin{aligned} \mathbf{P}(X=k \mid X+Y=n) &= \frac{e^{-\lambda_1}\lambda_1^k}{k!}\frac{e^{-\lambda_2}\lambda_2^{n-k}}{(n-k)!}\left[\frac{e^{-(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^n}{n!}\right]^{-1} \\ &= \frac{n!}{(n-k)!k!}\frac{\lambda_1^k\lambda_2^{n-k}}{(\lambda_1+\lambda_2)^n} \\ &= \binom{n}{k}\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^k\left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{n-k} \end{aligned}$$

Logo a distribuição condicional de X dado que $X+Y=n$ é uma binomial com parâmetros n e $\lambda_1/(\lambda_1+\lambda_2)$.

◁

6.6.2 Variáveis aleatórias contínuas

Se Y é contínua, em geral $\mathbf{P}(Y=y)=0$, de modo que não podemos dividir por essa probabilidade. Em vez disso, condicionamos em uma faixa estreita $y<Y\le y+\Delta y$. Para $\Delta y>0$,

$$f_{X|\{y<Y\le y+\Delta y\}}(x) = \frac{\displaystyle\int_y^{y+\Delta y} f_{XY}(x,s)\,ds}{\displaystyle\int_y^{y+\Delta y} f_Y(s)\,ds}.$$

Quando $\Delta y \to 0$, numerador e denominador são, em primeira ordem, $f_{XY}(x,y)\Delta y$ e $f_Y(y)\Delta y$. Isso motiva a definição seguinte.

Definição 6.6

Para todo y tal que $f_Y(y) > 0$, a **densidade condicional** de X dado $Y = y$ é

$$f_{X|Y}(x \mid y) = \frac{f_{XY}(x,y)}{f_Y(y)}.$$

Para cada y admissível, essa expressão é de fato uma densidade em x, pois

$$\int_{-\infty}^{\infty} f_{X|Y}(x \mid y)\, dx = \frac{1}{f_Y(y)} \int_{-\infty}^{\infty} f_{XY}(x,y)\, dx = 1.$$

De modo análogo,

$$f_{Y|X}(y \mid x) = \frac{f_{XY}(x,y)}{f_X(x)}.$$

Se X e Y são independentes, $f_{XY}(x,y) = f_X(x) f_Y(y)$, e portanto

$$f_{X|Y}(x \mid y) = f_X(x), \qquad f_{Y|X}(y \mid x) = f_Y(y).$$

Ou seja, condicionar não altera a distribuição quando as variáveis são independentes.

A função de distribuição condicional é obtida integrando a densidade condicional:

Definição 6.7

A **função de distribuição condicional** de X dado $Y = y$ é

$$F_{X|Y}(x \mid y) = \int_{-\infty}^{x} f_{X|Y}(z \mid y)\, dz.$$

Como consequência, as seguintes fórmulas são válidas:

$$F_{X,Y}(x,y) = \int_{-\infty}^{y} f_Y(z) F_{X|Y}(x \mid z) dz$$

$$F_X(x) = \int_{-\infty}^{\infty} f_Y(y) F_{X|Y}(x \mid y) dy$$

Exemplo 6.23.

Dado

$$f_{XY}(x,y) = \begin{cases} k & 0 < x < y < 1 \\ 0 & \text{caso contrário} \end{cases}$$

determine $f_{X|Y}(x \mid y)$ e $f_{Y|X}(y \mid x)$.

Solução.

A densidade conjunta é constante na região triangular sombreada da Figura 6.8.

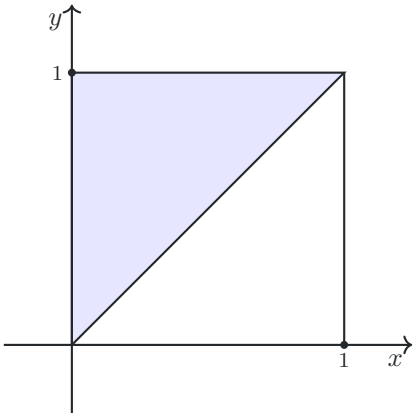


Figura 6.8: Região para o Exemplo 6.23

Podemos determinar k lembrando que a densidade deve integrar 1:

$$\iint f_{XY}(x,y)dxdy = \int_0^1 \int_0^y k dxdy = \int_0^1 kydy = \frac{k}{2} = 1 \Rightarrow k = 2$$

E agora podemos calcular as marginais:

$$f_X(x) = \int f_{XY}(x,y)dy = \int_x^1 2dy = 2(1-x) \quad 0 < x < 1$$

e

$$f_Y(y) = \int f_{XY}(x,y)dx = \int_0^y 2dx = 2y \quad 0 < y < 1$$

E assim

$$f_{X|Y}(x \mid y) = \frac{f_{XY}(x,y)}{f_Y(y)} = \frac{1}{y} \quad 0 < x < y < 1$$

e

$$f_{Y|X}(y \mid x) = \frac{f_{XY}(x,y)}{f_X(x)} = \frac{1}{1-x} \quad 0 < x < y < 1$$

◁

Exemplo 6.24.

As variáveis X e Y têm densidade dada por

$$f(x,y) = (x+y), \text{ se } 0 \leq x \leq 1{,}0 \leq y \leq 1$$

Calculemos a densidade condicional $f_{X|Y}$. Inicialmente, calculamos a marginal de Y. Temos

$$f_Y(y) = \int_0^1 (x+y)dx = y + \frac{1}{2}, 0 \leq y \leq 1$$

Então, para qualquer y fixado com $0 \leq y \leq 1$,

$$f_{X|Y}(x \mid y) = \frac{f(x,y)}{f_Y(y)} = \frac{x+y}{y+1/2} = \frac{2(x+y)}{2y+1}; \text{ para } 0 \leq x \leq 1$$

Fora desse intervalo, $f_{X|Y}(x \mid y)$ é zero.

A função de distribuição condicional pode ser obtida a partir da integração da densidade condicional. Assim, para qualquer y fixado com $0 \leq y \leq 1$ e $0 \leq x \leq 1$, vem:

$$\begin{aligned} F_{X|Y}(x \mid y) &= \int_{-\infty}^{x} \frac{2(z+y)}{2y+1}\, dz \\ &= \frac{x(2y+x)}{2y+1}. \end{aligned}$$

e, portanto,

$$F_{X|Y}(x \mid y) = \begin{cases} 0, & \text{se } x < 0, \\ \dfrac{x(2y+x)}{2y+1}, & \text{se } 0 \leq x < 1, \\ 1, & \text{se } x \geq 1. \end{cases}$$

A função de distribuição conjunta pode ser obtida a partir da condicional, da seguinte forma:

$$\begin{aligned} F_{X,Y}(x,y) &= \int_{-\infty}^{y} f_Y(z) F_{X|Y}(x \mid z)\, dz \\ &= \begin{cases} 0, & \text{se } x < 0 \text{ ou } y < 0, \\ \dfrac{x^2y + xy^2}{2}, & \text{se } 0 \leq x < 1,\ 0 \leq y < 1, \\ \dfrac{x + x^2}{2}, & \text{se } 0 \leq x < 1,\ y \geq 1, \\ \dfrac{y + y^2}{2}, & \text{se } x \geq 1,\ 0 \leq y < 1, \\ 1, & \text{se } x \geq 1,\ y \geq 1. \end{cases} \end{aligned}$$

Deixamos ao leitor a tarefa de verificar se essa expressão está correta, o que pode ser feito calculando seu valor diretamente pela densidade conjunta.

◁

6.6.3 Variáveis aleatórias mistas

Também é possível combinar uma variável contínua e outra discreta. Se X tem densidade f_X e N é discreta, escrevemos $p_{N|X}(n \mid x)$ para a probabilidade condicional de $N = n$ quando o valor de X é x. A forma mista da regra de Bayes é

$$f_{X|N}(x \mid n) = \frac{p_{N|X}(n \mid x) f_X(x)}{p_N(n)}, \qquad p_N(n) > 0.$$

Ela tem a mesma estrutura de sempre: *verossimilhança* $\times$ *densidade inicial, divididas pela constante de normalização.*

Exemplo 6.25 — Atualizando uma probabilidade de Bernoulli.

Suponha que realizemos $n+m$ tentativas de Bernoulli e observemos n sucessos e m fracassos. A probabilidade de sucesso é desconhecida. Para ilustrar o condicionamento misto, tratemos essa probabilidade como uma variável aleatória $\Theta \in (0,1)$ e adotemos, por simplicidade, uma priori uniforme:

$$f_\Theta(\theta) = 1, \qquad 0 < \theta < 1.$$

Essa é uma escolha de modelagem, não uma afirmação de ausência absoluta de informação.

Se N é o número de sucessos, então, dado $\Theta = \theta$,

$$N \mid (\Theta = \theta) \sim \mathrm{Bin}(n+m, \theta).$$

Logo, pela forma mista da regra de Bayes,

$$\begin{aligned} f_{\Theta|N}(\theta \mid n) &\propto p_{N|\Theta}(n \mid \theta) f_\Theta(\theta) \\ &\propto \theta^n (1-\theta)^m, \qquad 0 < \theta < 1. \end{aligned}$$

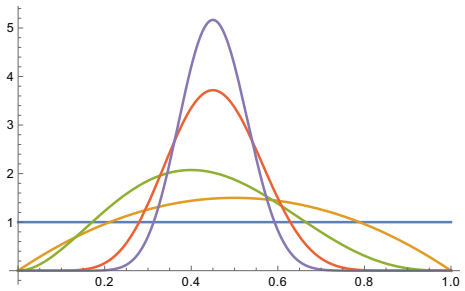
Normalizando,

$$f_{\Theta|N}(\theta \mid n) = \frac{\theta^n (1-\theta)^m}{B(n+1, m+1)},$$

portanto

$$\Theta \mid (N = n) \sim \mathrm{Beta}(n+1, m+1).$$

A conta mostra por que a família beta é natural para modelar probabilidades de sucesso: depois de observar dados de Bernoulli, permanecemos na mesma família.



Em geral, Beta(α,β) tem densidade

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, \qquad 0 < x < 1,$$

onde

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

◁

Exemplo 6.26 — Uma distribuição mista.

Considere $X \in \{1,2,3\}$ discreta e $Y \in [0,1]$ contínua. A lei conjunta é mista: para cada x, o valor $g(x,y)\,dy$ representa aproximadamente a probabilidade de $X = x$ e $Y \in [y, y+dy]$, onde

$$g(x,y) = \begin{cases} \dfrac{xy^{x-1}}{3}, & x \in \{1,2,3\},\ 0 \le y \le 1, \\ 0, & \text{caso contrário.} \end{cases}$$

Integrando em y, obtemos $p_X(x) = 1/3$ para $x = 1,2,3$. Somando em x,

$$f_Y(y) = \frac{1 + 2y + 3y^2}{3}, \qquad 0 \le y \le 1.$$

Se fixamos $X = 2$, a distribuição condicional de Y é contínua:

$$f_{Y|X}(y \mid 2) = \frac{g(2,y)}{p_X(2)} = 2y, \qquad 0 \le y \le 1.$$

Por outro lado, se fixamos $Y = 1/2$, a distribuição condicional de X é discreta. Como $f_Y(1/2) = 11/12$,

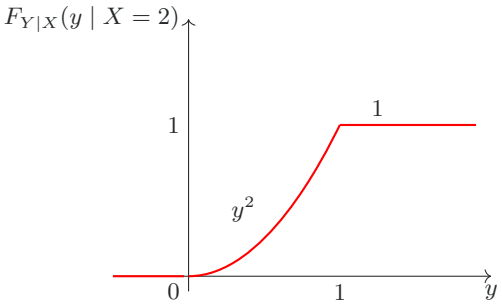
$$p_{X|Y}\left(x \mid \frac{1}{2}\right) = \frac{g(x,1/2)}{f_Y(1/2)} = \frac{4}{11} x \left(\frac{1}{2}\right)^{x-1}, \qquad x = 1,2,3.$$

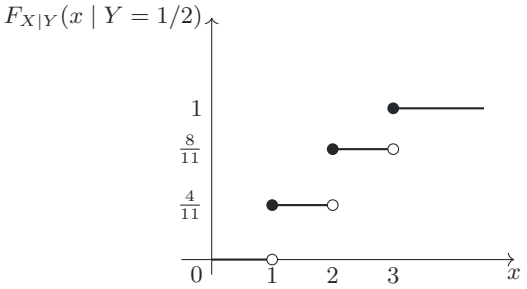
Assim,

$$F_{Y|X}(y \mid 2) = \begin{cases} 0, & y < 0, \\ y^2, & 0 \le y \le 1, \\ 1, & y > 1, \end{cases}$$

e

$$F_{X|Y}\left(x \mid \frac{1}{2}\right) = \begin{cases} 0, & x < 1, \\ \frac{4}{11}, & 1 \le x < 2, \\ \frac{8}{11}, & 2 \le x < 3, \\ 1, & x \ge 3. \end{cases}$$





Esse exemplo mostra por que, em modelos mistos, é importante distinguir massa condicional de densidade condicional.

◁

Distribuições conjuntas permitem responder três perguntas diferentes: como as coordenadas variam juntas, quando podem ser separadas por independência e como a informação sobre uma delas altera a distribuição da outra. Essas três ideias serão usadas repetidamente nos capítulos seguintes.

6.7 Exemplos integradores

Os modelos deste capítulo começam a ficar especialmente úteis quando o vetor aleatório não é apenas um par de números, mas registra a configuração inteira de um sistema. Os exemplos a seguir mostram três maneiras de explorar essa ideia: representar um objeto combinatório por indicadores, transformar contagens multinomiais em uma quantidade operacional e reconhecer dependência entre coordenadas que têm marginais simples.

Exemplo 6.27 — Um grafo aleatório como vetor de Bernoulli.

Considere o grafo aleatório de Erdős–Rényi $G(n,p)$. Para cada par $e = \{i,j\}$ de vértices, seja I_e a indicadora do evento "a aresta e está presente". O vetor

$$(I_e)_{e \in \binom{[n]}{2}}$$

é um vetor aleatório discreto com $\binom{n}{2}$ coordenadas. Como as arestas são escolhidas independentemente,

$$\mathbf{P}(I_e = i_e,\ e \in \binom{[n]}{2}) = \prod_{e} p^{i_e}(1-p)^{1-i_e}, \qquad i_e \in \{0,1\}.$$

Um objeto combinatório inteiro passa, assim, a ser uma única observação vetorial. Graus, número de arestas e número de triângulos são funções desse vetor; voltaremos a elas nos capítulos de transformações e esperança.

◁

Exemplo 6.28 — Balanceamento de carga e o problema de ocupação.

Suponha que n requisições sejam enviadas, independentemente e com a mesma probabilidade, para m servidores. Se N_j é o número de requisições que chegam ao servidor j, então

$$(N_1, \dots, N_m)$$

tem distribuição multinomial com parâmetros n e $p_1 = \cdots = p_m = 1/m$.

Uma pergunta natural é quantos servidores serão efetivamente utilizados. Defina

$$I_j = \mathbb{1}_{\{N_j > 0\}}, \qquad A = \sum_{j=1}^{m} I_j.$$

Como um servidor fica vazio somente quando nenhuma das n requisições o escolhe,

$$\mathbf{P}(I_j = 1) = 1 - \left(1 - \frac{1}{m}\right)^n.$$

Logo,

$$\mathbf{E}[A] = m\left[1 - \left(1 - \frac{1}{m}\right)^{n}\right].$$

A mesma conta descreve bolas lançadas em caixas, chaves distribuídas numa tabela de *hash* ou usuários distribuídos entre máquinas. O vetor multinomial guarda a carga completa; a função A extrai uma característica operacional do sistema.

◁

Exemplo 6.29 — Graus de vértices vizinhos.

No modelo $G(n,p)$, sejam D_u e D_v os graus de dois vértices distintos. Cada um tem distribuição

$$\mathrm{Bin}(n-1,p),$$

mas eles não são independentes, pois ambos contêm a indicadora $I_{\{u,v\}}$ da aresta comum. Escrevendo

$$D_u = I_{\{u,v\}} + \sum_{w \neq u,v} I_{\{u,w\}}, \qquad D_v = I_{\{u,v\}} + \sum_{w \neq u,v} I_{\{v,w\}},$$

todas as parcelas distintas são independentes. Portanto,

$$\mathrm{Cov}(D_u,D_v) = \mathrm{Var}(I_{\{u,v\}}) = p(1-p).$$

As marginais dos dois graus são iguais e muito simples, mas a distribuição conjunta registra a aresta que compartilham. É exatamente esse tipo de informação que se perde quando olhamos apenas para as marginais.

◁

Ao atravessar este capítulo

Distribuições conjuntas descrevem dependência. Marginais são obtidas somando ou integrando; condicionais renormalizam uma fatia; independência equivale à fatoração adequada. Vetores de indicadores também permitem representar objetos complexos, como grafos e sistemas de ocupação, por coordenadas elementares.

6.8 Exercícios

Exercício 6.1 — Tabela conjunta

A função de probabilidade conjunta de $X,Y \in \{0,1,2\}$ é proporcional a $x+y+1$. Determine a constante, as marginais, $\mathbf{P}(X < Y)$ e $\mathbf{P}(X = 1 \mid Y = 2)$. Decida se X e Y são independentes.

Exercício 6.2 — Região triangular

Se $f_{XY}(x,y) = c$ na região $0 < y < x < 3$, determine c, as densidades marginais e $\mathbf{P}(X + Y < 3)$.

Exercício 6.3 — Condicional contínua

Se $f_{XY}(x,y) = 2$, para $0 < y < x < 1$, obtenha $f_Y(y)$ e $f_{X|Y}(x \mid y)$. Calcule $\mathbf{P}(X > 3/4 \mid Y = 1/2)$.

Exercício 6.4 — Máximo e mínimo

Se X e Y são independentes e uniformes em $(0,1)$, determine as funções de distribuição de $M = \max(X,Y)$ e $N = \min(X,Y)$. Calcule $\mathbf{P}(M - N < 1/3)$.

Exercício 6.5 — Soma de Poisson

Se $X \sim \text{Poisson}(2{,}5)$ e $Y \sim \text{Poisson}(1{,}7)$ são independentes, obtenha a distribuição de $X + Y$ e calcule $\mathbf{P}(X + Y \leq 2)$.

CAPÍTULO

7

Funções de variáveis aleatórias

Nos capítulos anteriores transformamos uma variável de cada vez. Para uma função simples, a CDF resolve quase tudo. Mas o que muda quando transformamos um vetor inteiro? E se, em vez de combinar coordenadas, quisermos apenas ordená-las?

Essas duas perguntas conduzem às ideias do capítulo. Na mudança suave de coordenadas, precisamos corrigir a deformação de áreas: surge o jacobiano. Na ordenação, exploramos a estrutura especial de mínimos, máximos e posições numa amostra. São problemas diferentes, mas ambos começam com a mesma pergunta: como a distribuição muda quando aplicamos uma função ao vetor aleatório?

7.1 Funções de variáveis aleatórias contínuas

Dadas duas variáveis aleatórias X e Y, considere

$$Z = g(X, Y).$$

Se f_{XY} é conhecida, a função de distribuição de Z é obtida integrando

sobre a região em que $g(x,y) \leq z$:

$$F_Z(z) = \mathbf{P}(Z \leq z) = \mathbf{P}(g(X,Y) \leq z) = \mathbf{P}((X,Y) \in D_z)$$
$$= \iint_{(x,y)\in D_z} f_{XY}(x,y)\,dx\,dy,$$

onde $D_z = \{(x,y) : g(x,y) \leq z\}$. Depois de determinar F_Z, derivamos para obter f_Z, quando a derivada existe. O suporte merece atenção especial: ele determina os limites de integração e, muitas vezes, obriga a separar o cálculo em casos. O primeiro exemplo mostra o método em uma transformação unidimensional.

Exemplo 7.1.

Se $X \sim U[0,1]$, qual é a distribuição de $Y = -\log(X)$?

Solução. Como

$$0 < Y < \infty \Leftrightarrow 0 < X < 1,$$

e $\mathbf{P}(0 < X < 1) = 1$, temos que $F_Y(y) = 0$ para $y \leq 0$. Se $y > 0$, então

$$\mathbf{P}(Y \leq y) = \mathbf{P}(-\log(X) \leq y) = \mathbf{P}\left(X \geq e^{-y}\right) = 1 - e^{-y}.$$

Portanto, $Y \sim \exp(1)$, ou seja, Y tem uma distribuição exponencial com parâmetro 1.

◁

Exemplo 7.2.

Dado $Z = X - Y$, determine $f_Z(z)$.

Solução. Utilizando a expressão derivada anteriormente e a Figura 7.1, temos

$$F_Z(z) = \mathbf{P}(X - Y \leq z) = \int_{y=-\infty}^{\infty} \int_{x=-\infty}^{z+y} f_{XY}(x,y)\,dx\,dy,$$

onde a região de integração $x - y \leq z$ é ilustrada na Figura 7.1.

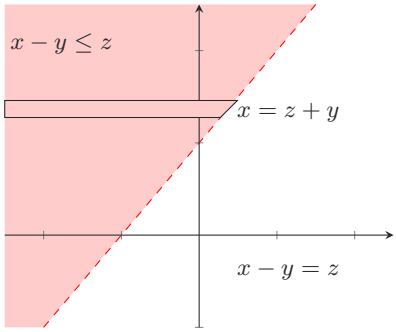


Figura 7.1: Região de integração para $x - y \leq z$

Assim, obtemos

$$f_Z(z) = \frac{dF_Z(z)}{dz} = \int_{-\infty}^{\infty} f_{XY}(z+y, y)\, dy. \tag{7.1}$$

Quando as variáveis X e Y são independentes, a Equação 7.1 se reduz a

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z+y) f_Y(y)\, dy = (f_X * f_{-Y})(z),$$

em que $f_{-Y}(u) = f_Y(-u)$. Portanto, a densidade de uma diferença é a convolução da densidade de X com a densidade de $-Y$.

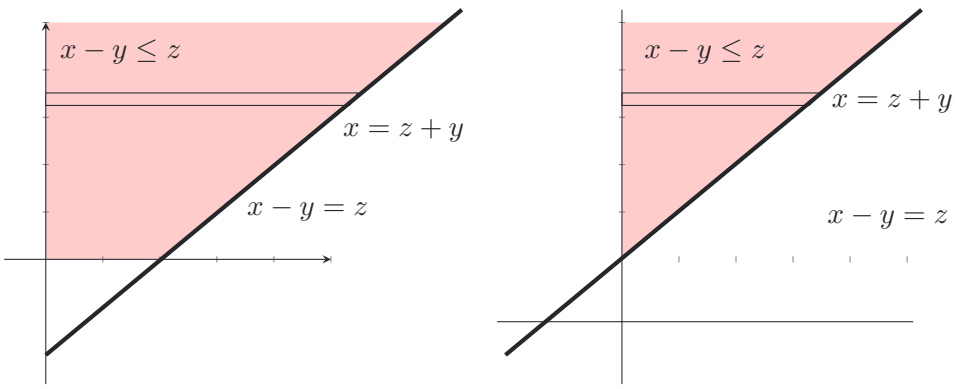
Como um caso especial, suponha que

$$f_X(x) = 0 \quad \text{para } x < 0, \quad f_Y(y) = 0 \quad \text{para } y < 0.$$

Nesse caso, z pode ser positivo ou negativo, o que requer uma análise separada para $z \geq 0$ e $z < 0$, pois as regiões de integração para esses casos diferem significativamente. Para $z \geq 0$, conforme mostrado na Figura 7.3a, temos

$$F_Z(z) = \int_{y=0}^{\infty} \int_{x=0}^{z+y} f_{XY}(x, y)\, dx\, dy.$$

◁



(a) (b)

Figura 7.3: .

e para $z < 0$, da Fig. 7.3b

$$F_Z(z) = \int_{y=-z}^{\infty} \int_{x=0}^{z+y} f_{XY}(x, y) dx dy$$

Após a diferenciação, temos

$$f_Z(z) = \begin{cases} \int_0^\infty f_{XY}(z+y, y)dy & z \geq 0 \\ \int_{-z}^\infty f_{XY}(z+y, y)dy & z < 0 \end{cases}$$

Para antecipar o exemplo do quociente, se X e Y são não negativas, a região $\{X/Y \leq z\}$, para $z \geq 0$, é a mostrada na Figura 7.4. Nesse caso,

$$F_Z(z) = \int_{y=0}^{\infty} \int_{x=0}^{yz} f_{XY}(x, y)dxdy$$

ou

$$f_Z(z) = \int_{y=0}^{\infty} y f_{XY}(yz, y)dy$$

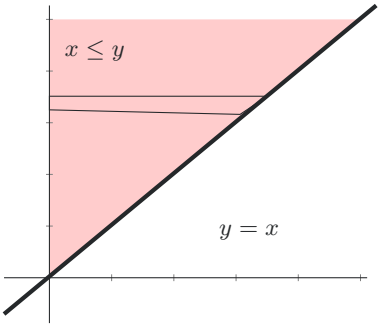


Figura 7.4

Exemplo 7.3.

Seja $Z = \frac{X}{Y}$ e suponha que $\mathbf{P}(Y = 0) = 0$. Vamos determinar a função de densidade $f_Z(z)$.

Solução. Iniciaremos pela distribuição acumulada:

$$F_Z(z) = \mathbf{P}\left(\frac{X}{Y} \leq z\right).$$

A desigualdade $\frac{X}{Y} \leq z$ pode ser reescrita como $X \leq Yz$ quando $Y > 0$, e $X \geq Yz$ quando $Y < 0$. Particionamos, então, o cálculo nos eventos $\{Y > 0\}$ e $\{Y < 0\}$; a hipótese $\mathbf{P}(Y = 0) = 0$ assegura que nenhum caso com probabilidade positiva foi omitido. Assim,

$$\begin{aligned} \mathbf{P}\left(\frac{X}{Y} \leq z\right) &= \mathbf{P}\left(\left\{\frac{X}{Y} \leq z\right\} \cap \{Y \neq 0\}\right) \\ &= \mathbf{P}\left(\frac{X}{Y} \leq z, Y > 0\right) + \mathbf{P}\left(\frac{X}{Y} \leq z, Y < 0\right) \\ &= \mathbf{P}\left(X \leq Yz, Y > 0\right) + \mathbf{P}\left(X \geq Yz, Y < 0\right). \end{aligned}$$

Podemos expressar as probabilidades anteriores como integrais sobre duas regiões, obtendo:

$$F_Z(z) = \int_{y=0}^{\infty} \int_{x=-\infty}^{yz} f_{XY}(x,y)\,dx\,dy + \int_{y=-\infty}^{0} \int_{x=yz}^{\infty} f_{XY}(x,y)\,dx\,dy. \quad (7.2)$$

Finalmente derivando a Equação 7.2, obtemos:

$$\begin{aligned} f_Z(z) &= \int_0^{\infty} y f_{XY}(yz,y)\,dy + \int_{-\infty}^{0} -y f_{XY}(yz,y)\,dy \\ &= \int_{-\infty}^{\infty} |y| f_{XY}(yz,y)\,dy. \end{aligned}$$

Observe que, se X e Y são variáveis aleatórias não negativas, a expressão anterior se reduz a:

$$F_Z(z) = \int_{y=0}^{\infty} \int_{x=0}^{yz} f_{XY}(x,y)\,dx\,dy \qquad \text{e}$$

$$f_Z(z) = \int_{y=0}^{\infty} y f_{XY}(yz,y)\,dy.$$

◁

Exemplo 7.4.

Sejam X e Y variáveis aleatórias conjuntamente normais com média zero, cuja função de densidade conjunta é dada por:

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} e^{-\left[\frac{1}{2(1-r^2)}\left(\frac{x^2}{\sigma_1^2} - \frac{2rxy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2}\right)\right]}.$$

Mostre que o quociente $Z = \frac{X}{Y}$ tem uma densidade de Cauchy centrada em $\frac{r\sigma_1}{\sigma_2}$.

Solução. Utilizando o fato de que $f_{XY}(-x,-y) = f_{XY}(x,y)$, obtemos:

$$f_Z(z) = \frac{2}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} \int_0^{\infty} y e^{-y^2/2\sigma_0^2}\,dy = \frac{\sigma_0^2}{\pi\sigma_1\sigma_2\sqrt{1-r^2}}.$$

onde

$$\sigma_0^2 = \frac{1-r^2}{\left(\frac{z^2}{\sigma_1^2}\right) - \left(\frac{2rz}{\sigma_1\sigma_2}\right) + \left(\frac{1}{\sigma_2^2}\right)}.$$

Assim, temos:

$$f_Z(z) = \frac{\sigma_1\sigma_2\sqrt{1-r^2}/\pi}{\sigma_2^2\left(z - \frac{r\sigma_1}{\sigma_2}\right)^2 + \sigma_1^2(1-r^2)}.$$

Essa é a densidade de Cauchy com parâmetro de localização $r\sigma_1/\sigma_2$ e parâmetro de escala $\sigma_1\sqrt{1-r^2}/\sigma_2$. Integrando de $-\infty$ até z, obtemos a função de distribuição correspondente:

$$F_Z(z) = \frac{1}{2} + \frac{1}{\pi}\arctan\left(\frac{\sigma_2 z - r\sigma_1}{\sigma_1\sqrt{1-r^2}}\right).$$

◁

Exemplo 7.5.

Sejam X e Y variáveis aleatórias independentes, com $X \sim \text{Gama}(m,\beta)$ e $Y \sim \text{Gama}(n,\beta)$, na parametrização forma–taxa. Mostre que $Z = \frac{X}{X+Y}$ segue uma distribuição beta.

Solução. A função de densidade conjunta de X e Y é dada por:

$$f_{XY}(x,y) = f_X(x)f_Y(y) = \frac{\beta^{m+n}}{\Gamma(m)\Gamma(n)}x^{m-1}y^{n-1}e^{-\beta(x+y)}, \qquad x>0,\ y>0.$$

Note que $0 < Z < 1$, uma vez que X e Y são variáveis aleatórias não negativas. Assim, a função de distribuição acumulada $F_Z(z)$ pode ser expressa como:

$$\begin{aligned} F_Z(z) &= P[Z \le z] = \mathbf{P}\left(\frac{X}{X+Y} \le z\right) = \mathbf{P}\left(X \le Y\frac{z}{1-z}\right) \\ &= \int_0^\infty \int_0^{\frac{yz}{1-z}} f_{XY}(x,y)\,dx\,dy, \end{aligned}$$

onde utilizamos a representação gráfica mostrada na Figura 7.4.

Diferenciando em relação a z, obtemos:

$$\begin{aligned} f_Z(z) &= \int_0^\infty \frac{y}{(1-z)^2} f_{XY}\left(\frac{yz}{1-z}, y\right) dy \\ &= \frac{\beta^{m+n}}{\Gamma(m)\Gamma(n)}\frac{z^{m-1}}{(1-z)^{m+1}}\int_0^\infty y^{m+n-1}e^{-\beta y/(1-z)}\,dy \\ &= \frac{z^{m-1}(1-z)^{n-1}}{\Gamma(m)\Gamma(n)}\int_0^\infty u^{m+n-1}e^{-u}\,du = \frac{\Gamma(m+n)}{\Gamma(m)\Gamma(n)}z^{m-1}(1-z)^{n-1} \\ &= \begin{cases} \frac{1}{\beta(m,n)}z^{m-1}(1-z)^{n-1} & \text{para } 0<z<1, \\ 0 & \text{caso contrário.} \end{cases} \end{aligned}$$

Essa expressão representa uma distribuição beta.

◁

Exemplo 7.6.
Seja $Z = X^2 + Y^2$. Determine $f_Z(z)$.
Solução. Temos que

$$F_Z(z) = \mathbf{P}(X^2 + Y^2 \leq z) = \iint_{x^2+y^2\leq z} f_{XY}(x, y)\, dx\, dy$$

Mas, $x^2 + y^2 \leq z$ representa a área de um círculo com raio $\sqrt{z}$ e, portanto,

$$F_Z(z) = \int_{-\sqrt{z}}^{\sqrt{z}} \int_{-\sqrt{z-y^2}}^{\sqrt{z-y^2}} f_{XY}(x, y)\, dx\, dy,$$

para $z > 0$; para $z \leq 0$, $F_Z(z) = 0$.

Logo

$$f_Z(z) = \int_{-\sqrt{z}}^{\sqrt{z}} \frac{1}{2\sqrt{z-y^2}} \left(f_{XY}\left(\sqrt{z-y^2}, y\right) + f_{XY}\left(-\sqrt{z-y^2}, y\right)\right) dy \tag{7.3}$$

◁

Exemplo 7.7.
X e Y são variáveis aleatórias normais independentes com média zero e variância comum σ^2. Determine $f_Z(z)$ para $Z = X^2 + Y^2$.
Solução. Usando 7.3, para $z > 0$ obtemos

$$\begin{aligned} f_Z(z) &= \frac{e^{-z/(2\sigma^2)}}{2\pi\sigma^2} \int_{-\sqrt{z}}^{\sqrt{z}} \frac{dy}{\sqrt{z-y^2}} \\ &= \frac{e^{-z/(2\sigma^2)}}{\pi\sigma^2} \int_0^{\pi/2} d\theta = \frac{1}{2\sigma^2} e^{-z/(2\sigma^2)}. \end{aligned}$$

onde usamos a substituição $y = \sqrt{z}\sin\theta$. Assim, $X^2 + Y^2$ tem distribuição exponencial com taxa $1/(2\sigma^2)$ e média $2\sigma^2$.

◁

7.2 Funções de variáveis contínuas: método do Jacobiano

O método da função de distribuição continua válido, mas pode obrigar a redesenhar a região de integração para cada valor. Há uma maneira de fazer a

geometria de uma vez só? Quando a transformação é diferenciável e invertível, o teorema de mudança de variáveis faz exatamente isso. O jacobiano registra quanto pequenas áreas são dilatadas ou contraídas. Se $h = g^{-1}$, um pequeno retângulo de área $dz\,dw$ corresponde, localmente, a uma área aproximada

$$|J(z,w)|\,dz\,dw$$

no plano original. Esse é o fator que preserva a massa de probabilidade.

Sejam $A_0, A \subset \mathbb{R}^2$ regiões abertas e $g : A_0 \to A$ uma bijeção. Escreveremos

$$(z, w) := g(x, y) = (g_1(x, y), g_2(x, y))\,. \tag{7.4}$$

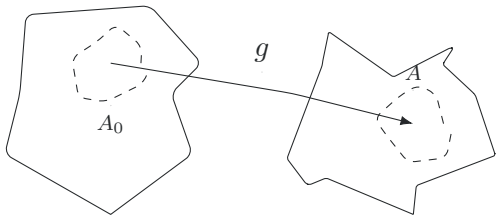


Figura 7.5: Representação de transformação entre regiões A_0 e A via g.

Como g é uma bijeção, existe a inversa $h = g^{-1} : A \to A_0$, dada por

$$x = h_1(z, w), \qquad y = h_2(z, w).$$

Suponha que as derivadas parciais

$$\frac{\partial x}{\partial z}, \quad \frac{\partial x}{\partial w}, \quad \frac{\partial y}{\partial z}, \quad \frac{\partial y}{\partial w}$$

existem e são contínuas em A.

Definimos o Jacobiano da transformação inversa h como o determinante

$$J(z, w) = \begin{vmatrix} \frac{\partial x}{\partial z} & \frac{\partial x}{\partial w} \\ \frac{\partial y}{\partial z} & \frac{\partial y}{\partial w} \end{vmatrix} = \det \begin{pmatrix} \frac{\partial x}{\partial z} & \frac{\partial x}{\partial w} \\ \frac{\partial y}{\partial z} & \frac{\partial y}{\partial w} \end{pmatrix}.$$

Teorema 7.1 — Mudança de variáveis em integrais duplas

Sejam $g : A_0 \to A$ uma bijeção diferenciável e $h = g^{-1}$, ambas com derivadas parciais contínuas e determinantes jacobianos não nulos. Para toda região $B \subset A$,

$$\iint_{h(B)} f(x,y)\,dx\,dy = \iint_B f(h_1(z,w),h_2(z,w))\ |J(z,w)|\ dz\,dw.$$

Podemos aplicar o Teorema de Mudança de Variáveis para densidades. Seja $f = f_{X,Y}$ a densidade conjunta das variáveis aleatórias X e Y, e assumamos que

$$\mathbf{P}\left((X,Y) \in A_0\right) = 1.$$

Sejam Z e W as variáveis transformadas, ou seja, $Z = g_1(X,Y)$ e $W = g_2(X,Y)$. Então, para todo evento $B \subset A$, temos

$$\begin{aligned}
\mathbf{P}((Z,W) \in B) &= \mathbf{P}((X,Y) \in h(B)) && (7.5)\\
&= \iint_{h(B)} f_{X,Y}(x,y)\,dx\,dy && (7.6)\\
&= \iint_B f_{X,Y}(h_1(z,w),h_2(z,w))\ |J(z,w)|\ dz\,dw. && (7.7)
\end{aligned}$$

Assim, pela definição de densidade, a função obtida é a densidade de (Z,W), e obtemos o seguinte teorema:

Teorema 7.2

Sob as condições dadas, a densidade conjunta de Z e W é:

$$f_{Z,W}(z,w) = \begin{cases} f_{X,Y}(h_1(z,w),h_2(z,w))\,|J(z,w)|\,, & \text{se } (z,w) \in A,\\ 0, & \text{caso contrário.}\end{cases}$$

O teorema conduz a um roteiro de cálculo: determinar o suporte transformado A, inverter as equações para escrever x e y em função de z,w, calcular o módulo do jacobiano da inversa e multiplicá-lo pela densidade original avaliada no ponto inverso. O suporte vem primeiro, pois uma fórmula algébrica correta fora de A ainda não define a densidade correta.

A principal dificuldade na aplicação do método do jacobiano está em determinar o conjunto A.

Exemplo 7.8 — De um uniforme a um tempo de espera.
Um gerador de números aleatórios costuma começar por uma variável $U \sim \text{Unif}(0,1)$. Como transformar esse número em um tempo de espera exponencial de taxa λ? A transformação

$$T = -\frac{1}{\lambda}\log U$$

faz exatamente isso.

Solução. A transformação leva $(0,1)$ em $(0,\infty)$ e sua inversa é $u = e^{-\lambda t}$. Portanto,

$$\left|\frac{du}{dt}\right| = \lambda e^{-\lambda t}.$$

Como a densidade de U vale 1 em $(0,1)$, obtemos

$$f_T(t) = \begin{cases} \lambda e^{-\lambda t}, & t > 0, \\ 0, & t \le 0. \end{cases}$$

Logo $T \sim \text{Exp}(\lambda)$. O cálculo é unidimensional, mas já contém a ideia central do método: ao mudar de coordenadas, a densidade precisa compensar a deformação produzida pela transformação.

◁

Em muitos exemplos, é mais fácil calcular o jacobiano da transformação direta. Como as matrizes derivadas de funções inversas são inversas uma da outra,

$$|J(z,w)| = \frac{1}{|J(x,y)|}$$

sendo

$$J(x,y) = \begin{vmatrix} \frac{\partial z}{\partial x} & \frac{\partial z}{\partial y} \\ \frac{\partial w}{\partial x} & \frac{\partial w}{\partial y} \end{vmatrix} \tag{7.8}$$

pois o determinante jacobiano da transformação direta costuma ser mais fácil de calcular. A identidade vale nos pontos em que ambos os determinantes são não nulos.

Nessa notação, a densidade conjunta pode ser expressa como

$$f_{ZW}(z,w) = |J(z,w)| f_{XY}\big(h_1(z,w), h_2(z,w)\big) = \frac{f_{XY}\big(h_1(z,w), h_2(z,w)\big)}{|J(x,y)|} \tag{7.9}$$

Exemplo 7.9 — Carga total e desequilíbrio.

Dois servidores recebem cargas normalizadas X e Y, modeladas como variáveis independentes e uniformes em (0,1). Em vez das duas cargas separadas, queremos acompanhar

$$U = X + Y, \qquad V = X - Y,$$

onde U mede a carga total e V o desequilíbrio entre os servidores. Quais pares (U,V) podem ocorrer e como se distribuem?

Solução. A transformação é linear e pode ser invertida imediatamente:

$$x = \frac{u+v}{2}, \qquad y = \frac{u-v}{2}.$$

Seu jacobiano direto é

$$\det \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = -2,$$

de modo que o módulo do jacobiano da inversa vale $1/2$. Como $f_{X,Y}(x,y) = 1$ no quadrado unitário, temos

$$f_{U,V}(u,v) = \frac{1}{2}$$

sempre que o ponto inverso pertence a esse quadrado.

A geometria torna o suporte transparente: a transformação leva os quatro vértices do quadrado (0,0),(0,1),(1,1),(1,0) nos pontos (0,0),(1,−1),(2,0),(1,1). Assim,

$$A = \{(u,v) : 0 < u < 2,\ \max(-u, u-2) < v < \min(u, 2-u)\},$$

e

$$f_{U,V}(u,v) = \begin{cases} \frac{1}{2}, & (u,v) \in A, \\ 0, & \text{caso contrário.} \end{cases}$$

O losango também revela algo que uma conta marginal esconderia: carga total e desequilíbrio não são independentes. Valores extremos de U restringem fortemente os valores possíveis de V.

◁

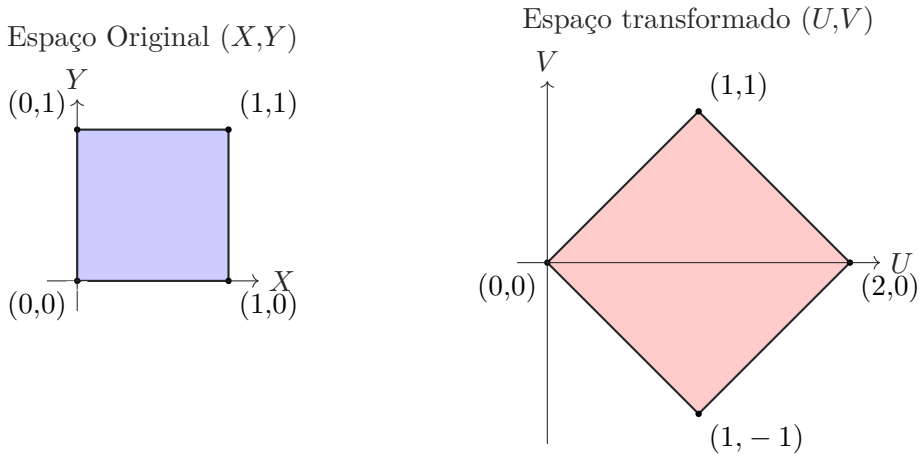


Figura 7.6: Cargas individuais e coordenadas de carga total e desequilíbrio.

Exemplo 7.10 — Tempo total e diferença entre etapas.

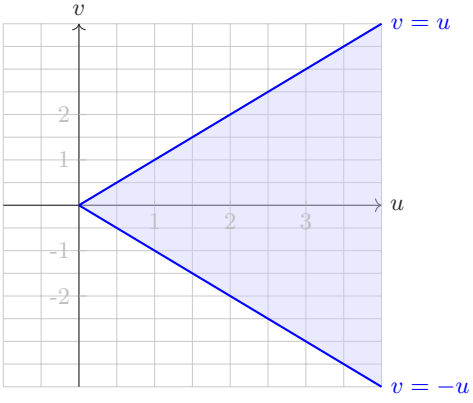
Duas etapas de um serviço têm durações independentes X e Y, ambas exponenciais de taxa λ. O tempo total é $U = X + Y$, enquanto

$$V = X - Y$$

registra qual etapa demorou mais e por quanto. Determine a densidade conjunta de (U,V) e as distribuições dessas duas quantidades.

Solução. A densidade original é

$$f_{X,Y}(x,y) = \lambda^2 e^{-\lambda(x+y)}, \qquad x > 0,\ y > 0.$$



A inversa é

$$x = \frac{u+v}{2}, \qquad y = \frac{u-v}{2},$$

e as condições $x > 0$, $y > 0$ equivalem a $u > 0$ e $|v| < u$. Como o módulo do jacobiano da inversa vale $1/2$,

$$f_{U,V}(u,v) = \frac{\lambda^2}{2} e^{-\lambda u}, \qquad u > 0,\ |v| < u.$$

Integrando em v,

$$f_U(u) = \lambda^2 u e^{-\lambda u}, \qquad u > 0,$$

logo $U \sim \text{Gama}(2,\lambda)$. Para a diferença,

$$f_V(v) = \int_{|v|}^{\infty} \frac{\lambda^2}{2} e^{-\lambda u}\, du = \frac{\lambda}{2} e^{-\lambda |v|}, \qquad v \in \mathbb{R}.$$

Portanto, a diferença entre duas exponenciais independentes de mesma taxa tem uma distribuição de Laplace centrada em zero. Apesar das formas simples das duas marginais, U e V não são independentes: o próprio suporte $|v| < u$ acopla total e diferença.

◁

Exemplo 7.11 — Tempo total e fração gasta na primeira tarefa.

Duas tarefas têm tempos de execução independentes X e Y, ambos exponenciais com média μ. Para planejar o trabalho, duas quantidades são mais informativas do que os tempos separados:

$$U = X + Y, \qquad V = \frac{X}{X+Y}.$$

Aqui U é o tempo total e V é a fração desse tempo gasta na primeira tarefa. Determine a densidade conjunta de (U,V).

Solução. A transformação é invertida por

$$X = UV, \qquad Y = U(1-V).$$

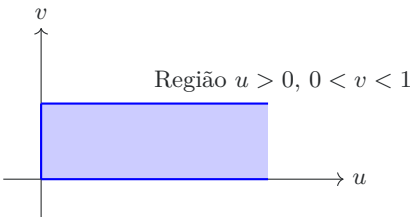
O módulo do jacobiano da inversa é u, pois

$$\left| \det \begin{pmatrix} v & u \\ 1-v & -u \end{pmatrix} \right| = u$$

no suporte, onde $u > 0$. Assim,

$$f_{U,V}(u,v) = u\, f_{X,Y}(uv, u(1-v)) \tag{7.10}$$

$$= \frac{u}{\mu^2} e^{-u/\mu}, \qquad u > 0,\ 0 < v < 1. \tag{7.11}$$



A densidade fatoriza:

$$f_{U,V}(u,v) = \underbrace{\frac{u}{\mu^2} e^{-u/\mu} \mathbb{1}_{(0,\infty)}(u)}_{f_U(u)}\ \underbrace{\mathbb{1}_{(0,1)}(v)}_{f_V(v)}.$$

Logo

$$U \sim \text{Gama}(2, 1/\mu), \qquad V \sim \text{Unif}(0,1),$$

e U e V são independentes.

O resultado é mais forte do que parece: depois de observar o tempo total, a fração desse tempo gasta na primeira tarefa continua uniformemente distribuída. Essa independência não é uma propriedade da transformação sozinha; o próximo exemplo mostra o que acontece quando mantemos U e V, mas mudamos a distribuição de X e Y.

◁

Exemplo 7.12 — A mesma transformação, outra geometria.

Agora suponha que X e Y sejam independentes e uniformes em $(0,1)$, mas mantenha exatamente as mesmas quantidades

$$U = X + Y, \qquad V = \frac{X}{X+Y}.$$

Será que soma e proporção continuam independentes?

Solução. A inversa continua sendo

$$X = UV, \qquad Y = U(1-V),$$

e o módulo do jacobiano continua sendo u. A diferença está no suporte: agora precisamos impor simultaneamente

$$0 < uv < 1, \qquad 0 < u(1-v) < 1.$$

Como $0 < u < 2$ e $0 < v < 1$, essas restrições dão

$$\begin{cases} 0 < v < 1, & 0 < u < 1, \\ \dfrac{u-1}{u} < v < \dfrac{1}{u}, & 1 < u < 2. \end{cases}$$

Portanto,

$$f_{U,V}(u,v) = u$$

nessa região e vale zero fora dela.

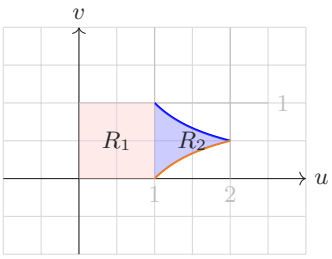


Figura 7.7: Suporte de (U,V) para $X,Y \sim \mathrm{Unif}(0,1)$.

Geometricamente, o quadrado unitário não se transforma agora em um retângulo: para $u > 1$, as restrições $X < 1$ e $Y < 1$ apertam progressivamente o intervalo possível de v. Como o suporte não é um produto de um intervalo em u por um intervalo em v, U e V não são independentes.

Comparado ao exemplo anterior, o jacobiano é o mesmo. O que mudou foi a combinação entre densidade e suporte. É essa combinação, e não a transformação isoladamente, que produziu a independência no caso exponencial.

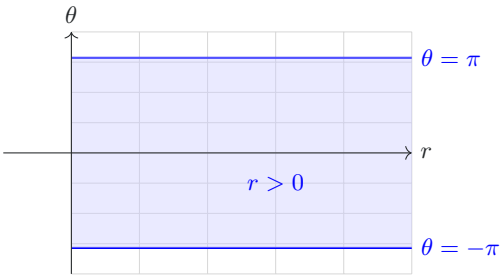
◁

Exemplo 7.13 — Erro de localização em duas dimensões.

Um sistema de posicionamento comete erros independentes X e Y nas direções horizontal e vertical, ambos com distribuição $N(0,\sigma^2)$. Para quem usa o sistema, porém, interessa mais a distância do erro à posição correta e sua direção. Defina

$$R = \sqrt{X^2 + Y^2}, \qquad \Theta = \operatorname{atan2}(Y,X), \qquad -\pi < \Theta \le \pi.$$

Determine a distribuição conjunta de (R,Θ).



Solução. A densidade conjunta cartesiana é

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right).$$

Em coordenadas polares,

$$x = r\cos\theta, \qquad y = r\sin\theta,$$

e o módulo do jacobiano da transformação inversa é

$$\left|\det\begin{pmatrix}\cos\theta & -r\sin\theta\\ \sin\theta & r\cos\theta\end{pmatrix}\right| = r.$$

Logo,

$$f_{R,\Theta}(r,\theta) = \frac{r}{2\pi\sigma^2}e^{-r^2/(2\sigma^2)}, \qquad r>0,\ -\pi<\theta\le\pi.$$

A expressão já se separa em duas densidades:

$$f_{R,\Theta}(r,\theta) = \underbrace{\frac{r}{\sigma^2}e^{-r^2/(2\sigma^2)}}_{f_R(r)}\ \underbrace{\frac{1}{2\pi}}_{f_\Theta(\theta)}.$$

Portanto,

$$R \sim \text{Rayleigh}(\sigma), \qquad \Theta \sim \text{Unif}(-\pi,\pi],$$

e R e Θ são independentes.

A interpretação é simples: a simetria circular do erro gaussiano torna todas as direções igualmente prováveis, e a magnitude do erro não carrega informação sobre a direção em que ele ocorreu.

◁

Exemplo 7.14 — A conexão Beta–Gama.
O exemplo das duas tarefas exponenciais sugere uma pergunta: a independência entre soma e proporção sobrevive além do caso exponencial? Sejam

$$X \sim \text{Gama}(m,\beta), \qquad Y \sim \text{Gama}(n,\beta)$$

independentes, com a mesma taxa β. Defina

$$S = X + Y, \qquad P = \frac{X}{X+Y}.$$

Determine a distribuição conjunta de (S,P).
Solução. A transformação inversa é

$$X = SP, \qquad Y = S(1-P),$$

com suporte $s > 0$, $0 < p < 1$. O módulo do jacobiano é s. Assim,

$$\begin{aligned} f_{S,P}(s,p) &= \frac{\beta^{m+n}}{\Gamma(m)\Gamma(n)}(sp)^{m-1}[s(1-p)]^{n-1}e^{-\beta s}s \\ &= \left[\frac{\beta^{m+n}}{\Gamma(m+n)}s^{m+n-1}e^{-\beta s}\right]\left[\frac{\Gamma(m+n)}{\Gamma(m)\Gamma(n)}p^{m-1}(1-p)^{n-1}\right]. \end{aligned}$$

Reconhecemos as duas parcelas como densidades. Portanto,

$$S \sim \text{Gama}(m+n,\beta), \qquad P \sim \text{Beta}(m,n),$$

e S e P são independentes.

Quando $m = n = 1$, recuperamos exatamente o exemplo das duas tarefas exponenciais: a soma é gama e a proporção é uniforme. O caso geral mostra que a independência não era um acidente de uma conta simples, mas parte de uma estrutura Beta–Gama.

◁

7.2.1 Variáveis auxiliares

Às vezes, interessa apenas $Z = g(X,Y)$, mas o método do jacobiano transforma duas coordenadas. Introduzimos então uma variável auxiliar W, escolhida de modo que a transformação $(X,Y) \mapsto (Z,W)$ possa ser invertida. Por exemplo, podemos tomar

$$Z = g(X,Y)$$

e completar a transformação com $W = X$ ou $W = Y$. Depois de obter $f_{Z,W}$, integramos em w para recuperar a marginal f_Z.

Exemplo 7.15 — Densidade da soma de variáveis aleatórias.

Considere duas variáveis aleatórias X e Y. Calcule $f_Z(z)$ sendo $Z = X + Y$ a soma das variáveis aleatórias X e Y.

Solução. Suponha $Z = X + Y$ e escolha $W = Y$ para que a transformação seja injetiva e a solução seja dada por $y = w, x = z - w$. O Jacobiano da transformação é dado por

$$J(x, y) = \begin{vmatrix} 1 & 1 \\ 0 & 1 \end{vmatrix} = 1$$

E assim

$$f_{ZW}(z, w) = f_{XY}(z - w, w).$$

E logo

$$f_Z(z) = \int f_{ZW}(z, w)dw = \int_{-\infty}^{+\infty} f_{XY}(z - w, w)dw.$$

Se X e Y forem independentes, a última integral é a convolução de f_X e f_Y; veja a Definição 6.4.

◁

Exemplo 7.16 — Densidade do produto de variáveis aleatórias.

Considere duas variáveis aleatórias contínuas X e Y. Determine a densidade de $Z = XY$.

Solução.

Seja $Z = XY$. Com $W = X$, o sistema $xy = z$, $x = w$ possui uma única solução: $x_1 = w$ e $y_1 = \frac{z}{w}$. Nesse caso, temos $J(x, y) = -w$, e consequentemente:

$$f_{ZW}(z, w) = \frac{1}{|w|} f_{XY}\left(w, \frac{z}{w}\right).$$

Portanto, a densidade da variável aleatória $Z = XY$ é dada por:

$$f_Z(z) = \int_{-\infty}^{\infty} \frac{1}{|w|} f_{XY}\left(w, \frac{z}{w}\right) dw.$$

Caso especial Agora assumimos que as variáveis aleatórias X e Y são independentes e possuem distribuição uniforme no intervalo (0,1). Nesse caso, temos que $z < w$ e

$$f_{XY}\left(w, \frac{z}{w}\right) = f_X(w) f_Y\left(\frac{z}{w}\right) = 1.$$

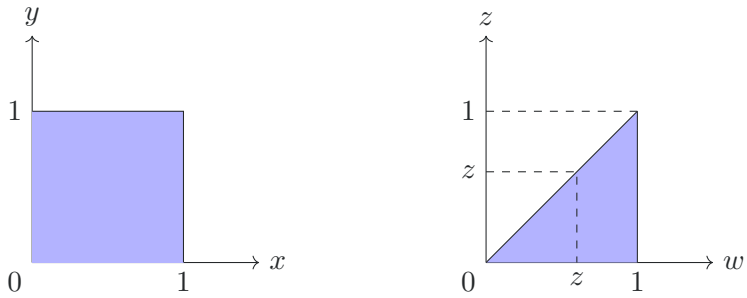
Assim, obtemos (veja a Figura 7.8):

$$f_{ZW}(z,w) = \begin{cases} \frac{1}{w} & \text{se } 0 < z < w < 1 \\ 0 & \text{caso contrário} \end{cases}.$$

Logo,

$$f_Z(z) = \int_z^1 \frac{1}{w}\, dw = \begin{cases} -\ln z & \text{se } 0 < z < 1 \\ 0 & \text{caso contrário} \end{cases}.$$

◁



(a) (b)

Figura 7.8: Figura para o cálculo de $f_Z(z)$ sendo $Z = X \cdot Y$.

Funções não injetivas Podemos utilizar o método do jacobiano também em casos em que a função g não é injetiva, bastando que g seja injetiva quando restrita a cada uma de k regiões abertas disjuntas cuja união contêm o valor de X com probabilidade um. Para tanto, suponhamos que $G, G_1, \ldots, G_k$ sejam subregiões abertas do $\mathbb{R}^n$ tais que $G_1, \ldots, G_k$ sejam disjuntas e valha

$$\mathbf{P}\left(X \in \bigcup_{i=1}^{k} G_i\right) = 1,$$

e tais que a função $g|_{G_i}$, a restrição de g a G_i, seja uma correspondência biunívoca entre G_i e $G, \forall i = 1, \ldots, k$. (Neste caso podemos dizer que a função g é k a 1.) Além disso, suponhamos que a função inversa de $g|_{G_i}$, denotada por $h^{(i)}$, satisfaça todas as condições da função h do caso anterior, e indiquemos com $J_i(X, y)$ o jacobiano da função $h^{(i)}$. (Este jacobiano é função de $y \in G$. Notemos que $h^{(i)} : G \to G_i$ é uma bijeção.) Temos, então, o seguinte esquema:

Desde que

$$\mathbf{P}\left(X \in \bigcup_{i=1}^{k} G_i\right) = 1 \quad \text{e} \quad \left[X \in \bigcup_{i=1}^{k} G_i\right] \subset [Y \in G]$$

temos $\mathbf{P}(Y \in G) = 1$, i.e., Y toma valores só em G (pelo menos com probabilidade 1).

Teorema 7.3

Sob as condições dadas acima, se X tem densidade $f(x_1, \ldots, x_n)$, então Y tem densidade

$$f_Y(y) = \begin{cases} \displaystyle\sum_{i=1}^{k} f\left(h^{(i)}(y)\right) \cdot \left|J_i(x,y)\right|, & \text{se } y \in G \\ 0, & \text{se } Y \notin G. \end{cases}$$

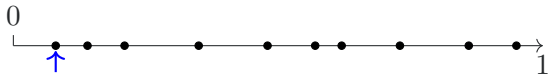
7.3 Estatísticas de ordem

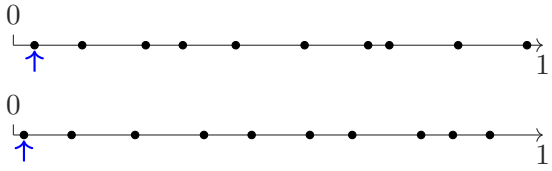
Até aqui as coordenadas tinham nomes. Em muitas perguntas, porém, a identidade de cada observação deixa de importar: queremos apenas a menor, a maior ou a k-ésima menor. O que acontece com a distribuição quando substituímos os rótulos pela ordem? O tempo do vencedor de uma corrida é o mínimo dos tempos, a vida útil de um sistema em série é o mínimo das vidas de seus componentes e a de um sistema em paralelo é o máximo. Ordenar uma amostra transforma, portanto, um vetor aleatório em novas variáveis aleatórias chamadas **estatísticas de ordem**.

Nesta seção, partiremos dos extremos, cujas distribuições são mais imediatas, e depois obteremos a densidade da k-ésima estatística e a distribuição conjunta de duas posições ordenadas. Os exemplos de confiabilidade e amplitude mostram como essas fórmulas são usadas.

Exemplo 7.17.

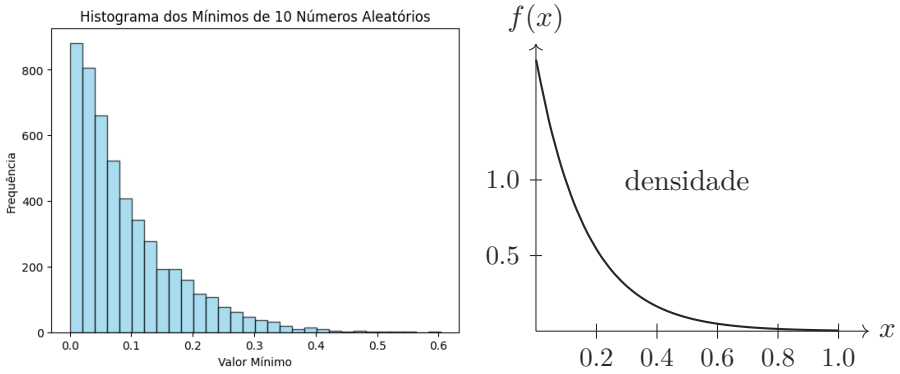
Sejam X_1,X_2,...,X_{10} uma amostra aleatória de tamanho 10 de uma distribuição uniforme em (0,1). Abaixo aparecem três realizações e, em cada uma, o menor valor está destacado.





Reunindo os três mínimos em um único gráfico, obtemos:

Não surpreendentemente, valores do mínimo próximos de zero são mais prováveis do que valores próximos de 1. Se repetirmos o experimento com amostras de tamanho 10, a densidade do mínimo tem o gráfico a seguir.



◁

Considere uma amostra aleatória $X_1, \ldots, X_n$ de uma distribuição contínua com função de distribuição F e densidade f. Definimos:

$$X_{(1)} = \min\left(X_1, X_2, \ldots, X_n\right),$$

que representa o menor valor observado entre as variáveis X_i, e:

$$X_{(n)} = \max\left(X_1, X_2, \ldots, X_n\right),$$

correspondendo ao maior valor observado. De modo geral,

$$X_{(k)} \stackrel{\text{def}}{=} \text{o } k\text{-ésimo menor valor entre } X_1, \ldots, X_n, \qquad k = 1, \ldots, n,$$

é a k-ésima estatística de ordem. A sequência

$$\left(X_{(1)}, X_{(2)}, \ldots, X_{(n)}\right)$$

é a amostra ordenada.

De modo geral, temos $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$, implicando que valores de ordem podem coincidir. Entretanto, se a distribuição comum da amostra $\{X_1, X_2, \ldots, X_n\}$ for contínua, a probabilidade de coincidência é nula. Portanto, nas análises subsequentes, assumiremos que todos os valores observados são distintos, garantindo as seguintes desigualdades estritas:

$$X_{(1)} < X_{(2)} < \cdots < X_{(n)}.$$

Estatísticas de ordem extremas É direto descrever as distribuições de $X_{(1)}$ e $X_{(n)}$. Começando pelo máximo,

$$\mathbf{P}\left[X_{(n)} \leq x\right] = \mathbf{P}\left(X_1 \leq x, X_2 \leq x, \ldots, X_n \leq x\right),$$

pois a condição $X_{(n)} \leq x$ implica que todos os valores X_i devem ser menores ou iguais a x, e o contrário também é verdadeiro. Como $X_1, X_2, \ldots, X_n$ são variáveis independentes, podemos escrever:

$$\mathbf{P}\left[X_{(n)} \leq x\right] = [F(x)]^n.$$

Denotando a função de densidade de $X_{(n)}$ por $g_n(x)$, obtemos, ao derivar ambos os lados da expressão anterior:

$$g_n(x) = n[F(x)]^{n-1} f(x).$$

A função de densidade de $X_{(1)}$, que chamamos de $g_1(x)$, pode ser encontrada de forma semelhante. Consideramos:

$$\begin{aligned}
\mathbf{P}\left[X_{(1)} \leq x\right] &= 1 - \mathbf{P}\left[X_{(1)} > x\right] \\
&= 1 - \mathbf{P}\left(X_1 > x, X_2 > x, \ldots, X_n > x\right) \\
&= 1 - [1 - F(x)]^n.
\end{aligned}$$

Portanto, ao derivar, obtemos:

$$g_1(x) = n[1 - F(x)]^{n-1} f(x).$$

Exemplo 7.18.

Suponha que os tempos, em segundos, de oito velocistas sejam variáveis aleatórias independentes, todas com a distribuição uniforme $\mathcal{U}(9.6{,}10.0)$. Queremos encontrar a probabilidade de que o resultado do vencedor seja menor que 9.69, ou seja, calcular $\mathbf{P}(X_{(1)} < 9.69)$.

Solução. A função densidade de probabilidade da variável aleatória X é dada por $f_X(x) = 2.5 \cdot \mathbb{1}_{(9.6,10.0)}(x)$, onde

$$\mathbb{1}_{(9.6,10.0)}(x) = \begin{cases} 1 & x \in (9.6, 10.0) \\ 0 & \text{caso contrário.} \end{cases}$$

Para o valor específico de $x = 9.69$, podemos calcular a função de distribuição acumulada $F_X(x)$.

A probabilidade desejada é dada por:

$$\mathbf{P}(X_{(1)} < x) = 1 - (1 - F_X(x))^8$$

Substituindo $F_X(x)$ na expressão, temos:

$$F_X(x) = \int_{9.6}^{9.69} f_X(t)\,dt = \int_{9.6}^{9.69} 2.5\,dt = 2.5 \cdot (9.69 - 9.6) = 2.5 \cdot 0.09 = 0.225$$

Portanto, substituindo $F_X(9.69)$ na equação da probabilidade:

$$\mathbf{P}(X_{(1)} < 9.69) = 1 - (1 - 0.225)^8 \equiv 1 - (0.775)^8 \approx 0.86986.$$

Assim, concluímos que a probabilidade de que o tempo do vencedor seja inferior a 9.69 segundos é aproximadamente 0.86986, ou 86.986%.

◁

Outras estatísticas de ordem O evento $\{X_{(j)} \leq x\}$ ocorre se pelo menos j das n observações forem menores ou iguais a x. Como cada observação satisfaz essa condição com probabilidade $F(x)$, a contagem correspondente é binomial. Portanto,

$$\begin{aligned} \mathbf{P}\left[X_{(j)} \leq x\right] &= \sum_{k=j}^{n} \mathbf{P}\left(k \text{ das } n \text{ observações são menores ou iguais a } x\right) \\ &= \sum_{k=j}^{n} \binom{n}{k} [F(x)]^k [1 - F(x)]^{n-k}. \end{aligned} \tag{7.12}$$

Ao derivar (7.12), os termos intermediários se cancelam pela identidade $\binom{n}{k+1}(k+1) = \binom{n}{k}(n-k)$. Resta

$$\begin{aligned} g_j(x) &= \binom{n}{j} j[F(x)]^{j-1}[1-F(x)]^{n-j} f(x) \\ &= \frac{n!}{(j-1)!(n-j)!}[F(x)]^{j-1}[1-F(x)]^{n-j} f(x). \end{aligned}$$

Há também uma interpretação local: para $X_{(j)}$ cair em um pequeno intervalo ao redor de x, devem existir $j-1$ observações abaixo de x, uma nesse intervalo e $n-j$ acima de x. As três categorias têm pesos aproximados $F(x)$, $f(x)\,dx$ e $1-F(x)$, o que reproduz a fórmula anterior por uma contagem multinomial.

Se generalizarmos para a densidade conjunta de $X_{(i)}$ e $X_{(j)}$, com $i<j$, a mesma contagem multinomial fornece, para $x_i < x_j$,

$$\begin{aligned} g_{ij}\left(x_i, x_j\right) =& \frac{n!}{(i-1)!(j-i-1)!(n-j)!} \\ & \times \left[F\left(x_i\right)\right]^{i-1}\left[F\left(x_j\right)-F\left(x_i\right)\right]^{j-i-1}\left[1-F\left(x_j\right)\right]^{n-j} f\left(x_i\right) f\left(x_j\right). \end{aligned}$$

Fora da região $x_i < x_j$, essa densidade é zero.

Em particular, a função de densidade conjunta para $X_{(1)}$ e $X_{(n)}$ torna-se:

$$g_{1n}\left(x_1, x_n\right) = \frac{n!}{(n-2)!}\left[F\left(x_n\right)-F\left(x_1\right)\right]^{n-2} f\left(x_1\right) f\left(x_n\right).$$

O mesmo método pode ser aplicado para encontrar a densidade conjunta de $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$, que resulta em:

$$g_{12\cdots n}\left(x_1, x_2, \ldots, x_n\right) = \begin{cases} n! f\left(x_1\right) f\left(x_2\right) \cdots f\left(x_n\right), & x_1 < x_2 < \cdots < x_n, \\ 0, & \text{caso contrário.} \end{cases}$$

A partir dessa função de densidade conjunta, é possível obter a função de densidade marginal para qualquer uma das estatísticas de ordem, embora esse tópico não seja abordado neste texto.

Exemplo 7.19.

Considere componentes eletrônicos de um certo tipo, cujo tempo de vida X em horas é descrito por uma densidade de probabilidade:

$$f(x) = \begin{cases} \frac{1}{100} e^{-x/100}, & x > 0, \\ 0, & \text{caso contrário.} \end{cases}$$

Dessa forma temos que o tempo médio de vida é de 100 horas. Suponha que dois desses componentes operem de forma independente e em série em um sistema; isto é, o sistema falha assim que um dos componentes falha. Determine a função de densidade para Y, o tempo de vida do sistema.

Solução. Como o sistema falha com a falha do primeiro componente, temos $Y = \min(X_1, X_2)$, onde X_1 e X_2 são variáveis aleatórias independentes com a densidade dada. Sabemos que $F(x) = 1 - e^{-x/100}$ para $x \geq 0$, então:

$$\begin{aligned} f_Y(y) &= g_1(y) \\ &= 2[1 - F(y)]f(y) \\ &= 2e^{-y/100} \left(\frac{1}{100}\right) e^{-y/100} \\ &= \begin{cases} \frac{1}{50} e^{-y/50}, & y > 0, \\ 0, & \text{caso contrário.} \end{cases} \end{aligned}$$

Portanto, o mínimo de duas variáveis aleatórias exponenciais também segue uma distribuição exponencial, porém com uma média reduzida pela metade.

◁

Exemplo 7.20.

Considere agora que, na situação anterior, os componentes operem em paralelo: o sistema só falha depois que ambos falharem. Determine a densidade de Y, o tempo de vida do sistema.

Solução. Neste caso, $Y = \max(X_1, X_2)$, e temos:

$$\begin{aligned} f_Y(y) &= g_2(y) \\ &= 2\left[F(y)\right] f(y) \\ &= 2\left(1 - e^{-y/100}\right)\left(\frac{1}{100}\right) e^{-y/100} \\ &= \begin{cases} \frac{1}{50}\left(e^{-y/100} - e^{-y/50}\right), & y > 0, \\ 0, & \text{caso contrário.} \end{cases} \end{aligned}$$

Observe que o máximo de duas variáveis aleatórias exponenciais não é, necessariamente, uma variável aleatória exponencial.

O tempo médio de vida do sistema é

$$\begin{aligned}\mathbf{E}[Y] &= \int_0^\infty y \cdot \frac{1}{50}\left(e^{-y/100} - e^{-y/50}\right) dy \\ &= \frac{1}{50}\left(\int_0^\infty y e^{-y/100}\, dy - \int_0^\infty y e^{-y/50}\, dy\right) \\ &= 150 \text{ horas.}\end{aligned}$$

◁

Exemplo 7.21.

Suponha que $X_1, X_2, \ldots, X_n$ sejam variáveis aleatórias independentes, cada uma com distribuição uniforme no intervalo (0,1).

1. Encontre a função de densidade de probabilidade da amplitude $R = X_{(n)} - X_{(1)}$.
2. Calcule a média e a variância de R.

Solução.

1. No caso de uma distribuição uniforme, temos:

$$f(x) = \begin{cases} 1, & 0 < x < 1, \\ 0, & \text{caso contrário.} \end{cases}$$

e

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

Assim, a função de densidade conjunta de $X_{(1)}$ e $X_{(n)}$ é:

$$f_{1n}(x_1, x_n) = \begin{cases} n(n-1)\left[x_n - x_1\right]^{n-2}, & 0 < x_1 < x_n < 1, \\ 0, & \text{caso contrário.} \end{cases}$$

Para encontrar a densidade de $R = X_{(n)} - X_{(1)}$, fixamos o valor de x_1 e definimos $R = X_n - x_1$. Portanto, $R = g(X_n) = X_n - x_1$, e a função inversa é $X_n = h(R) = R + x_1$. Usando métodos de transformação, temos:

$$f_{1r}(x_1, r) = n(n-1)r^{n-2}(1).$$

Ainda é uma função de x_1 e r, e devemos ter $x_1 \leq 1 - r$. Ao integrar em relação a x_1, obtemos:

$$\begin{aligned} f_R(r) &= \int_0^{1-r} n(n-1) r^{n-2} \mathrm{d}x_1 \\ &= \begin{cases} n(n-1) r^{n-2}(1-r), & 0 < r < 1, \\ 0, & \text{caso contrário.} \end{cases} \end{aligned}$$

Para o item 2, usando integrais beta,

$$\mathbf{E}[R] = \int_0^1 r f_R(r)\, dr = \frac{n-1}{n+1}, \qquad \mathbf{E}[R^2] = \frac{n(n-1)}{(n+1)(n+2)}.$$

Consequentemente,

$$\operatorname{Var}(R) = \mathbf{E}[R^2] - \mathbf{E}[R]^2 = \frac{2(n-1)}{(n+1)^2(n+2)}.$$

◁

7.4 Exemplos integradores

Neste capítulo, as aplicações centrais são mudanças suaves de coordenadas e estatísticas de ordem. Os exemplos a seguir mostram essas duas ideias em ação, primeiro pelo jacobiano e depois por máximos e mínimos.

Exemplo 7.22 — Uma transformação polar clássica.
Sejam U e V independentes e uniformes em (0,1), e defina

$$R = \sqrt{-2 \log U}, \qquad \Theta = 2\pi V.$$

A transformação inversa é

$$U = e^{-R^2/2}, \qquad V = \frac{\Theta}{2\pi}.$$

Pelo método do jacobiano,

$$f_{R,\Theta}(r,\theta) = \frac{1}{2\pi}\, r e^{-r^2/2}, \qquad r > 0, \quad 0 < \theta < 2\pi.$$

Agora considere

$$X = R \cos \Theta, \qquad Y = R \sin \Theta.$$

Na mudança de coordenadas polares, o fator de área é r. Assim,

$$f_{X,Y}(x,y) = \frac{1}{2\pi} e^{-(x^2+y^2)/2} = \left(\frac{1}{\sqrt{2\pi}} e^{-x^2/2}\right)\left(\frac{1}{\sqrt{2\pi}} e^{-y^2/2}\right).$$

Portanto, X e Y são normais padrão independentes.

Essa transformação, conhecida como Box–Muller, mostra o caminho inverso ao usual: em vez de descobrir apenas a distribuição de uma transformação, escolhemos a transformação para construir uma distribuição desejada.

◁

Exemplo 7.23 — O máximo de observações uniformes.

Se $U_1, \ldots, U_n$ são independentes e uniformes em (0,1), seja

$$M_n = \max(U_1, \ldots, U_n).$$

Para $0 \le x \le 1$,

$$\mathbf{P}(M_n \le x) = \mathbf{P}(U_1 \le x, \ldots, U_n \le x) = x^n.$$

Logo,

$$f_{M_n}(x) = n x^{n-1}, \qquad 0 < x < 1,$$

e

$$\mathbf{E}[M_n] = \frac{n}{n+1}.$$

À medida que a amostra cresce, o máximo se aproxima da fronteira 1. O exemplo antecipa uma ideia geral: extremos têm escalas e comportamentos próprios.

◁

Exemplo 7.24 — Redundância e tempo até o primeiro resultado.

Suponha que a mesma tarefa seja realizada simultaneamente por k unidades independentes, cada uma com tempo de conclusão exponencial de taxa λ. Se o trabalho termina assim que chega o primeiro resultado, então

$$T = \min(T_1, \ldots, T_k).$$

Como

$$\mathbf{P}(T > t) = \prod_{i=1}^{k} \mathbf{P}(T_i > t) = e^{-k\lambda t},$$

temos

$$T \sim \text{Exp}(k\lambda), \qquad \mathbf{E}[T] = \frac{1}{k\lambda}.$$

Executar a tarefa em paralelo reduz o tempo médio até o primeiro resultado, embora use mais recursos. A mesma matemática de confiabilidade em série reaparece aqui com uma interpretação computacional.

◁

Ao atravessar este capítulo

Para obter a lei de uma transformação, pode-se trabalhar com a função de distribuição, convolução ou mudança de variáveis. O suporte transformado e o módulo do jacobiano são partes essenciais do cálculo. Estatísticas de ordem mostram que mínimos e máximos também são apenas funções do vetor aleatório.

7.5 Exercícios

Exercício 7.1 — Quociente

Se X e Y são normais padrão independentes, use a mudança $Z = X/Y$, $W = Y$ para obter a densidade de Z. Identifique a distribuição.

Exercício 7.2 — Coordenadas polares

Um ponto (X,Y) é escolhido uniformemente no disco unitário. Para $R = \sqrt{X^2 + Y^2}$ e $\Theta = \text{atan2}(Y,X)$, obtenha a densidade conjunta de (R,Θ), suas marginais e verifique se são independentes.

Exercício 7.3 — Estatísticas de ordem

Se $X_1, \ldots, X_6$ são independentes e uniformes em $(0,1)$, calcule $\mathbf{P}(X_{(1)} > 0{,}2)$, $\mathbf{P}(X_{(6)} < 0{,}8)$ e a densidade de $X_{(6)}$.

CAPÍTULO

8

Esperança

A distribuição de uma variável contém muita informação. Mas e se quisermos apenas um número que represente seu centro? A esperança faz essa compressão por meio de uma média ponderada. No caso discreto, os pesos são probabilidades; no contínuo, a densidade desempenha esse papel. Se X é discreta e assume valores no conjunto enumerável C, então

$$\mathbf{E}[X] = \sum_{x \in C} x\mathbf{P}(X = x) \tag{8.1}$$

Se X tem densidade f_X, a fórmula correspondente é

$$\mathbf{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx \tag{8.2}$$

Essas fórmulas são o ponto de partida do capítulo. Depois de estabelecer suas propriedades, calcularemos esperanças de funções de uma ou várias variáveis, estudaremos variância, covariância e correlação e usaremos o condicionamento para decompor problemas complexos. Momentos e funções geradoras encerram o capítulo, sempre com somas e integrais usuais do cálculo.

8.1 Definição e propriedades fundamentais

As fórmulas já são conhecidas. O que precisamos agora é entender quais contas podemos fazer com elas sem reconstruir uma distribuição inteira a

cada passo. A linearidade, as indicadoras e a fórmula das caudas serão as primeiras ferramentas desse tipo.

Definição 8.1

Se X é discreta, com valores $x_1, x_2, \ldots$, definimos

$$\mathbf{E}[X] = \sum_i x_i \mathbf{P}(X = x_i),$$

desde que a soma seja absolutamente convergente. Se X possui densidade f_X, definimos

$$\mathbf{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx,$$

desde que $\int_{-\infty}^{\infty} |x| f_X(x)\, dx < \infty$. Nos dois casos dizemos que X é *integrável* quando $\mathbf{E}[|X|] < \infty$.

Se a distribuição é mista, calculamos separadamente a contribuição das massas pontuais e a da densidade. Assim, se X assume valores x_i com massas p_i e tem também uma parte contínua de densidade f, então

$$\mathbf{E}[g(X)] = \sum_i g(x_i) p_i + \int_{-\infty}^{\infty} g(x) f(x)\, dx,$$

sempre que a soma e a integral convergirem absolutamente.

Proposição 8.1 — Propriedades básicas

Para variáveis integráveis X, Y e constantes a, b, valem:

1. $\mathbf{E}[aX + bY] = a\mathbf{E}[X] + b\mathbf{E}[Y]$;
2. se $X \le Y$, então $\mathbf{E}[X] \le \mathbf{E}[Y]$;
3. $|\mathbf{E}[X]| \le \mathbf{E}[|X|]$;
4. para todo evento A, $\mathbf{E}[\mathbb{1}_A] = \mathbf{P}(A)$.

Teorema 8.2 — Fórmula das caudas

Se $X \geq 0$, então

$$\mathbf{E}[X] = \int_0^\infty \mathbf{P}(X > t)\, dt. \tag{8.3}$$

Se, além disso, X assume valores inteiros não negativos,

$$\mathbf{E}[X] = \sum_{k=1}^{\infty} \mathbf{P}(X \geq k). \tag{8.4}$$

Demonstração. A ideia é a mesma nos casos discreto e contínuo: um valor $x \geq 0$ pode ser medido pelo comprimento do intervalo que fica abaixo dele,

$$x = \int_0^\infty \mathbb{1}_{\{t<x\}}\, dt.$$

Se X é discreta, com valores não negativos x_i e probabilidades p_i, então

$$\begin{aligned}\mathbf{E}[X] &= \sum_i x_i p_i = \sum_i p_i \int_0^{x_i} dt \\ &= \int_0^\infty \sum_{i:x_i>t} p_i\, dt = \int_0^\infty \mathbf{P}(X > t)\, dt.\end{aligned}$$

Se X possui densidade f_X, então

$$\begin{aligned}\mathbf{E}[X] &= \int_0^\infty x f_X(x)\, dx = \int_0^\infty \int_0^x f_X(x)\, dt\, dx \\ &= \int_0^\infty \int_t^\infty f_X(x)\, dx\, dt = \int_0^\infty \mathbf{P}(X > t)\, dt.\end{aligned}$$

Como os integrandos são não negativos, a troca da ordem de integração não cria problemas de cancelamento.

Finalmente, se X assume valores inteiros não negativos, então para $k-1 \leq t < k$ vale $\mathbf{P}(X > t) = \mathbf{P}(X \geq k)$. Particionando a integral em intervalos unitários,

$$\int_0^\infty \mathbf{P}(X > t)\, dt = \sum_{k=1}^{\infty} \mathbf{P}(X \geq k),$$

o que prova a segunda fórmula. ∎

A interpretação é geométrica: em vez de somar os valores de X diretamente, podemos somar suas "camadas". Para cada altura t, contamos a probabilidade de a variável ultrapassar esse nível. No caso inteiro, essas camadas são os níveis $1,2,\ldots$; no caso contínuo, elas variam continuamente com t.

8.2 Esperança de funções de vetores aleatórios

Sejam X,Y variáveis aleatórias e $g : \mathbb{R}^2 \to \mathbb{R}$. Uma estratégia possível seria descobrir primeiro a distribuição de $g(X,Y)$ e só depois calcular sua média. Mas precisamos realmente fazer esse desvio?

Não. Podemos calcular $\mathbf{E}[g(X,Y)]$ diretamente da distribuição conjunta, atribuindo a cada par (x,y) o valor $g(x,y)$ e usando como peso a probabilidade ou a densidade conjunta.

Teorema 8.3 — Esperança de uma função de um vetor aleatório
Seja $g : \mathbb{R}^2 \to \mathbb{R}$.

1. Se X,Y são discretas e
$$\sum_x \sum_y |g(x,y)|\mathbf{P}(X = x,Y = y) < \infty,$$
então
$$\mathbf{E}[g(X,Y)] = \sum_x \sum_y g(x,y)\mathbf{P}(X = x,Y = y).$$

2. Se (X,Y) possui densidade conjunta f_{XY} e
$$\iint_{\mathbb{R}^2} |g(x,y)| f_{XY}(x,y)\,dx\,dy < \infty,$$
então
$$\mathbf{E}[g(X,Y)] = \iint_{\mathbb{R}^2} g(x,y) f_{XY}(x,y)\,dx\,dy.$$

Demonstração. No caso discreto, agrupe os pares (x,y) segundo o valor $z = g(x,y)$. Se $Z = g(X,Y)$, então
$$\mathbf{P}(Z = z) = \sum_{(x,y):g(x,y)=z} \mathbf{P}(X = x,Y = y).$$
Logo,
$$\begin{aligned}\mathbf{E}[Z] &= \sum_z z\mathbf{P}(Z = z) \\ &= \sum_z z \sum_{(x,y):g(x,y)=z} \mathbf{P}(X = x,Y = y) \\ &= \sum_x \sum_y g(x,y)\mathbf{P}(X = x,Y = y).\end{aligned}$$

A convergência absoluta garante que o reagrupamento das somas não altera o resultado.

No caso contínuo, suponha primeiro $g \geq 0$. Pela fórmula das caudas,

$$\mathbf{E}[g(X,Y)] = \int_0^\infty \mathbf{P}(g(X,Y) > t)\, dt.$$

Como

$$\mathbf{P}(g(X,Y) > t) = \iint_{\{(x,y):g(x,y)>t\}} f_{XY}(x,y)\, dx\, dy,$$

trocando a ordem de integração obtemos

$$\begin{aligned}\mathbf{E}[g(X,Y)] &= \iint_{\mathbb{R}^2} \left(\int_0^{g(x,y)} dt \right) f_{XY}(x,y)\, dx\, dy \\ &= \iint_{\mathbb{R}^2} g(x,y) f_{XY}(x,y)\, dx\, dy.\end{aligned}$$

Para uma função que assume ambos os sinais, escreva $g = g^+ - g^-$ e aplique o caso não negativo às duas parcelas. A hipótese de integrabilidade garante que ambas sejam finitas. ∎

Para uma única variável, a mesma regra fica:

Teorema 8.4

Se $g(X)$ é integrável, então

$$\mathbf{E}[g(X)] = \begin{cases} \sum_i g(x_i)\mathbf{P}(X = x_i), & X \text{ discreta}, \\ \int_{-\infty}^{\infty} g(x) f_X(x)\, dx, & X \text{ com densidade}. \end{cases} \tag{8.5}$$

A Tabela 8.1 resume os casos especiais mais importantes de 8.5.

Discreto X	$\mathbf{E}[g(X)] = \sum_i g(t_i)\, \mathbf{P}(X = t_i)$
Absolutamente Contínuo X	$\mathbf{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$

Tabela 8.1: Fórmulas computacionais para $\mathbf{E}[g(X)]$

Exemplo 8.1.

Um acidente ocorre em um ponto X, que é distribuído uniformemente ao longo de uma estrada de comprimento L. No momento do acidente, uma ambulância está no ponto Y, que também é distribuído uniformemente ao longo da mesma estrada. Suponha que X e Y sejam variáveis independentes. A tarefa é determinar a distância esperada entre a ambulância e o local do acidente.

Solução. Para resolver o problema, precisamos encontrar $\mathbf{E}[|X - Y|]$, que representa a distância esperada entre os pontos X e Y. Como X e Y são independentes e uniformemente distribuídos, a função densidade conjunta é:

$$f(x, y) = \frac{1}{L^2}, \quad 0 < x < L, \quad 0 < y < L.$$

De acordo com o Teorema 8.3, podemos expressar a distância esperada como:

$$\mathbf{E}[|X - Y|] = \frac{1}{L^2} \int_0^L \int_0^L |x - y| \, dy \, dx.$$

Vamos calcular a integral interna primeiro:

$$\int_0^L |x - y| \, dy = \int_0^x (x - y) \, dy + \int_x^L (y - x) \, dy.$$

Resolvendo cada uma das integrais separadamente, temos:

$$\int_0^x (x - y) \, dy = \left[xy - \frac{y^2}{2} \right]_0^x = \frac{x^2}{2},$$

$$\int_x^L (y - x) \, dy = \left[\frac{y^2}{2} - xy \right]_x^L = \frac{L^2}{2} - \frac{x^2}{2} - x(L - x).$$

Somando os resultados:

$$\int_0^L |x - y| \, dy = \frac{x^2}{2} + \frac{L^2}{2} - \frac{x^2}{2} - x(L - x) = \frac{L^2}{2} + x^2 - xL.$$

Agora, substituímos na integral externa para encontrar $\mathbf{E}[|X - Y|]$:

$$\mathbf{E}[|X - Y|] = \frac{1}{L^2} \int_0^L \left(\frac{L^2}{2} + x^2 - xL \right) dx.$$

Calculando a integral:

$$\mathbf{E}[|X - Y|] = \frac{1}{L^2} \left[\frac{L^2 x}{2} + \frac{x^3}{3} - \frac{L x^2}{2} \right]_0^L = \frac{L}{3}.$$

Portanto, a distância esperada entre a ambulância e o local do acidente é:

$$\mathbf{E}[|X-Y|] = \frac{L}{3}.$$

◁

Funções de variáveis aleatórias independentes Quando X e Y são independentes, a distribuição conjunta fatora. Isso permite calcular uma esperança por somas ou integrais sucessivas usando apenas as distribuições marginais.

Teorema 8.5

Sejam X e Y independentes e $h : \mathbb{R}^2 \to \mathbb{R}$ não negativa ou integrável. No caso discreto,

$$\mathbf{E}[h(X,Y)] = \sum_x \sum_y h(x,y) p_X(x) p_Y(y).$$

Se X e Y têm densidades,

$$\mathbf{E}[h(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x,y) f_X(x) f_Y(y)\, dx\, dy. \tag{8.6}$$

Demonstração. Pela independência, a distribuição conjunta fatora. No caso discreto,

$$\mathbf{P}(X = x, Y = y) = p_X(x) p_Y(y),$$

e, portanto, o Teorema 8.3 fornece imediatamente

$$\mathbf{E}[h(X,Y)] = \sum_x \sum_y h(x,y) p_X(x) p_Y(y).$$

No caso contínuo, a independência equivale a

$$f_{XY}(x,y) = f_X(x) f_Y(y).$$

Substituindo essa fatoração na fórmula do mesmo teorema,

$$\mathbf{E}[h(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x,y) f_X(x) f_Y(y)\, dx\, dy.$$

A hipótese de não negatividade ou integrabilidade é justamente a que garante a validade das somas ou integrais iteradas utilizadas acima. ■

Em particular, a esperança do produto de funções de variáveis aleatórias independentes é o produto de suas esperanças.

Corolário 8.6

Sejam X, Y independentes e sejam g_1, g_2 funções integráveis. Então,

$$\mathbf{E}\left[g_1(X)g_2(Y)\right] = \mathbf{E}\left[g_1(X)\right]\mathbf{E}\left[g_2(Y)\right] \tag{8.7}$$

Demonstração. Aplique o teorema anterior a $h(x,y) = g_1(x)g_2(y)$. Tanto no caso discreto quanto no contínuo, a soma ou integral iterada fatora no produto de duas expressões:

$$\mathbf{E}[g_1(X)g_2(Y)] = \mathbf{E}[g_1(X)]\,\mathbf{E}[g_2(Y)].$$

■

Corolário 8.7

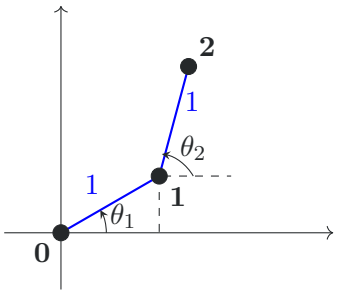
Sejam X, Y independentes. Então,

$$\mathbf{E}\left[XY\right] = \mathbf{E}\left[X\right]\mathbf{E}\left[Y\right] \tag{8.8}$$

A diferença $\mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]$ é denominada covariância. Ela mede associação linear; covariância nula, por si só, não implica independência.

Exemplo 8.2 — Caminhada aleatória no plano.

Considere uma partícula inicialmente localizada em um ponto dado no plano. Suponha que a partícula execute uma sequência de passos, todos de mesmo comprimento, mas em direções escolhidas de forma completamente aleatória. Especificamente, após cada passo, a nova posição da partícula é alcançada a uma distância de exatamente uma unidade da posição anterior, sendo o ângulo do movimento uniformemente distribuído no intervalo $(0, 2\pi)$, como ilustrado na Figura 8.1. Determine o valor esperado do quadrado da distância da partícula à origem após n passos.



0 = posição inicial
1 = posição após o primeiro passo
2 = posição após o segundo passo

Figura 8.1: Caminhada aleatória no plano

Solução. Se (X_i, Y_i) representa, em coordenadas retangulares, a mudança de posição no i-ésimo passo, com $i = 1, \ldots, n$, então temos:

$$\begin{aligned} X_i &= \cos\theta_i, \\ Y_i &= \sin\theta_i, \end{aligned}$$

onde θ_i, para $i = 1, \ldots, n$, é, por hipótese, uma variável aleatória independente e uniformemente distribuída no intervalo $(0, 2\pi)$. Como a posição da partícula após n passos tem coordenadas retangulares $(\sum_{i=1}^n X_i, \sum_{i=1}^n Y_i)$, o quadrado da distância D^2 em relação à origem é dado por:

$$\begin{aligned} D^2 &= \left(\sum_{i=1}^n X_i\right)^2 + \left(\sum_{i=1}^n Y_i\right)^2 \\ &= \sum_{i=1}^n \left(X_i^2 + Y_i^2\right) + \sum_{i\neq j} \left(X_i X_j + Y_i Y_j\right) \\ &= n + \sum_{i\neq j} \left(\cos\theta_i \cos\theta_j + \sin\theta_i \sin\theta_j\right), \end{aligned}$$

onde usamos o fato de que $\cos^2\theta_i + \sin^2\theta_i = 1$.

Ao calcular as esperanças, utilizamos a independência de θ_i e θ_j para $i \neq j$, além das seguintes propriedades de integração sobre o intervalo $(0, 2\pi)$:

$$\begin{aligned} 2\pi \mathbf{E}\left[\cos\theta_i\right] &= \int_0^{2\pi} \cos u \, du = \sin(2\pi) - \sin(0) = 0, \\ 2\pi \mathbf{E}\left[\sin\theta_i\right] &= \int_0^{2\pi} \sin u \, du = \cos(0) - \cos(2\pi) = 0. \end{aligned}$$

Portanto, obtemos o resultado:

$$\mathbf{E}[D^2] = n.$$

◁

Exemplo 8.3 — Número Esperado de Pareamentos.

Suponha que N pessoas joguem seus chapéus no centro de uma sala. Os chapéus são misturados aleatoriamente, e cada pessoa seleciona um chapéu ao acaso. Qual é o número esperado de pessoas que recuperam o próprio chapéu?

Solução. Vamos definir a variável aleatória X como o número de pessoas que selecionam seus próprios chapéus. Podemos expressar X como a soma de variáveis indicadoras individuais:

$$X = X_1 + X_2 + \cdots + X_N,$$

onde, para cada $i = 1, 2, \ldots, N$,

$$X_i = \begin{cases} 1 & \text{se a pessoa } i \text{ selecionar o próprio chapéu,} \\ 0 & \text{caso contrário.} \end{cases}$$

Cada variável X_i indica o evento de a pessoa i recuperar seu próprio chapéu. A probabilidade desse evento é

$$\mathbf{P}[X_i = 1] = \frac{1}{N},$$

pois há N chapéus igualmente prováveis, e apenas um deles pertence à pessoa i.

Assim, o valor esperado de X_i é

$$\mathbf{E}[X_i] = 0 \times \mathbf{P}[X_i = 0] + 1 \times \mathbf{P}[X_i = 1] = \frac{1}{N}.$$

Embora as variáveis X_i não sejam independentes (porque a seleção de chapéus por uma pessoa afeta as opções disponíveis para as outras), a linearidade da esperança matemática não requer independência. Portanto, podemos somar os valores esperados individuais:

$$\mathbf{E}[X] = \mathbf{E}[X_1] + \mathbf{E}[X_2] + \cdots + \mathbf{E}[X_N] = N \times \frac{1}{N} = 1.$$

Portanto, o número esperado de pessoas que recuperam seus próprios chapéus é 1.

◁

Este problema é conhecido como o "problema dos chapéus" ou "problema das cartas desordenadas" e é um exemplo clássico em probabilidade. Ele ilustra a utilidade da linearidade da esperança matemática mesmo quando os eventos não são independentes.

Exemplo 8.4 — Problema do Colecionador de Cupons.

Suponha que existam N tipos diferentes de cupons de desconto, e que cada vez que alguém recolhe um cupom, este tenha igual probabilidade de ser qualquer um dos N tipos. Determine o número esperado de cupons que alguém precisa acumular antes de conseguir um conjunto completo que contenha pelo menos um de cada tipo.

Solução.

Seja X o número total de cupons acumulados até que se obtenha pelo menos um de cada tipo. Vamos calcular $\mathbf{E}[X]$ decompondo X como a soma de variáveis aleatórias X_i, onde X_i representa o número de cupons necessários para passar de possuir $i-1$ tipos distintos para possuir i tipos distintos de cupons.

Assim, temos:

$$X = X_1 + X_2 + \cdots + X_N$$

Inicialmente, não temos nenhum cupom, então o primeiro cupom sempre será de um novo tipo. Portanto, $X_1 = 1$.

Para $i \geq 2$, suponha que já tenhamos $i-1$ tipos distintos de cupons. A probabilidade de que o próximo cupom seja de um tipo novo é:

$$p_i = \frac{N-(i-1)}{N} = \frac{N-i+1}{N}$$

Logo, X_i é uma variável aleatória geométrica com parâmetro p_i. A esperança matemática de uma variável geométrica é dada por $\mathbf{E}[X_i] = \frac{1}{p_i}$, portanto:

$$\mathbf{E}[X_i] = \frac{N}{N-i+1}$$

O número esperado total de cupons necessários é a soma das esperanças individuais:

$$\mathbf{E}[X] = \sum_{i=1}^{N} \mathbf{E}[X_i] = \sum_{i=1}^{N} \frac{N}{N-i+1}$$

Para simplificar a soma, realizamos a substituição $k = N - i + 1$, o que implica $i = N - k + 1$. Quando $i = 1$, $k = N$; quando $i = N$, $k = 1$. Assim, a soma torna-se:

$$\mathbf{E}[X] = \sum_{k=1}^{N} \frac{N}{k} = N \sum_{k=1}^{N} \frac{1}{k} = N \left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{N}\right).$$

◁

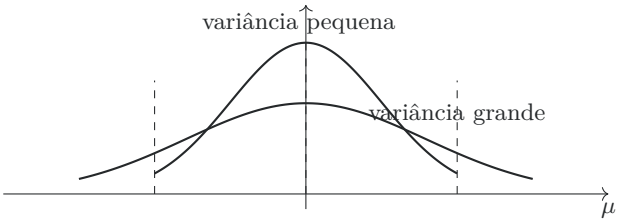
8.3 Variância e covariância

A esperança localiza um centro. Mas duas variáveis com a mesma média podem parecer completamente diferentes: uma quase não oscila, outra se espalha por uma faixa enorme. Como registrar essa diferença? A variância mede o desvio quadrático médio em torno da esperança; a covariância fará o mesmo papel para a variação conjunta de duas variáveis.

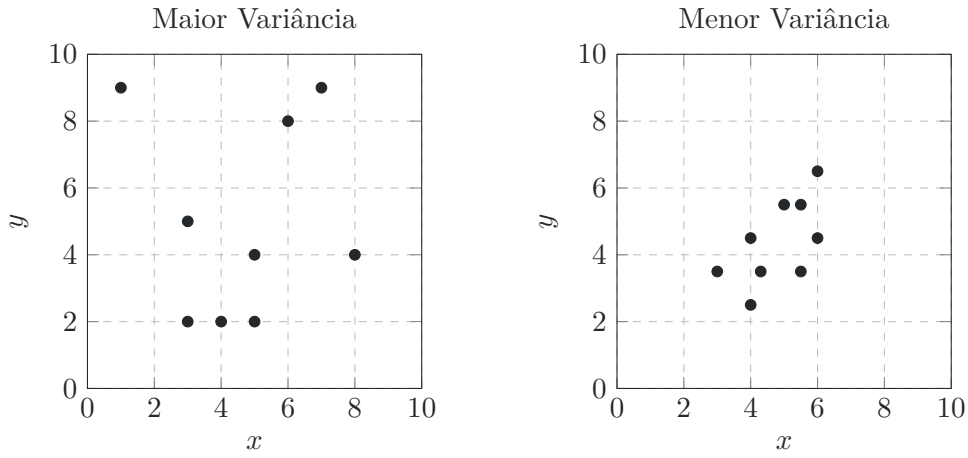
Definição 8.2

Se X é uma variável aleatória com média μ, então a variância de X, representada por $\mathrm{Var}(X)$, é definida como

$$\mathrm{Var}(X) = \mathbf{E}\left[(X - \mu)^2\right]$$



Quanto maior a variância, maior tende a ser a dispersão. Sua raiz quadrada é o **desvio padrão**, expresso na mesma unidade de X, e por isso mais fácil de interpretar diretamente.



Uma fórmula alternativa para $\mathrm{Var}(X)$ é deduzida a seguir:

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbf{E}\left[(X - \mathbf{E}[X])^2\right] \\
&= \mathbf{E}\left[X^2 - 2X\mathbf{E}[X] + \mathbf{E}[X]^2\right] \\
&= \mathbf{E}\left[X^2\right] - 2\mathbf{E}[X]\mathbf{E}[X] + \mathbf{E}[X]^2 \\
&= \mathbf{E}\left[X^2\right] - 2\mathbf{E}[X]^2 + \mathbf{E}[X]^2 \\
&= \mathbf{E}\left[X^2\right] - \mathbf{E}[X]^2
\end{aligned}$$

Exemplo 8.5.

Considere cinco lançamentos de uma moeda honesta. Seja H o número de caras. Calcule $\mathbf{E}[H]$ e $\mathrm{Var}(H)$. Lembramos que

$$\mathbf{P}(H = 0) = \frac{1}{2}^5 = \frac{1}{32}$$

$$\mathbf{P}(H = 1) = \binom{5}{1}\frac{1}{2}^5 = \frac{5}{32}$$

$$\mathbf{P}(H = 2) = \binom{5}{2}\frac{1}{2}^5 = \frac{10}{32}$$

$$\mathbf{P}(H = 3) = \binom{5}{3}\frac{1}{2}^5 = \frac{10}{32}$$

$$\mathbf{P}(H = 4) = \binom{5}{4}\frac{1}{2}^5 = \frac{5}{32}$$

$$\mathbf{P}(H=5) = \binom{5}{5}\frac{1}{2}^5 = \frac{1}{32}$$

Logo a esperança é

$$\mathbf{E}[H] = 1\frac{5}{32} + 2\frac{10}{32} + 3\frac{10}{32} + 4\frac{5}{32} + 5\frac{1}{32} = \frac{80}{32} = \frac{5}{2}$$

Para o calculo da variância começaremos calculando

$$\mathbf{E}[H^2] = 1\frac{5}{32} + 4\frac{10}{32} + 9\frac{10}{32} + 16\frac{5}{32} + 25\frac{1}{32} = \frac{15}{2}$$

Logo a variância é

$$\mathrm{Var}(H) = \frac{15}{2} - \frac{25}{4} = \frac{5}{4}$$

◁

A variância trata de uma variável; para duas variáveis, a covariância descreve como seus desvios em relação às médias variam em conjunto. Ela mede associação linear, não independência.

A covariância pode assumir valores positivos, negativos ou zero:

- Um valor positivo indica que as duas variáveis tendem a aumentar juntas.
- Um valor negativo indica que uma variável tende a aumentar quando a outra diminui.
- Um valor próximo de zero indica que não há uma relação linear clara entre as duas variáveis.

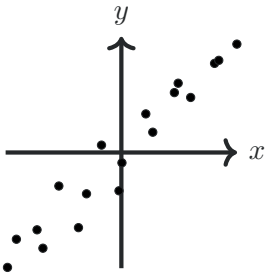


Figura 8.2: a) Covariância positiva

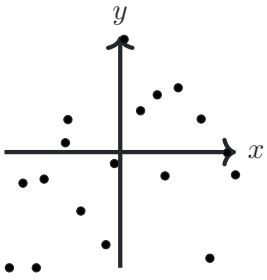


Figura 8.3: b) Não correlacionadas

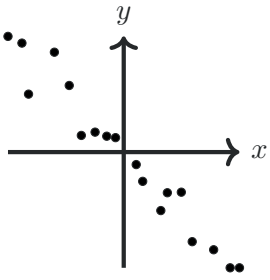


Figura 8.4: c) Covariância negativa

Definição 8.3

A covariância entre X e Y, representada por $\text{Cov}(X,Y)$, é

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

Expandindo o lado direito da definição anterior, vemos que

$$\begin{aligned}\text{Cov}(X,Y) &= \mathbf{E}[XY - \mathbf{E}[X]Y - X\mathbf{E}[Y] + \mathbf{E}[Y]\mathbf{E}[X]] \\ &= \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y] - \mathbf{E}[X]\mathbf{E}[Y] + \mathbf{E}[X]\mathbf{E}[Y] \\ &= \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]\end{aligned}$$

Se X e Y são independentes, então $\text{Cov}(X,Y) = 0$. A recíproca é falsa. Para ver isso, seja X tal que

$$\mathbf{P}(X = 0) = \mathbf{P}(X = 1) = \mathbf{P}(X = -1) = \frac{1}{3}$$

e definindo-se

$$Y = \begin{cases} 0 & \text{se } X \neq 0 \\ 1 & \text{se } X = 0 \end{cases}$$

Como $XY = 0$, temos $\mathbf{E}[XY] = 0$; além disso, $\mathbf{E}[X] = 0$. Logo,

$$\text{Cov}(X,Y) = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y] = 0$$

Entretanto, X e Y são claramente dependentes. A proposição a seguir lista algumas propriedades da covariância.

Proposição 8.8

1. $\text{Cov}(X,Y) = \text{Cov}(Y,X)$
2. $\text{Cov}(X,X) = \text{Var}(X)$
3. $\text{Cov}(aX,Y) = a\,\text{Cov}(X,Y)$
4. $\text{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} \text{Cov}\left(X_i, Y_j\right)$

Demonstração. As proposições 1 e 2resultam imediatamente da definição da covariância, e a proposição 3é deixada como exercício para o leitor. Para demonstrar a proposição 4 , que diz que a operação de cálculo da covariância é aditiva (como é a operação de cálculo do valor esperado), suponha que $\mu_i = \mathbf{E}[X_i]$ e $v_j = \mathbf{E}[Y_j]$. Então,

$$\mathbf{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mu_i, \quad \mathbf{E}\left[\sum_{j=1}^{m} Y_j\right] = \sum_{j=1}^{m} v_j$$

e

$$\begin{aligned}
\operatorname{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) &= \mathbf{E}\left[\left(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu_i\right)\left(\sum_{j=1}^{m} Y_j - \sum_{j=1}^{m} v_j\right)\right] && (8.9)\\
&= \mathbf{E}\left[\sum_{i=1}^{n} (X_i - \mu_i) \sum_{j=1}^{m} (Y_j - v_j)\right] && (8.10)\\
&= \mathbf{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{m} (X_i - \mu_i)(Y_j - v_j)\right] && (8.11)\\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{E}[(X_i - \mu_i)(Y_j - v_j)] && (8.12)
\end{aligned}$$

onde a última igualdade é obtida porque o valor esperado da soma de variáveis aleatórias é igual à soma de seus valores esperados. ■

Tomando $Y_j = X_j$ nas propriedades 2 e 4 da Proposição 8.8, obtemos

$$\begin{aligned}
\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) &= \operatorname{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right)\\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \operatorname{Cov}(X_i, X_j)\\
&= \sum_{i=1}^{n} \operatorname{Var}(X_i) + \sum\sum_{i\neq j} \operatorname{Cov}(X_i, X_j)
\end{aligned}$$

Como cada par de índices $i, j, i \neq j$, aparece duas vezes no somatório duplo, a fórmula anterior é equivalente a

$$\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \operatorname{Var}(X_i) + 2\sum_{i<j} \operatorname{Cov}(X_i, X_j)$$

Se $X_1, \ldots, X_n$ são independentes por pares, no sentido de X_i e X_j serem independentes para $i \neq j$, então a Equação 8.3 reduz-se para

$$\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \operatorname{Var}(X_i)$$

Exemplo 8.6.

Calcule a variância de uma variável aleatória binomial X com parâmetros n e p.

Solução. Como tal variável aleatória representa o número de sucessos em n tentativas independentes quando cada tentativa tem a mesma probabilidade de sucesso p, podemos escrever

$$X = X_1 + \ldots + X_n$$

onde X_i representa variáveis independentes de Bernoulli tais que

$$X_i = \begin{cases} 1 & \text{se a } i\text{-ésima tentativa é um sucesso} \\ 0 & \text{caso contrário} \end{cases}$$

Com isso, da Equação 8.3, obtemos

$$\operatorname{Var}(X) = \operatorname{Var}(X_1) + \ldots + \operatorname{Var}(X_n)$$

Mas

$$\begin{aligned} \operatorname{Var}(X_i) &= \mathbf{E}\left[X_i^2\right] - (\mathbf{E}[X_i])^2 \\ &= \mathbf{E}[X_i] - (\mathbf{E}[X_i])^2 \qquad \text{já que } X_i^2 = X_i \\ &= p - p^2 \end{aligned}$$

Logo,

$$\operatorname{Var}(X) = np(1-p)$$

◁

Exemplo 8.7.

Sejam $X_1, \ldots, X_n$ variáveis aleatórias independentes identicamente distribuídas com valor esperado μ e variância σ^2, e suponha que $\bar{X} = \sum_{i=1}^{n} X_i/n$ seja a média amostral. As grandezas $X_i - \bar{X}, i = 1, \ldots, n$, são chamadas de desvios, e elas são iguais às diferenças entre os valores individuais dos dados e a média amostral. A variável aleatória

$$S^2 = \sum_{i=1}^{n} \frac{\left(X_i - \bar{X}\right)^2}{n-1}$$

é chamada de variância amostral. Determine

1. $\mathrm{Var}(\bar{X})$ e
2. $\mathbf{E}\left[S^2\right]$.

Solução. 1

$$\begin{aligned}
\mathrm{Var}(\bar{X}) &= \left(\frac{1}{n}\right)^2 \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \\
&= \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} \mathrm{Var}\left(X_i\right) \qquad \text{pela independência} \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

2 Começamos com a seguinte identidade algébrica:

$$\begin{aligned}
(n-1)S^2 &= \sum_{i=1}^{n} \left(X_i - \mu + \mu - \bar{X}\right)^2 \\
&= \sum_{i=1}^{n} (X_i - \mu)^2 + \sum_{i=1}^{n} (\bar{X} - \mu)^2 - 2(\bar{X} - \mu) \sum_{i=1}^{n} (X_i - \mu) \\
&= \sum_{i=1}^{n} (X_i - \mu)^2 + n(\bar{X} - \mu)^2 - 2(\bar{X} - \mu) n (\bar{X} - \mu) \\
&= \sum_{i=1}^{n} (X_i - \mu)^2 - n(\bar{X} - \mu)^2
\end{aligned}$$

Calculando a esperança da equação anterior, obtemos

$$\begin{aligned}
(n-1)\mathbf{E}\left[S^2\right] &= \sum_{i=1}^{n} \mathbf{E}\left[(X_i - \mu)^2\right] - n\mathbf{E}\left[(\bar{X} - \mu)^2\right] \\
&= n\sigma^2 - n\,\mathrm{Var}(\bar{X}) \\
&= (n-1)\sigma^2
\end{aligned}$$

onde usamos o item 1 e o fato de que $\mathbf{E}[\bar{X}] = \mu$. Dividindo por $n-1$, concluímos que $\mathbf{E}[S^2] = \sigma^2$: a variância amostral é um estimador não viesado da variância da população.

◁

Exemplo 8.8 — Modelo Multinomial.

Considere m tentativas independentes, em que cada uma pode resultar em um dos r possíveis desfechos, com probabilidades associadas $p_1, p_2, \ldots, p_r$, satisfazendo $\sum_{i=1}^{r} p_i = 1$. Definimos N_i, para $i = 1, \ldots, r$, como a quantidade de tentativas que resultam no desfecho i. Nessas condições, as variáveis $N_1, N_2, \ldots, N_r$ seguem uma distribuição multinomial, dada por:

$$\mathbf{P}\left(N_1 = n_1, N_2 = n_2, \ldots, N_r = n_r\right) = \frac{m!}{n_1!\, n_2! \ldots n_r!} p_1^{n_1} p_2^{n_2} \cdots p_r^{n_r},$$

onde $\sum_{i=1}^{r} n_i = m$. Agora, calcule a covariância entre N_i e N_j.

Solução. Podemos começar observando que, para $i \neq j$, é razoável supor que, à medida que N_i aumenta, N_j tende a diminuir. Assim, é intuitivo concluir que essas variáveis apresentam uma covariância negativa. Vamos calcular sua covariância usando a Proposição 8.8 e as representações

$$N_i = \sum_{k=1}^{m} I_i(k) \quad \text{e} \quad N_j = \sum_{k=1}^{m} I_j(k)$$

onde

$$I_i(k) = \begin{cases} 1 & \text{se a tentativa } k \text{ levar ao resultado } i \\ 0 & \text{caso contrário} \end{cases}$$

$$I_j(k) = \begin{cases} 1 & \text{se a tentativa } k \text{ levar ao resultado } j \\ 0 & \text{caso contrário} \end{cases}$$

Da Proposição 8.8, temos

$$\operatorname{Cov}\left(N_i, N_j\right) = \sum_{\ell=1}^{m} \sum_{k=1}^{m} \operatorname{Cov}\left(I_i(k), I_j(\ell)\right)$$

Agora, por um lado, quando $k \neq \ell$,

$$\operatorname{Cov}\left(I_i(k), I_j(\ell)\right) = 0$$

pois o resultado da tentativa k é independente do resultado da tentativa ℓ. Por outro lado,

$$\begin{aligned}\operatorname{Cov}\left(I_i(\ell), I_j(\ell)\right) &= \mathbf{E}\left[I_i(\ell)I_j(\ell)\right] - \mathbf{E}\left[I_i(\ell)\right]\mathbf{E}\left[I_j(\ell)\right] \\ &= 0 - p_i p_j = -p_i p_j\end{aligned}$$

onde a equação usa o fato de que $I_i(\ell)I_j(\ell) = 0$, já que a tentativa ℓ não pode levar simultaneamente aos resultados $i\mathbf{E}j$. Com isso, obtemos

$$\operatorname{Cov}\left(N_i, N_j\right) = -mp_i p_j$$

o que está de acordo com nossa intuição de que as variáveis N_i e N_j são negativamente correlacionados.

◁

Exemplo 8.9.

Considere X e Y definidos como:

$$X = \cos\Theta, \quad Y = \sin\Theta,$$

onde Θ é uma variável aleatória uniformemente distribuída no intervalo $(0, 2\pi)$.

1. Mostre que X e Y são não-correlacionadas.
2. Mostre que X e Y não são independentes.

X e Y são não-correlacionadas.

Sabemos que

$$f_\Theta(\theta) = \begin{cases} \frac{1}{2\pi}, & 0 < \theta < 2\pi, \\ 0, & \text{caso contrário.} \end{cases}$$

Então:

$$\mathbf{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx = \int_0^{2\pi} \cos\theta f_\Theta(\theta)\, d\theta = \frac{1}{2\pi}\int_0^{2\pi} \cos\theta\, d\theta = 0.$$

De forma semelhante:

$$\mathbf{E}[Y] = \frac{1}{2\pi}\int_0^{2\pi} \sin\theta\, d\theta = 0.$$

Agora, calculemos $\mathbf{E}[XY]$:

$$\mathbf{E}[XY] = \frac{1}{2\pi}\int_0^{2\pi} \cos\theta\sin\theta\, d\theta = \frac{1}{4\pi}\int_0^{2\pi} \sin(2\theta)\, d\theta = 0.$$

Portanto, como $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$, concluímos que X e Y são não-correlacionadas.

X e Y não são independentes.

Calculando:

$$\mathbf{E}[X^2] = \frac{1}{2\pi}\int_0^{2\pi} \cos^2\theta\, d\theta = \frac{1}{4\pi}\int_0^{2\pi} (1+\cos(2\theta))\, d\theta = \frac{1}{2}.$$

De forma semelhante:

$$\mathbf{E}[Y^2] = \frac{1}{2\pi}\int_0^{2\pi} \sin^2\theta\, d\theta = \frac{1}{4\pi}\int_0^{2\pi} (1-\cos(2\theta))\, d\theta = \frac{1}{2}.$$

Por fim:

$$\mathbf{E}[X^2Y^2] = \frac{1}{2\pi}\int_0^{2\pi} \cos^2\theta \sin^2\theta\, d\theta = \frac{1}{16\pi}\int_0^{2\pi} (1-\cos(4\theta))\, d\theta = \frac{1}{8}.$$

Logo:

$$\mathbf{E}[X^2Y^2] = \frac{1}{8} \neq \frac{1}{4} = \mathbf{E}[X^2]\mathbf{E}[Y^2].$$

Isso implica que X e Y não são independentes.

◁

Consideremos agora dois resultados que são de interesse para a Estatística. Suponhamos que se deseje escolher uma constante real c para "predizer" o valor de uma variável aleatória X. Qual c é o melhor preditor? Se queremos minimizar nosso erro absoluto médio (i.e., a média de $|X-c|$), o melhor preditor é a mediana (veja a definição adiante). Mas se queremos minimizar o erro quadrático médio $\mathbf{E}(X-c)^2$, o melhor preditor é a média:

Proposição 8.9

Seja X uma variável aleatória com $\mathbf{E}[X^2] < \infty$ e $\mu = \mathbf{E}[X]$. Então μ minimiza o erro quadrático médio:

$$\operatorname{Var} X = \mathbf{E}(X-\mu)^2 = \min_{c\in\mathbb{R}} \mathbf{E}(X-c)^2.$$

Demonstração. $(X-c)^2 = (X-\mu+\mu-c)^2 = (X-\mu)^2+2(\mu-c)(X-\mu)+(\mu-c)^2$, logo (pela linearidade da esperança)

$$\begin{aligned}\mathbf{E}(X-c)^2 =& \mathbf{E}(X-\mu)^2 + 2(\mu-c)(\mathbf{E}[X]-\mu)+ \\ &+ (\mu-c)^2 = \operatorname{Var} X + (\mu-c)^2.\end{aligned}$$

Logo$\mathbf{E}(X-c)^2 \geq \mathbf{E}(X-\mu)^2, \forall c \in \mathbb{R}$. ■

Proposição 8.10

Seja X integrável e seja m uma mediana de X. Então m minimiza o erro absoluto médio:

$$\mathbf{E}|X-m| = \min_{c\in\mathbb{R}} \mathbf{E}|X-c|.$$

Por definição, m é uma mediana de X se $\mathbf{P}(X \geq m) \geq 1/2$ e $\mathbf{P}(X \leq m) \geq 1/2$.

8.3.1 Correlação

A covariância não é uma medida normalizada e pode ser difícil de interpretar diretamente, pois depende das unidades das variáveis. Por isso, é comum normalizá-la dividindo pelo produto dos desvios padrão das variáveis, o que resulta no coeficiente de correlação, uma medida normalizada que varia entre -1 e 1 e é mais fácil de interpretar.

Definição 8.4

A **correlação** das duas variáveis aleatórias X e Y, representada por $\rho(X,Y)$, é definida, desde que $\operatorname{Var}(X)$ e $\operatorname{Var}(Y)$ sejam positivas, como

$$\rho(X,Y) = \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}}$$

Pode-se mostrar que

Proposição 8.11

$$-1 \leq \rho(X, Y) \leq 1 \tag{8.13}$$

Demonstração. Para demonstrar a Equação 8.13, suponha que X e Y tenham variâncias dadas por σ_x^2 e σ_y^2, respectivamente. Então, por um lado,

$$\begin{aligned} 0 &\leq \operatorname{Var}\left(\frac{X}{\sigma_x} + \frac{Y}{\sigma_y}\right) \\ &= \frac{\operatorname{Var}(X)}{\sigma_x^2} + \frac{\operatorname{Var}(Y)}{\sigma_y^2} + \frac{2\operatorname{Cov}(X, Y)}{\sigma_x \sigma_y} \\ &= 2[1 + \rho(X, Y)] \end{aligned}$$

o que implica que

$$-1 \leq \rho(X, Y)$$

Por outro lado,

$$\begin{aligned} 0 &\leq \operatorname{Var}\left(\frac{X}{\sigma_x} - \frac{Y}{\sigma_y}\right) \\ &= \frac{\operatorname{Var}(X)}{\sigma_x^2} + \frac{\operatorname{Var} Y}{(-\sigma_y)^2} - \frac{2\operatorname{Cov}(X, Y)}{\sigma_x \sigma_y} \\ &= 2[1 - \rho(X, Y)] \end{aligned}$$

o que implica que

$$\rho(X, Y) \leq 1$$

o que completa a demonstração da Equação 8.13. ■

De fato, como $\operatorname{Var}(Z) = 0$ implica que Z é constante com probabilidade 1, a igualdade ocorre exatamente quando $Y = a + bX$ com probabilidade 1: temos $\rho(X,Y) = 1$ se $b > 0$ e $\rho(X,Y) = -1$ se $b < 0$.

A correlação mede associação *linear*. Valores próximos de 1 ou -1 indicam forte alinhamento linear; o sinal indica a direção. Já $\rho(X,Y) = 0$ significa apenas que as variáveis são não correlacionadas. Não significa independência nem ausência de relação: o exemplo anterior tem covariância zero e, ainda assim, dependência entre X e Y.

8.4 Esperança condicional

A esperança depende da informação que temos. Considere dois dados independentes: X é a soma dos resultados e Y é o resultado do primeiro dado. Antes de observar qualquer dado,

$$\mathbf{E}[X] = 7.$$

Se descobrimos que $Y = 5$, então

$$X = 5 + \text{resultado do segundo dado}, \qquad \mathbf{E}[X \mid Y = 5] = 8{,}5.$$

A variável X não mudou; mudou a informação disponível sobre ela.

A **esperança condicional** é justamente a média depois dessa atualização: primeiro condicionamos a distribuição de X à informação $Y = y$ e, em seguida, calculamos a esperança com essa distribuição.

Definição 8.5

- Se X e Y são conjuntamente discretas, para $p_Y(y) > 0$,

$$\mathbf{E}[X \mid Y = y] = \sum_x x\, p_{X|Y}(x \mid y).$$

- Se X e Y são conjuntamente contínuas, para $f_Y(y) > 0$,

$$\mathbf{E}[X \mid Y = y] = \int_{-\infty}^{\infty} x\, f_{X|Y}(x \mid y)\, dx.$$

Há uma distinção importante. Para y fixado, $\mathbf{E}[X \mid Y = y]$ é um número. Se deixamos y variar, obtemos uma função

$$\varphi(y) = \mathbf{E}[X \mid Y = y].$$

No exemplo dos dados, $\varphi(y) = y + 3{,}5$. Antes de observar Y, portanto, $\varphi(Y)$ ainda é aleatória; essa variável será denotada por

$$\mathbf{E}[X \mid Y].$$

Em resumo: $\mathbf{E}[X \mid Y = y]$ é uma média para uma informação já conhecida; $\mathbf{E}[X \mid Y]$ registra qual média será usada quando Y for observado.

No caso contínuo, por exemplo,

$$\mathbf{E}[Y \mid X = x] = \int_{-\infty}^{\infty} y\, f_{Y|X}(y \mid x)\, dy.$$

Esse número é o centro da distribuição condicional de Y para o valor observado $X = x$. À medida que x varia, esses centros formam a função

$$\varphi(x) = \mathbf{E}[Y \mid X = x],$$

frequentemente chamada de **linha de regressão**; veja a Figura 8.5.

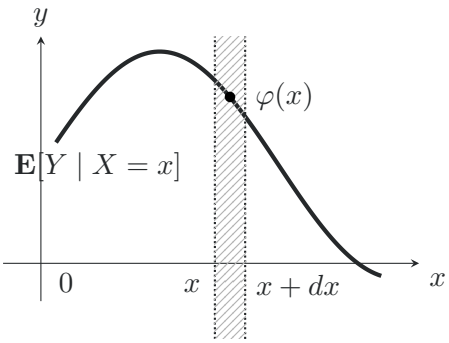


Figura 8.5: Linha de regressão

Exemplo 8.10.

Seja

$$f_{XY}(x,y) = \begin{cases} 1 & 0 < |y| < x < 1 \\ 0 & \text{caso contrário} \end{cases}$$

Determine $\mathbf{E}(X \mid Y)$ e $\mathbf{E}(Y \mid X)$.

Solução. Como a Figura 8.6 mostra, $f_{XY}(x,y) = 1$ na área sombreada e zero em outros lugares. Portanto,

$$f_X(x) = \int_{-x}^{x} f_{XY}(x,y)dy = 2x \quad 0 < x < 1$$

e

$$f_Y(y) = \int_{|y|}^{1} 1dx = 1 - |y| \quad |y| < 1$$

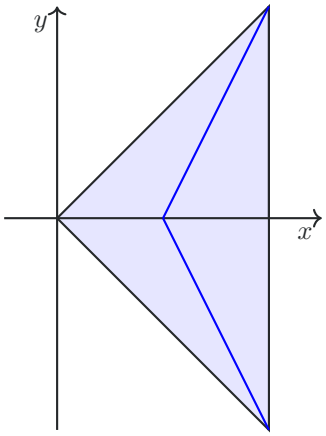


Figura 8.6: Região $0 < |y| < x < 1$.

Isso dá

$$f_{X|Y}(x \mid y) = \frac{1}{1-|y|}, \qquad |y| < x < 1,$$

e

$$f_{Y|X}(y \mid x) = \frac{1}{2x}, \qquad -x < y < x.$$

Portanto,

$$\mathbf{E}[X \mid Y = y] = \int_{|y|}^{1} \frac{x}{1-|y|}\, dx = \frac{1+|y|}{2}, \qquad |y| < 1,$$

enquanto, por simetria,

$$\mathbf{E}[Y \mid X = x] = \int_{-x}^{x} \frac{y}{2x}\, dy = 0, \qquad 0 < x < 1.$$

Logo,

$$\mathbf{E}[X \mid Y] = \frac{1+|Y|}{2}, \qquad \mathbf{E}[Y \mid X] = 0.$$

◁

Se Y não traz informação sobre X, a média também não deve mudar.

Proposição 8.12

Se X e Y são independentes, então

$$\mathbf{E}[X \mid Y] = \mathbf{E}[X].$$

Mais geralmente,

$$\mathbf{E}[g(X) \mid Y] = \mathbf{E}[g(X)]$$

sempre que as esperanças existirem.

Demonstração. A independência implica que a distribuição condicional de X dado $Y = y$ coincide com a distribuição marginal de X: no caso discreto, $p_{X|Y}(x \mid y) = p_X(x)$, e no contínuo, $f_{X|Y}(x \mid y) = f_X(x)$. Substituindo na definição de esperança condicional obtemos o resultado. ∎

Exemplo 8.11.

Se X e Y são variáveis aleatórias binomiais independentes com parâmetros n e p idênticos, calcule o valor esperado condicional de X dado que $X+Y = m$.
Solução. Vamos primeiro calcular a função de probabilidade condicional de X dado que $X + Y = m$. Para $k \leq \min(n, m)$,

$$\begin{aligned}
\mathbf{P}(X = k \mid X + Y = m) &= \frac{\mathbf{P}(X = k, X + Y = m)}{\mathbf{P}(X + Y = m)} \\
&= \frac{\mathbf{P}(X = k, Y = m - k)}{\mathbf{P}(X + Y = m)} \\
&= \frac{\mathbf{P}(X = k)\mathbf{P}(Y = m - k)}{\mathbf{P}(X + Y = m)} \\
&= \frac{\binom{n}{k} p^k (1-p)^{n-k} \binom{n}{m-k} p^{m-k} (1-p)^{n-m+k}}{\binom{2n}{m} p^m (1-p)^{2n-m}} \\
&= \frac{\binom{n}{k}\binom{n}{m-k}}{\binom{2n}{m}}
\end{aligned}$$

onde usamos o fato de que $X + Y$ é uma variável aleatória binomial com parâmetros $2n$ e p. Com isso, a distribuição condicional de X dado que $X+Y = m$ é hipergeométrica. Mais precisamente, trata-se de uma hipergeométrica com população $2n$, n elementos de um dos tipos e amostra de tamanho m. Assim, pela fórmula da esperança da hipergeométrica, obtemos

$$\mathbf{E}[X \mid X + Y = m] = \frac{m}{2}.$$

◁

Exemplo 8.12.

Suponha que a densidade conjunta de X e Y seja dada por

$$f(x,y) = \frac{e^{-x/y}e^{-y}}{y} \quad 0 < x < \infty, 0 < y < \infty$$

Calcule $\mathbf{E}[X \mid Y = y]$.

Solução. Começamos calculando a densidade condicional

$$\begin{aligned} f_{X|Y}(x \mid y) &= \frac{f_{XY}(x,y)}{f_Y(y)} \\ &= \frac{f_{XY}(x,y)}{\int_{-\infty}^{\infty} f_{XY}(u,y)\,du} \\ &= \frac{(1/y)e^{-x/y}e^{-y}}{\int_0^{\infty}(1/y)e^{-x/y}e^{-y}dx} \\ &= \frac{(1/y)e^{-x/y}}{\int_0^{\infty}(1/y)e^{-x/y}dx} \\ &= \frac{1}{y}e^{-x/y} \end{aligned}$$

Com isso, a distribuição condicional de X, dado que $Y = y$, é uma distribuição exponencial com média y. Assim,

$$\mathbf{E}[X \mid Y = y] = \int_0^{\infty} \frac{x}{y}e^{-x/y}dx = y$$

◁

Uma vez fixado $Y = y$, calculamos com a distribuição condicional exatamente como calcularíamos com qualquer outra distribuição. Em particular,

$$\mathbf{E}[g(X) \mid Y = y] = \begin{cases} \sum_x g(x)\, p_{X|Y}(x \mid y), & \text{caso discreto,} \\ \int_{-\infty}^{\infty} g(x)\, f_{X|Y}(x \mid y)\, dx, & \text{caso contínuo,} \end{cases}$$

e a linearidade continua válida. No caso contínuo, embora $\mathbf{P}(Y = y) = 0$, a densidade condicional fornece a regra correta; não estamos dividindo por $\mathbf{P}(Y = y)$.

8.4.1 Esperança condicional como variável aleatória

Agora formalizamos a distinção vista no exemplo inicial.

Definição 8.6

A esperança condicional $\mathbf{E}[X \mid Y]$ é a variável aleatória $\varphi(Y)$, em que

$$\varphi(y) = \mathbf{E}[X \mid Y = y].$$

Para cada informação possível $Y = y$, escolhemos a média correspondente. Qual é a média dessas médias? A resposta é a lei da esperança total.

Proposição 8.13

Se X é integrável, então

$$\mathbf{E}[X] = \mathbf{E}[\mathbf{E}[X \mid Y]]. \tag{8.14}$$

Demonstração. No caso contínuo, substituímos a densidade condicional:

$$\begin{aligned}\mathbf{E}[\mathbf{E}[X \mid Y]] &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} x \, \frac{f_{X,Y}(x,y)}{f_Y(y)} \, dx\right] f_Y(y) \, dy \\ &= \int_{-\infty}^{\infty} x \left[\int_{-\infty}^{\infty} f_{X,Y}(x,y) \, dy\right] dx \\ &= \int_{-\infty}^{\infty} x f_X(x) \, dx = \mathbf{E}[X].\end{aligned}$$

A integrabilidade de X justifica a troca da ordem de integração. No caso discreto, o mesmo cálculo é feito trocando integrais por somas. ∎

Se Y é uma variável aleatória discreta, então a Equação 8.14 diz que

$$\mathbf{E}[X] = \sum_y \mathbf{E}[X \mid Y = y]\mathbf{P}(Y = y) \tag{8.15}$$

Por outro lado, se Y é contínua com densidade $f_Y(y)$, a Equação 8.14 diz que

$$\mathbf{E}[X] = \int_{-\infty}^{\infty} \mathbf{E}[X \mid Y = y] f_Y(y) dy \tag{8.16}$$

A identidade (8.14) é chamada de **lei da esperança total**. Ela calcula a média global ponderando as médias condicionais pela distribuição de Y. Sua utilidade vem do fato de que, em muitos problemas, fixar Y elimina a principal fonte de dificuldade. Os exemplos seguintes exploram exatamente essa estratégia.

Exemplo 8.13.

Em um experimento, um ratinho é colocado em um labirinto com três portas. A primeira porta conduz a um túnel que o leva à saída após 12 minutos. A segunda porta leva a um túnel que o fará retornar ao labirinto após 36 minutos. Já a terceira porta conduz a um túnel que também o faz retornar ao labirinto, mas após 48 minutos. Supondo que o ratinho escolha qualquer uma das portas com igual probabilidade, qual é o tempo esperado até que ele consiga sair do labirinto?

Solução. Seja X o tempo (em minutos) até que o ratinho alcance a saída, e seja Y a porta escolhida inicialmente. O tempo esperado $\mathbf{E}[X]$ pode ser expresso como:

$$\begin{aligned}\mathbf{E}[X] = \mathbf{E}[X \mid Y = 1]\mathbf{P}(Y = 1) + \mathbf{E}[X \mid Y = 2]\mathbf{P}(Y = 2) \\ + \mathbf{E}[X \mid Y = 3]\mathbf{P}(Y = 3).\end{aligned}$$

Como as portas são escolhidas com igual probabilidade, temos:

$$\mathbf{E}[X] = \frac{1}{3}\left(\mathbf{E}[X \mid Y = 1] + \mathbf{E}[X \mid Y = 2] + \mathbf{E}[X \mid Y = 3]\right).$$

Agora, analisamos cada caso condicionado à escolha da porta inicial:

$$\begin{aligned}\mathbf{E}[X \mid Y = 1] &= 12, \\ \mathbf{E}[X \mid Y = 2] &= 36 + \mathbf{E}[X], \\ \mathbf{E}[X \mid Y = 3] &= 48 + \mathbf{E}[X].\end{aligned} \tag{8.17}$$

A justificativa para as equações acima é a seguinte: se o ratinho escolher a primeira porta, ele sairá do labirinto em exatamente 12 minutos, logo $\mathbf{E}[X \mid Y = 1] = 12$. Se escolher a segunda porta, gastará 36 minutos no túnel antes de retornar ao labirinto, e então o problema será idêntico ao inicial, com um tempo esperado adicional $\mathbf{E}[X]$. Portanto, $\mathbf{E}[X \mid Y = 2] = 36 + \mathbf{E}[X]$. Um raciocínio semelhante aplica-se à terceira porta, resultando em $\mathbf{E}[X \mid Y = 3] = 48 + \mathbf{E}[X]$.

Substituindo essas expressões na equação para $\mathbf{E}[X]$, obtemos:

$$\mathbf{E}[X] = \frac{1}{3}\left(12 + (36 + \mathbf{E}[X]) + (48 + \mathbf{E}[X])\right).$$

Simplificando:

$$\mathbf{E}[X] = \frac{1}{3}\left(96 + 2\mathbf{E}[X]\right).$$

Multiplicando ambos os lados por 3 e isolando $\mathbf{E}[X]$, temos:

$$3\mathbf{E}[X] = 96 + 2\mathbf{E}[X] \quad \Longrightarrow \quad \mathbf{E}[X] = 96.$$

Portanto, o tempo esperado para que o ratinho saia do labirinto é de 96 minutos.

◁

Exemplo 8.14 — Esperança da soma de um número aleatório de v.a..

Considere uma cafeteria onde o número de clientes que entram em um determinado dia é uma variável aleatória com média igual a 70. Suponha também que os valores gastos pelos clientes sejam variáveis aleatórias independentes, com uma média comum de $R\$60{,}00$. Além disso, assuma que os gastos de cada cliente são independentes do número total de clientes que entram na cafeteria. Qual é o valor esperado do total gasto na cafeteria em um dia?

Solução.

Seja N o número de clientes que entram na cafeteria e X_i o valor gasto pelo i-ésimo cliente. O gasto total no dia pode ser expresso como $\sum_{i=1}^{N} X_i$. Calculamos a expectativa da soma da seguinte forma:

$$\mathbf{E}\left[\sum_{i=1}^{N} X_i\right] = \mathbf{E}\left[\mathbf{E}\left[\sum_{i=1}^{N} X_i \mid N = n\right]\right].$$

Por outro lado:

$$\mathbf{E}\left[\sum_{i=1}^{N} X_i \mid N = n\right] = \mathbf{E}\left[\sum_{i=1}^{n} X_i \mid N = n\right].$$

Devido à independência entre X_i e N, temos:

$$\mathbf{E}\left[\sum_{i=1}^{n} X_i \mid N = n\right] = \mathbf{E}\left[\sum_{i=1}^{n} X_i\right].$$

Como as variáveis X_i são independentes e identicamente distribuídas, segue que:

$$\mathbf{E}\left[\sum_{i=1}^{n} X_i\right] = n\mathbf{E}[X], \quad \text{onde } \mathbf{E}[X] = \mathbf{E}[X_i].$$

Assim, concluímos que:

$$\mathbf{E}\left[\sum_{i=1}^{N} X_i \mid N\right] = N\mathbf{E}[X].$$

Logo:

$$\mathbf{E}\left[\sum_{i=1}^{N} X_i\right] = \mathbf{E}[N\mathbf{E}[X]] = \mathbf{E}[N]\mathbf{E}[X].$$

No contexto do exemplo, temos $\mathbf{E}[N] = 70$ e $\mathbf{E}[X] = 60$. Portanto, a quantidade esperada de dinheiro gasto na cafeteria em um dia é:

$$70 \times R\$60{,}00 = R\$4.200{,}00.$$

◁

Exemplo 8.15.

O tempo necessário para descarregar um trem que transporta cereais é distribuído uniformemente entre a e b horas, isto é, $T \sim U(a, b)$, quando não há interrupções. No entanto, a grua utilizada na descarga é pouco confiável e está sujeita a avarias que ocorrem de acordo com uma distribuição de Poisson com média de λ avarias por hora. Cada vez que a grua falha, são necessárias d horas para consertá-la. Nosso objetivo é determinar o tempo médio real de descarga do trem, considerando todas essas condições.

Solução. Seja T_R o tempo real de descarga do trem, cujo valor esperado $\mathbf{E}(T_R)$ desejamos encontrar. Observamos que T_R depende do tempo de descarga sem interrupções T e do número de avarias N_T que ocorrem durante esse período. O número de avarias N_T segue uma distribuição de Poisson com parâmetro λT, ou seja, $N_T \sim \text{Poisson}(\lambda T)$.

Para calcular $\mathbf{E}(T_R)$, utilizamos a propriedade da esperança condicional:

$$\mathbf{E}(T_R) = \mathbf{E}[\mathbf{E}(T_R \mid T)].$$

Primeiro, determinamos $\mathbf{E}(T_R \mid T)$. Dado que o tempo de descarga sem interrupções é T e ocorrem $N_T = n$ avarias, o tempo real de descarga é:

$$T_R = T + nd.$$

Assim, a esperança condicional é:

$$\mathbf{E}(T_R \mid T) = \sum_{n=0}^{\infty} (T + nd) \cdot \mathbf{P}(N_T = n).$$

Podemos separar a soma em duas partes:

$$\mathbf{E}(T_R \mid T) = T \sum_{n=0}^{\infty} \mathbf{P}(N_T = n) + d \sum_{n=0}^{\infty} n\mathbf{P}(N_T = n).$$

Sabemos que:

- $\sum_{n=0}^{\infty} \mathbf{P}(N_T = n) = 1$, pois é a soma das probabilidades de todas as possíveis quantidades de avarias.

- $\mathbf{E}[N_T \mid T] = \sum_{n=0}^{\infty} n\mathbf{P}(N_T = n) = \lambda T$, já que N_T é Poisson com média λT.

  Portanto:

$$\mathbf{E}(T_R \mid T) = T \cdot 1 + d \cdot \lambda T = T(1 + d\lambda).$$

  Agora, calculamos $\mathbf{E}(T_R)$:

$$\mathbf{E}(T_R) = \mathbf{E}[\mathbf{E}(T_R \mid T)] = \mathbf{E}[T(1 + d\lambda)] = (1 + d\lambda)\mathbf{E}(T).$$

  Como T é uma variável aleatória uniforme em $[a, b]$, seu valor esperado é:

$$\mathbf{E}(T) = \frac{a + b}{2}.$$

Substituindo, obtemos o tempo médio real de descarga:

$$\mathbf{E}(T_R) = \frac{(a + b)(1 + d\lambda)}{2}.$$

O tempo médio real de descarga do trem é o tempo médio sem interrupções, $\frac{a+b}{2}$, ajustado pelo fator $(1 + d\lambda)$, que representa o impacto médio das avarias da grua no processo de descarga.

◁

Exemplo 8.16.

Três máquinas estão nas posições 0, 1 e 2 metros de uma linha. A cada chamado, qualquer uma delas pode falhar com probabilidade 1/3, independentemente da máquina atendida no chamado anterior. Um trabalhador compara duas estratégias:

1. permanecer na máquina que acabou de reparar até o próximo chamado;
2. retornar sempre à posição central, 1, depois de cada reparo.

Se X é a distância percorrida entre dois chamados consecutivos, qual estratégia produz a menor distância média?

Estratégia 1 Sejam I e J as posições das máquinas nos dois chamados. Elas são independentes e uniformes em $\{0,1,2\}$, e $X = |I - J|$. Portanto,

$$\mathbf{E}_1[X] = \frac{1}{9}\sum_{i=0}^{2}\sum_{j=0}^{2}|i - j| = \frac{8}{9} \text{ metro.}$$

Estratégia 2 Entre dois chamados, o trabalhador vai da posição central até a máquina J e retorna ao centro; assim, $X = 2|J - 1|$. Logo,

$$\mathbf{E}_2[X] = \frac{1}{3}(2 + 0 + 2) = \frac{4}{3} \text{ metro.}$$

Nesse modelo, a estratégia 1 é melhor: ela reduz a distância média de $4/3$ para $8/9$ de metro por intervalo entre chamados.

◁

Exemplo 8.17 — Variância da distribuição geométrica.

Considere uma sequência de tentativas independentes, nas quais cada tentativa possui uma probabilidade de sucesso p. Denotemos por N o momento em que ocorre o primeiro sucesso. O objetivo é determinar a variância de N, ou seja, $\mathrm{Var}(N)$, sendo N uma geométrica.

Solução. Para abordar a questão, consideremos uma variável indicadora Y definida da seguinte forma:

- $Y = 1$ se a primeira tentativa resultar em sucesso (neste caso, $N = 1$),
- $Y = 0$ caso contrário (neste caso, N será maior ou igual a 2).

A variância de N pode ser expressa pela seguinte fórmula:

$$\mathrm{Var}(N) = \mathbf{E}[N^2] - (\mathbf{E}[N])^2.$$

Para calcular $\mathbf{E}[N^2]$, utilizamos a lei da esperança condicional:

$$\mathbf{E}[N^2] = \mathbf{E}[\mathbf{E}[N^2 \mid Y]].$$

Agora, vamos considerar os dois casos possíveis para Y:

1. **Caso em que** $Y = 1$ (a primeira tentativa é um sucesso): - Neste caso, temos $N = 1$, logo:
$$\mathbf{E}[N^2 \mid Y = 1] = 1^2 = 1.$$

2. **Caso em que** $Y = 0$ (a primeira tentativa é um fracasso): Se a primeira tentativa falhar, já se passou uma tentativa sem sucesso. O número total de tentativas necessárias para o primeiro sucesso será igual a 1 (a tentativa falha) mais o número de tentativas adicionais necessárias até o primeiro sucesso, que também segue a mesma distribuição que N. Assim, podemos expressar N como:
$$N = 1 + N',$$
onde N' é a variável aleatória que conta o número de tentativas até o primeiro sucesso após a primeira tentativa fracassada. Portanto, temos:
$$\mathbf{E}[N^2 \mid Y = 0] = \mathbf{E}[(1 + N)^2].$$

Com isso,

$$\begin{aligned}\mathbf{E}\left[N^2\right] &= \mathbf{E}\left[N^2 \mid Y=1\right]\mathbf{P}(Y=1) + \mathbf{E}\left[N^2 \mid Y=0\right]\mathbf{P}(Y=0)\\ &= p + (1-p)\mathbf{E}\left[(1+N)^2\right]\\ &= 1 + (1-p)\mathbf{E}\left[2N+N^2\right]\end{aligned}$$

Como $\mathbf{E}[N] = 1/p$ temos que

$$\mathbf{E}\left[N^2\right] = 1 + \frac{2(1-p)}{p} + (1-p)\mathbf{E}\left[N^2\right]$$

ou

$$\mathbf{E}\left[N^2\right] = \frac{2-p}{p^2}$$

Consequentemente,

$$\begin{aligned}\mathrm{Var}(N) &= \mathbf{E}\left[N^2\right] - (\mathbf{E}[N])^2\\ &= \frac{2-p}{p^2} - \left(\frac{1}{p}\right)^2\\ &= \frac{1-p}{p^2}\end{aligned}$$

◁

Exemplo 8.18 — Problema das Paradas de Ônibus.

O número de pessoas que embarcaram num ônibus é uma variável aleatória de Poisson com média $\mu = 5$. Um ônibus passa por $N = 10$ pontos, e cada passageiro tem a mesma probabilidade de descer em qualquer um dos N pontos, independentemente de onde os outros passageiros descem. Calcule o número esperado de paradas que o ônibus fará antes que todos os passageiros desembarquem.

Solução. Seja X o número de pessoas esperando no ponto de ônibus. Assim, $X \sim \text{Poisson}(5)$. Seja Y o número de pontos nos quais o ônibus para para desembarcar passageiros. (X e Y são variáveis aleatórias.)

$$\mathbf{E}[Y] = \sum_{k=0}^{\infty} \mathbf{E}[Y \mid X=k]\mathbf{P}(X=k)$$

A função de probabilidade de X é

$$\mathbf{P}(X=k) = e^{-5}\frac{5^k}{k!}.$$

Seja Y_i uma variável aleatória indicadora, onde $Y_i = 1$ se o ônibus para no i-ésimo ponto, e $Y_i = 0$ caso contrário.

$$\mathbf{E}[Y \mid X = k] = \mathbf{E}[Y_1 + Y_2 + \cdots + Y_{10} \mid X = k]$$

$$\mathbf{E}[Y_i \mid X = k] = 1 - \left(1 - \frac{1}{10}\right)^k$$

Pela linearidade da esperança:

$$\mathbf{E}[Y \mid X = k] = 10 \cdot \left(1 - \left(1 - \frac{1}{10}\right)^k\right).$$

Finalmente, juntando tudo:

$$\begin{aligned}
\mathbf{E}[Y] &= \sum_{k=0}^{\infty} \left(10 \cdot \left(1 - \left(1 - \frac{1}{10}\right)^k\right) \cdot e^{-5}\frac{5^k}{k!}\right) && (8.18)\\
&= 10e^{-5}\left(\sum_{k=0}^{\infty}\frac{5^k}{k!} - \sum_{k=0}^{\infty}\left(\left(\frac{9}{10}\right)^k \cdot \frac{5^k}{k!}\right)\right) && (8.19)\\
&= 10 - 10e^{-5}\sum_{k=0}^{\infty}\frac{\left(\frac{9}{2}\right)^k}{k!} && (8.20)\\
&= 10 - 10e^{-5}e^{\frac{9}{2}} = 10 - \frac{10}{\sqrt{e}} && (8.21)
\end{aligned}$$

◁

Variância condicional

Definição 8.7 — Variância Condicional

A *variância condicional* de X, dado $Y = y$, é

$$\begin{aligned}
\operatorname{Var}(X \mid Y = y) &= \mathbf{E}\left[(X - \mathbf{E}[X \mid Y = y])^2 \mid Y = y\right] \\
&= \mathbf{E}[X^2 \mid Y = y] - (\mathbf{E}[X \mid Y = y])^2,
\end{aligned}$$

supondo a existência das esperanças envolvidas.

Definição 8.8 — Covariância Condicional

A *covariância condicional* entre X e Y, dado $Z = z$, é

$$\mathrm{Cov}(X,Y \mid Z = z) = \mathbf{E}[XY \mid Z = z] - \mathbf{E}[X \mid Z = z]\mathbf{E}[Y \mid Z = z],$$

supondo a existência das esperanças envolvidas.

Como ocorre com a esperança condicional, $\mathrm{Var}(X \mid Y)$ e $\mathrm{Cov}(X,Y \mid Z)$ são funções das variáveis usadas no condicionamento e, portanto, também são variáveis aleatórias.

Proposição 8.14 — Lei da variância total

Para variáveis aleatórias X e Y, com $\mathbf{E}[X^2] < \infty$,

$$\mathrm{Var}(X) = \mathbf{E}[\mathrm{Var}(X \mid Y)] + \mathrm{Var}(\mathbf{E}[X \mid Y]). \tag{8.22}$$

Podemos provar que essa fórmula é válida começando pela definição de variância e utilizando a fórmula da expectativa condicional duas vezes:

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbf{E}[X^2] - \mathbf{E}[X]^2 \\
&= \mathbf{E}[\mathbf{E}(X^2 \mid Y)] - \mathbf{E}[\mathbf{E}(X \mid Y)]^2 \\
&= \mathbf{E}\Big[\mathbf{E}(X^2 \mid Y) - \mathbf{E}(X \mid Y)^2\Big] + \mathbf{E}[\mathbf{E}(X \mid Y)^2] - \mathbf{E}[\mathbf{E}(X \mid Y)]^2 \\
&= \mathbf{E}[\mathrm{Var}(X \mid Y)] + \mathrm{Var}(\mathbf{E}[X \mid Y]).
\end{aligned}$$

Na equação intermediária, adicionamos e subtraímos o termo necessário para reorganizar como expectativa e variância.

Exemplo 8.19.

Seja N o número de clientes que visitam uma determinada cafeteria em um dia. Suponha que conhecemos $\mathbf{E}[N]$ e $\mathrm{Var}(N)$. Seja X_i o valor médio que o i-ésimo cliente gasta. Admitimos que os X_i's são independentes entre si e também independentes de N. Assumimos ainda que possuem a mesma média e variância:

$$\mathbf{E}[X_i] = \mathbf{E}[X], \quad \mathrm{Var}(X_i) = \mathrm{Var}(X).$$

Seja Y o total de vendas da loja, ou seja,

$$Y = \sum_{i=1}^{N} X_i.$$

Determine $\mathbf{E}[Y]$ e $\mathrm{Var}(Y)$.

Solução. Para encontrar $\mathbf{E}[Y]$, não podemos utilizar diretamente a linearidade da expectativa, pois N é aleatório. No entanto, condicionado a $N = n$, podemos aplicar a linearidade e encontrar $\mathbf{E}[Y \mid N = n]$. Assim, utilizamos a lei da expectativa iterada:

$$\mathbf{E}[Y] = \mathbf{E}[\mathbf{E}[Y \mid N]] \quad \text{(lei da expectativa iterada)}.$$

Substituímos Y pela soma condicional:

$$\mathbf{E}[Y] = \mathbf{E}\left[\mathbf{E}\left[\sum_{i=1}^{N} X_i \mid N\right]\right].$$

Pela linearidade da expectativa:

$$\mathbf{E}[Y] = \mathbf{E}\left[\sum_{i=1}^{N} \mathbf{E}[X_i \mid N]\right].$$

Como X_i's são independentes de N:

$$\mathbf{E}[Y] = \mathbf{E}\left[\sum_{i=1}^{N} \mathbf{E}[X_i]\right].$$

Substituímos $\mathbf{E}[X_i] = \mathbf{E}[X]$:

$$\mathbf{E}[Y] = \mathbf{E}\left[N \cdot \mathbf{E}[X]\right].$$

Como $\mathbf{E}[X]$ é constante:

$$\mathbf{E}[Y] = \mathbf{E}[X] \cdot \mathbf{E}[N].$$

Para encontrar $\mathrm{Var}(Y)$, utilizamos a lei da variância total:

$$\mathrm{Var}(Y) = \mathbf{E}[\mathrm{Var}(Y \mid N)] + \mathrm{Var}(\mathbf{E}[Y \mid N]).$$

Sabemos que:

$$\mathrm{Var}(Y) = \mathbf{E}[N \cdot \mathrm{Var}(X)] + \mathrm{Var}(N \cdot \mathbf{E}[X]).$$

Calculando os termos:

$$\mathrm{Var}(Y) = \mathbf{E}[N] \cdot \mathrm{Var}(X) + (\mathbf{E}[X])^2 \cdot \mathrm{Var}(N).$$

Portanto:

$$\mathrm{Var}(Y) = \mathbf{E}[N] \cdot \mathrm{Var}(X) + (\mathbf{E}[X])^2 \cdot \mathrm{Var}(N).$$

◁

Exemplo 8.20.

Uma máquina em uma fábrica produz chips de computador. Cada chip possui uma probabilidade θ de ser defeituoso, e os chips são produzidos de forma independente, ou seja, a probabilidade de um chip ser defeituoso não depende do status dos outros chips. Suponha que a máquina produza 100 chips por dia. Toda manhã, quando a máquina é ligada, há um valor diferente de θ. De fato, θ é uma variável aleatória que segue uma distribuição Beta(1,4). Se Y representa o número de defeituosos produzidos em um dia escolhido aleatoriamente, qual é o valor esperado de Y?

Solução. Não precisamos calcular a função de probabilidade de Y. Como $Y \mid \theta \sim \text{Bin}(100,\theta)$, usamos

$$\mathbf{E}(Y) = \mathbf{E}[\mathbf{E}(Y \mid \theta)] = \mathbf{E}[100\theta] = 20,$$

pois o valor esperado de θ é $1/5$.

Qual é o desvio padrão do número de defeituosos produzidos em um dia escolhido aleatoriamente? Usando a fórmula para a variância, temos:

$$\begin{aligned}
\text{Var}(Y) &= \mathbf{E}[\text{Var}(Y \mid \theta)] + \text{Var}(\mathbf{E}[Y \mid \theta]) \\
&= \mathbf{E}[100\theta(1-\theta)] + \text{Var}(100\theta) \\
&= 100\mathbf{E}(\theta) - 100\mathbf{E}(\theta^2) + 10{,}000\,\text{Var}(\theta) \\
&= 100\mathbf{E}(\theta) - 100[\text{Var}(\theta) + \mathbf{E}(\theta)^2] + 10{,}000\,\text{Var}(\theta) \\
&= 20 - 100\left(\frac{2}{75} + \frac{1}{25}\right) + 10{,}000\frac{2}{75} \\
&= 280.
\end{aligned}$$

Portanto, o desvio padrão do número de defeituosos é aproximadamente 16.7.

◁

8.5 Momentos

Média e variância usam X e X^2. É natural perguntar o que outras potências podem revelar. Os momentos reúnem essas quantidades numa mesma linguagem e registram, quando existem, diferentes aspectos da forma de uma distribuição.

Definição 8.9

Seja X uma variável aleatória. Para $k = 1, 2, \ldots$, o **momento de ordem** k da variável X é definido como:

$$m_k = \mathbf{E}[X^k],$$

desde que esta quantidade seja finita. O momento de ordem 1, $\mathbf{E}[X]$, é a média da variável aleatória.
Se X é integrável e $\mu = \mathbf{E}[X]$, definimos o **momento central de ordem** k como

$$\mathbf{E}[(X - \mu)^k],$$

desde que essa quantidade também exista. Este momento descreve a dispersão e outras propriedades da variável ao redor de sua média.
De forma semelhante, o **momento absoluto de ordem** k é dado por:

$$\mathbf{E}[|X|^k].$$

A existência dos momentos depende da convergência da integral. Em particular, operações do tipo $\infty - \infty$ tornam o momento indefinido.

Exemplo 8.21 — Momentos de uma Distribuição Gama.

Considere $X \sim \text{Gama}(\alpha, \beta)$. Vamos calcular os momentos de ordem k desta variável.

$$\mathbf{E}[X^k] = \int_0^\infty x^k \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} dx.$$

Simplificando:

$$\mathbf{E}[X^k] = \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{k+\alpha-1} e^{-\beta x} dx.$$

Utilizando a definição da função Gama:

$$\mathbf{E}[X^k] = \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(k+\alpha)}{\beta^{k+\alpha}} = \frac{\Gamma(k+\alpha)}{\Gamma(\alpha)\beta^k}.$$

Escrevendo o resultado como produto de fatores sucessivos e usando a propriedade

A propriedade

$$\Gamma(z+1) = z\Gamma(z)$$

pode ser obtida pela integração por partes da definição da função gama:

$$\begin{aligned}\Gamma(z+1) &= -e^{-t}t^z\Big|_0^\infty + z\int_0^\infty t^{z-1}e^{-t}\,dt \\ &= z\Gamma(z)\end{aligned}$$

que

$$\Gamma(\alpha+1) = \alpha\,\Gamma(\alpha),$$

temos:

$$\mathbf{E}[X^k] = \frac{\alpha(\alpha+1)\cdots(\alpha+k-1)}{\beta^k}.$$

Se $\alpha = 1$, X segue uma distribuição Exponencial com parâmetro β. Nesse caso:

$$\mathbf{E}[X^k] = \frac{k!}{\beta^k}.$$

Assim, os momentos podem ser obtidos de forma simples para este caso específico.

◁

Exemplo 8.22 — Momentos da Distribuição Normal.

Seja a função densidade de probabilidade da variável aleatória normal dada por:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma^2}}, \tag{8.23}$$

onde a média é 0 e $\sigma^2 > 0$ é a variância.

Os momentos da distribuição normal são definidos como:

$$\mathbf{E}[X^n] = \int_{-\infty}^{\infty} x^n f(x)\,dx,$$

onde $f(x)$ é a densidade fornecida em (8.23).

Mostraremos que os momentos de ordem n da variável normal são:

$$\mathbf{E}[X^n] = \begin{cases} 0, & \text{se } n = 2k+1, \\ 1\cdot 3\cdot\cdots\cdot(n-1)\sigma^n, & \text{se } n = 2k. \end{cases} \tag{8.24}$$

Os momentos de ordem ímpar ($n = 2k+1$) são iguais a zero devido à simetria da distribuição normal em torno da média zero. Como $f(-x) = f(x)$ e x^{2k+1} é uma função ímpar, a integral sobre todo o eixo real é zero:

$$\mathbf{E}[X^{2k+1}] = \int_{-\infty}^{\infty} x^{2k+1} f(x)\,dx = 0.$$

Para os momentos pares, é mais simples usar uma recorrência por integração por partes. A normalização da densidade equivale à identidade

$$\int_{-\infty}^{\infty} e^{-x^2/(2\sigma^2)}\,dx = \sigma\sqrt{2\pi}. \tag{8.25}$$

Para $k \geq 1$, tome $u = x^{2k-1}$ e $dv = xe^{-x^2/(2\sigma^2)}\,dx$. Como $v = -\sigma^2 e^{-x^2/(2\sigma^2)}$ e o termo de fronteira é zero,

$$\mathbf{E}[X^{2k}] = (2k-1)\sigma^2\mathbf{E}[X^{2k-2}].$$

Partindo de $\mathbf{E}[X^0] = 1$, a recorrência fornece

$$\mathbf{E}[X^{2k}] = (2k-1)(2k-3)\cdots 3\cdot 1\,\sigma^{2k} = (2k-1)!!\,\sigma^{2k}. \tag{8.26}$$

Isso confirma a segunda parte de (8.24).

◁

8.6 Funções geradoras de momento

Calcular momentos um a um pode ser trabalhoso. Podemos reuni-los em um único objeto e recuperá-los por derivação? A função geradora de momentos faz isso e, além disso, transforma somas independentes em produtos.

Definição 8.10

A **função geradora de momentos** de X é

$$\mathrm{M}_X(t) = \mathbf{E}[e^{tX}],$$

desde que essa esperança seja finita para todo t em algum intervalo $(-h,h)$, com $h > 0$.

O nome vem de sua primeira propriedade: derivando em $t = 0$, recuperamos os momentos,

$$\mathbf{E}[X^n] = M_X^{(n)}(0).$$

Além disso, quando existe em um intervalo ao redor de zero, M_X determina a distribuição de X. Para somas de variáveis independentes, a função geradora da soma é o produto das funções geradoras individuais. Essas duas propriedades tornam a ferramenta especialmente eficiente para identificar distribuições de somas.

As fórmulas de cálculo seguem diretamente da definição de esperança:

$$\mathrm{M}_X(t) = \mathbf{E}\left[e^{tX}\right] = \begin{cases} \sum_i e^{tx_i} p(x_i) & \text{se } X \text{ é discreto} \\ \int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx & \text{se } X \text{ tem densidade.} \end{cases}$$

Exemplo 8.23.

A **distribuição degenerada de probabilidade** é a distribuição de uma variável aleatória constante. Se $X = c$ com probabilidade 1, então

$$M_X(t) = e^{ct}.$$

◁

Exemplo 8.24.

A função geradora de momentos de uma variável aleatória discreta X de Bernoulli com parâmetro p pode ser calculada lembrando que X assume valores 0 e 1, com as respectivas probabilidades $\mathbf{P}(X = 0) = 1 - p$ e $\mathbf{P}(X = 1) = p$. Assim, temos:

$$\begin{aligned} M_X(t) = \mathbf{E}[e^{tX}] = \sum_x e^{tx} \mathbf{P}(X = x) & \\ &= e^{t\cdot 0}(1-p) + e^{t\cdot 1} p \\ &= (1-p) + pe^t \end{aligned}$$

◁

Teorema 8.15

Se X possui função geradora de momentos em um intervalo aberto que contém zero, então

$$\mathbf{E}\left(X^n\right) = \mathrm{M}_X^{(n)}(0) = \left.\frac{d^n \mathrm{M}_X}{dt^n}\right|_{t=0}.$$

Demonstração.

$$\mathrm{M}_X^{(n)}(t) = \frac{d^n}{dt^n}\mathbf{E}\left(e^{tX}\right) = \mathbf{E}\left(\frac{d^n}{dt^n}e^{tX}\right) = \mathbf{E}\left(X^n e^{tX}\right).$$

Avaliando em $t = 0$, obtemos

$$\mathbf{E}\left(X^n\right) = \mathrm{M}_X^{(n)}(0) = \left.\frac{d^n \mathrm{M}_X}{dt^n}\right|_{t=0}$$

■

Exemplo 8.25 — Distribuição binomial com parâmetros n e p.

Se X é uma variável aleatória binomial com parâmetros n e p, então

$$\begin{aligned}
\mathrm{M}(t) &= \mathbf{E}\left[e^{tX}\right] \\
&= \sum_{k=0}^{n} e^{tk} \binom{n}{k} p^k (1-p)^{n-k} \\
&= \sum_{k=0}^{n} \binom{n}{k} \left(pe^t\right)^k (1-p)^{n-k} \\
&= \left(pe^t + 1 - p\right)^n
\end{aligned}$$

onde a última igualdade resulta do teorema binomial. Calculando a derivada, obtemos

$$\mathrm{M}'(t) = n\left(pe^t + 1 - p\right)^{n-1} pe^t$$

Assim,

$$\mathbf{E}[X] = \mathrm{M}'(0) = np$$

Calculando a derivada segunda, obtemos

$$\mathrm{M}''(t) = n(n-1)\left(pe^t + 1 - p\right)^{n-2}\left(pe^t\right)^2 + n\left(pe^t + 1 - p\right)^{n-1} pe^t$$

então

$$\mathbf{E}\left[X^2\right] = \mathrm{M}''(0) = n(n-1)p^2 + np$$

A variância de X é dada por

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbf{E}\left[X^2\right] - (\mathbf{E}[X])^2 \\
&= n(n-1)p^2 + np - n^2p^2 \\
&= np(1-p)
\end{aligned}$$

o que verifica o resultado obtido anteriormente.

◁

Exemplo 8.26 — Distribuição de Poisson com média λ.

Se X é uma variável aleatória de Poisson com parâmetro λ, então

$$\begin{aligned}
\mathrm{M}(t) &= \mathbf{E}\left[e^{tX}\right] \\
&= \sum_{n=0}^{\infty} \frac{e^{tn} e^{-\lambda} \lambda^n}{n!} \\
&= e^{-\lambda} \sum_{n=0}^{\infty} \frac{(\lambda e^t)^n}{n!} \\
&= e^{-\lambda} e^{\lambda e^t} \\
&= e^{\lambda\left(e^t - 1\right)}
\end{aligned}$$

Calculando a derivada, obtemos

$$\begin{aligned}
\mathrm{M}'(t) &= \lambda e^t e^{\lambda\left(e^t - 1\right)} \\
\mathrm{M}''(t) &= \left(\lambda e^t\right)^2 e^{\lambda\left(e^t - 1\right)} + \lambda e^t e^{\lambda\left(e^t - 1\right)}
\end{aligned}$$

Assim,

$$\begin{aligned}
\mathbf{E}[X] &= \mathrm{M}'(0) = \lambda \\
\mathbf{E}\left[X^2\right] &= \mathrm{M}''(0) = \lambda^2 + \lambda \\
\mathrm{Var}(X) &= \mathbf{E}\left[X^2\right] - (\mathbf{E}[X])^2 \\
&= \lambda
\end{aligned}$$

Portanto, a média e a variância de uma variável aleatória de Poisson são iguais a λ.

◁

Exemplo 8.27 — Distribuição exponencial com parâmetro λ.

$$\begin{aligned}
\mathrm{M}(t) &= \mathbf{E}\left[e^{tX}\right] \\
&= \int_0^{\infty} e^{tx} \lambda e^{-\lambda x} dx \\
&= \lambda \int_0^{\infty} e^{-(\lambda - t)x} dx \\
&= \frac{\lambda}{\lambda - t} \qquad \text{para } t < \lambda
\end{aligned}$$

Observamos dessa dedução que, para a distribuição exponencial, $\mathrm{M}(t)$ é definida apenas para valores de t menores que λ. Calculando a derivada de

$\mathrm{M}(t)$, obtemos

$$\mathrm{M}'(t) = \frac{\lambda}{(\lambda - t)^2} \quad \mathrm{M}''(t) = \frac{2\lambda}{(\lambda - t)^3}$$

Portanto,

$$\mathbf{E}[X] = \mathrm{M}'(0) = \frac{1}{\lambda} \quad \mathbf{E}\left[X^2\right] = \mathrm{M}''(0) = \frac{2}{\lambda^2}$$

A variância de X é dada por

$$\begin{aligned} \mathrm{Var}(X) &= \mathbf{E}\left[X^2\right] - (\mathbf{E}[X])^2 \\ &= \frac{1}{\lambda^2} \end{aligned}$$

◁

Teorema 8.16

Se a função geradora de momentos é finita em algum intervalo aberto que contém 0, ela determina de forma única a distribuição de probabilidade.

O teorema anterior afirma que, se duas variáveis aleatórias X e Y têm a mesma função geradora de momento, isto é, $M_X(t) = M_Y(t)$ para todos os t em um intervalo ao redor de 0, então elas têm a mesma distribuição de probabilidade. Assim por exemplo uma variável X tal que $M_X(t) = e^{\lambda\left(e^t - 1\right)}$ é uma variável de Poisson de taxa λ.

Teorema 8.17

Se X e Y são variáveis aleatórias independentes com função geradora de momento $\mathrm{M}_X(t)$ e $\mathrm{M}_Y(t)$ respectivamente, então $\mathrm{M}_{X+Y}(t)$, a função geradora de momento $X + Y$ é dada por

$$\mathrm{M}_{X+Y}(t) = \mathrm{M}_X(t)\mathrm{M}_Y(t).$$

Demonstração.

$$\begin{aligned} \mathrm{M}_{X+Y}(t) &= \mathbf{E}\left[e^{t(X+Y)}\right] \\ &= \mathbf{E}\left[e^{tX}e^{tY}\right] \\ &= \mathbf{E}\left[e^{tX}\right]\mathbf{E}\left[e^{tY}\right] \\ &= \mathrm{M}_X(t)\mathrm{M}_Y(t). \end{aligned}$$

■

Exemplo 8.28.

A função geradora de momentos de uma variável aleatória binomial também pode ser deduzida a partir da função geradora de uma Bernoulli. Se $X \sim \mathrm{Bin}(n,p)$, então X é a soma de n variáveis independentes $X_1, X_2, \ldots, X_n$, onde cada $X_i \sim \mathrm{Bernoulli}(p)$.

Assim, temos:

$$X = \sum_{i=1}^{n} X_i$$

Logo, a função geradora de momentos de $X \sim \mathrm{Bin}(n,p)$ é:

$$M_X(t) = (M_{X_i}(t))^n$$

Substituímos $M_{X_i}(t) = 1 - p + pe^t$:

$$M_X(t) = \left(1 - p + pe^t\right)^n$$

◁

Teorema 8.18 — Função geradora de momento da variável aleatória normal

Seja $\zeta \sim N(\mu, \sigma^2)$. Então, a função geradora de momento de ζ é dada por:

$$M_\zeta(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}.$$

Demonstração. A função geradora de momento $M_\zeta(t)$ é definida como:

$$M_\zeta(t) = \mathbf{E}[e^{t\zeta}] = \int_{-\infty}^{\infty} e^{tx} f_\zeta(x)\, dx,$$

onde $f_\zeta(x)$ é a função densidade de probabilidade da distribuição normal:

$$f_\zeta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Substituindo $f_\zeta(x)$ na expressão de $M_\zeta(t)$:

$$M_\zeta(t) = \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{tx - \frac{(x-\mu)^2}{2\sigma^2}}\, dx.$$

Agora, combinamos os expoentes:

$$e^{tx - \frac{(x-\mu)^2}{2\sigma^2}} = e^{-\frac{1}{2\sigma^2}\left((x-\mu)^2 - 2\sigma^2 tx\right)}.$$

Expansão do quadrado e rearranjo dos termos:

$$\begin{aligned}(x-\mu)^2 - 2\sigma^2 tx &= x^2 - 2\mu x + \mu^2 - 2\sigma^2 tx \\ &= x^2 - 2(\mu + \sigma^2 t)x + \mu^2.\end{aligned}$$

Completando o quadrado:

$$\begin{aligned}x^2 - 2(\mu + \sigma^2 t)x + \mu^2 &= \Big(x - (\mu + \sigma^2 t)\Big)^2 - (\mu + \sigma^2 t)^2 + \mu^2 \\ &= \Big(x - (\mu + \sigma^2 t)\Big)^2 - [(\mu + \sigma^2 t)^2 - \mu^2].\end{aligned}$$

Calculando $(\mu + \sigma^2 t)^2 - \mu^2$:

$$\begin{aligned}(\mu + \sigma^2 t)^2 - \mu^2 &= \mu^2 + 2\mu\sigma^2 t + \sigma^4 t^2 - \mu^2 \\ &= 2\mu\sigma^2 t + \sigma^4 t^2.\end{aligned}$$

Portanto, o expoente torna-se:

$$\begin{aligned}tx - \frac{(x-\mu)^2}{2\sigma^2} &= -\frac{1}{2\sigma^2}\left((x-\mu)^2 - 2\sigma^2 tx\right) \\ &= -\frac{1}{2\sigma^2}\left(\Big(x - (\mu + \sigma^2 t)\Big)^2 - 2\mu\sigma^2 t - \sigma^4 t^2\right) \\ &= -\frac{\big(x - (\mu + \sigma^2 t)\big)^2}{2\sigma^2} + \mu t + \frac{\sigma^2 t^2}{2}.\end{aligned}$$

Substituindo de volta na integral:

$$M_\zeta(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\mu t + \frac{\sigma^2 t^2}{2}} \int_{-\infty}^{\infty} e^{-\frac{\left(x-(\mu+\sigma^2 t)\right)^2}{2\sigma^2}} \, dx.$$

Reconhecemos que a integral é a integral da função densidade de uma distribuição normal com média $\mu + \sigma^2 t$ e variância σ^2, portanto:

$$\int_{-\infty}^{\infty} e^{-\frac{\left(x-(\mu+\sigma^2 t)\right)^2}{2\sigma^2}} \, dx = \sqrt{2\pi\sigma^2}.$$

Assim, simplificando:

$$M_\zeta(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}.$$

■

Teorema 8.19 — Soma de Variáveis Aleatórias Normais Independentes

Sejam $\zeta_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ e $\zeta_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ duas variáveis aleatórias independentes com distribuições normais. Então, a soma $\zeta = \zeta_1 + \zeta_2$ é também uma variável aleatória normal, isto é,

$$\zeta \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

Em outras palavras, a soma de variáveis aleatórias normais independentes é uma variável normal cuja média é a soma das médias e cuja variância é a soma das variâncias.

Demonstração. Para demonstrar que $\zeta = \zeta_1 + \zeta_2$ segue uma distribuição normal com média $\mu = \mu_1 + \mu_2$ e variância $\sigma^2 = \sigma_1^2 + \sigma_2^2$, utilizaremos a propriedade da função geradora de momento. A função geradora de momento de uma variável aleatória ζ é definida por

$$M_\zeta(t) = \mathbf{E}[e^{t\zeta}].$$

Primeiramente, lembramos que a função geradora de momento de uma variável normal $\zeta_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ é dada por

$$M_{\zeta_i}(t) = \exp\left(\mu_i t + \frac{\sigma_i^2 t^2}{2}\right).$$

Como ζ_1 e ζ_2 são independentes, a função geradora de momento da soma $\zeta = \zeta_1 + \zeta_2$ é o produto das FGMs individuais:

$$M_\zeta(t) = M_{\zeta_1+\zeta_2}(t) = M_{\zeta_1}(t) \cdot M_{\zeta_2}(t).$$

Substituindo as expressões das FGMs:

$$\begin{aligned} M_\zeta(t) &= \exp\left(\mu_1 t + \frac{\sigma_1^2 t^2}{2}\right) \cdot \exp\left(\mu_2 t + \frac{\sigma_2^2 t^2}{2}\right) \\ &= \exp\left(\mu_1 t + \mu_2 t + \frac{\sigma_1^2 t^2}{2} + \frac{\sigma_2^2 t^2}{2}\right) \\ &= \exp\left((\mu_1 + \mu_2)t + \frac{(\sigma_1^2 + \sigma_2^2)t^2}{2}\right). \end{aligned}$$

Observamos que $M_\zeta(t)$ é a função geradora de momento de uma variável normal com média $\mu = \mu_1 + \mu_2$ e variância $\sigma^2 = \sigma_1^2 + \sigma_2^2$, pois a função geradora de momento de uma variável normal $\mathcal{N}(\mu, \sigma^2)$ é

$$M_\zeta(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right).$$

Como as funções geradoras são finitas em um intervalo ao redor de zero e coincidem nesse intervalo, o teorema de unicidade permite concluir que ζ segue uma distribuição normal com os parâmetros desejados:

$$\zeta \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

■

A propriedade de que a soma de variáveis normais independentes é também uma variável normal é fundamental em estatística e probabilidade. Essa característica permite, por exemplo, que somas de erros aleatórios normalmente distribuídos permaneçam dentro da estrutura da distribuição normal, facilitando análises em diversas áreas como física, engenharia e ciências sociais.

Além disso, o fato de as variâncias se somarem reflete a natureza da dispersão acumulada ao combinar variáveis aleatórias independentes.

Exemplo 8.29.

Suponha que duas máquinas produzam peças com dimensões que seguem distribuições normais independentes: a primeira com média $\mu_1 = 10$ cm e desvio padrão $\sigma_1 = 0{,}2$ cm, e a segunda com média $\mu_2 = 15$ cm e desvio padrão $\sigma_2 = 0{,}3$ cm. A soma das dimensões das peças produzidas por ambas as máquinas será uma variável normal com média $\mu = 10 + 15 = 25$ cm e desvio padrão $\sigma = \sqrt{0{,}2^2 + 0{,}3^2} \approx 0{,}3606$ cm.

◁

Distribuição	Função geradora de momento $\mathrm{M}_X(t)$
Degenerada δ_a	e^{ta}
Bernoulli $\mathbf{P}(X = 1) = p$	$1 - p + pe^t$
Geométrica $(1-p)^{k-1}p$	$\frac{p}{1-(1-p)e^t}, t < -\ln(1-p)$
Binomial $B(n,p)$	$\left(1 - p + pe^t\right)^n$
Binomial negativa $\mathrm{NB}(r,p)$	$\left(\frac{p}{1-e^t+pe^t}\right)^r, t < -\ln(1-p)$
Poisson Pois(λ)	$e^{\lambda\left(e^t-1\right)}$
Uniforme (contínua) $\mathrm{U}(a,b)$	$\frac{e^{tb}-e^{ta}}{t(b-a)}$
Uniforme (discreta) $\mathrm{DU}(a,b)$	$\frac{e^{at}-e^{(b+1)t}}{(b-a+1)(1-e^t)}$
Normal $N\left(\mu,\sigma^2\right)$	$e^{t\mu+\frac{1}{2}\sigma^2t^2}$
Gama Gama(α,β) (taxa)	$\left(\frac{\beta}{\beta-t}\right)^\alpha, t < \beta$
Exponencial $\mathrm{Exp}(\lambda)$	$\left(1 - t\lambda^{-1}\right)^{-1}, t < \lambda$

8.7 Exemplos integradores

A força da esperança aparece com mais clareza quando a distribuição completa da variável de interesse é difícil, mas sua decomposição em partes simples é fácil. Indicadores resolvem contagens combinatórias; variância e covariância organizam fontes de ruído; e modelos mistos surgem naturalmente em aplicações.

Exemplo 8.30 — Pontos fixos de uma permutação aleatória.

Escolha uniformemente uma permutação π de $\{1,\ldots,n\}$, e seja X o número de pontos que permanecem em sua posição:

$$X = \#\{i : \pi(i) = i\}.$$

Defina $I_i = \mathbb{1}_{\{\pi(i)=i\}}$. Então

$$X = \sum_{i=1}^{n} I_i.$$

Para cada i,

$$\mathbf{P}(\pi(i) = i) = \frac{1}{n}.$$

Pela linearidade,

$$\mathbf{E}[X] = \sum_{i=1}^{n} \mathbf{E}[I_i] = n \cdot \frac{1}{n} = 1.$$

O resultado independe de n. As indicadoras não são independentes, mas a linearidade da esperança não precisa dessa hipótese.

◁

Exemplo 8.31 — Número esperado de arestas e de triângulos.

Em $G(n,p)$, seja M o número total de arestas. Se I_e indica a presença da aresta e,

$$M = \sum_{e \in \binom{[n]}{2}} I_e,$$

e portanto

$$\mathbf{E}[M] = \binom{n}{2} p.$$

Para triângulos, seja T a quantidade de subconjuntos de três vértices que formam um triângulo. Se J_A indica que as três arestas do conjunto $A \in \binom{[n]}{3}$ estão presentes,

$$T = \sum_{A \in \binom{[n]}{3}} J_A.$$

Como $\mathbf{E}[J_A] = p^3$,

$$\mathbf{E}[T] = \binom{n}{3} p^3.$$

A distribuição de T é bem menos simples que a de M, mas sua média continua saindo em uma linha.

◁

Exemplo 8.32 — Dois sensores e uma fonte comum de ruído.

Dois sensores medem a mesma grandeza θ. Suponha que

$$X = \theta + B + \varepsilon_1, \qquad Y = \theta + B + \varepsilon_2,$$

onde B,ε_1,ε_2 têm média zero, são independentes,

$$\operatorname{Var}(B) = \tau^2, \qquad \operatorname{Var}(\varepsilon_1) = \operatorname{Var}(\varepsilon_2) = \sigma^2.$$

Então

$$\mathbf{E}[X] = \mathbf{E}[Y] = \theta,$$

mas os erros dos sensores não são independentes, pois ambos contêm B. De fato,

$$\operatorname{Cov}(X,Y) = \tau^2$$

e

$$\operatorname{Corr}(X,Y) = \frac{\tau^2}{\tau^2 + \sigma^2}.$$

A correlação mede, nesse modelo, a parcela da variabilidade causada por uma fonte comum.

◁

Exemplo 8.33 — Diversificação de uma carteira.

Se X e Y são os retornos de dois ativos e uma fração a do capital é aplicada no primeiro, o retorno da carteira é

$$R = aX + (1-a)Y.$$

Pela linearidade,

$$\mathbf{E}[R] = a\mathbf{E}[X] + (1-a)\mathbf{E}[Y],$$

enquanto

$$\operatorname{Var}(R) = a^2 \operatorname{Var}(X) + (1-a)^2 \operatorname{Var}(Y) + 2a(1-a)\operatorname{Cov}(X,Y).$$

A última parcela mostra por que diversificação não significa apenas possuir muitos ativos: o efeito depende de como seus retornos variam juntos.

◁

Exemplo 8.34 — Contagem com detecção imperfeita.

Suponha que existam N objetos numa região e que cada um seja detectado, independentemente, com probabilidade p. Se D é o número de detecções verdadeiras,

$$D \sim \mathrm{Bin}(N,p).$$

Admita ainda que o procedimento produza um número F de falsos positivos, independente de D, com distribuição Poisson(λ). A contagem observada é

$$Y = D + F.$$

Então

$$\mathbf{E}[Y] = Np + \lambda, \qquad \mathrm{Var}(Y) = Np(1-p) + \lambda.$$

O modelo separa duas fontes de erro: objetos reais que deixam de ser detectados e objetos inexistentes acrescentados pelo procedimento.

◁

Exemplo 8.35 — Seguro com franquia: uma variável naturalmente mista.

Se $X \geq 0$ representa o valor de um prejuízo e o contrato possui franquia $d > 0$, a seguradora paga

$$Y = (X-d)^+ = \max\{X-d,0\}.$$

Mesmo que X tenha densidade, Y não é puramente contínua. Há uma massa em zero:

$$\mathbf{P}(Y=0) = \mathbf{P}(X \leq d) = F_X(d),$$

pois todos os prejuízos abaixo da franquia produzem o mesmo pagamento. Para $y > 0$,

$$f_Y(y) = f_X(y+d).$$

Assim,

$$\mathbf{E}[Y] = \int_d^\infty (x-d) f_X(x)\,dx.$$

Usando a fórmula pela cauda,

$$\mathbf{E}[Y] = \int_d^\infty \mathbf{P}(X > t)\,dt.$$

Se $X \sim \mathrm{Exp}(\lambda)$, segue que

$$\mathbf{E}[Y] = \frac{e^{-\lambda d}}{\lambda}.$$

A franquia produz, portanto, uma distribuição mista de maneira completamente natural e prepara a formulação unificada da esperança.

◁

8.8 Uma visão unificada da esperança

Até aqui, calculamos esperança por duas fórmulas que parecem pertencer a mundos diferentes. Para uma variável aleatória discreta, somamos valores ponderados por probabilidades; para uma variável com densidade, integramos valores ponderados pela densidade:

$$\mathbf{E}[X] = \sum_x x\,\mathbf{P}(X = x), \qquad \mathbf{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\,dx.$$

Na maior parte dos problemas, essa distinção é exatamente o que precisamos. Há, porém, situações em que uma mesma variável possui uma parte discreta e uma parte contínua. Elas oferecem uma boa oportunidade para perguntar se as duas fórmulas acima são realmente definições diferentes ou apenas duas faces de uma construção mais geral.

Esta seção é uma janela para essa construção. A ideia central é simples: em vez de começar pela forma da distribuição de X, começaremos pelos eventos em que X assume determinados valores. Isso conduz naturalmente à integral de Lebesgue com respeito à probabilidade. Não desenvolveremos aqui uma teoria geral de integração; nosso objetivo é apenas enxergar a estrutura comum que já estava presente nos cálculos do capítulo.

Quando soma e integral aparecem juntas

Considere o tempo X, em milissegundos, necessário para responder a uma requisição. Com probabilidade $1/4$, a resposta vem de um cache e leva exatamente 1 ms. Nos demais $3/4$ dos casos, a requisição é processada pelo servidor e o tempo é uniforme no intervalo [2,6].

A distribuição de X não é discreta: ela assume um contínuo de valores entre 2 e 6. Também não é puramente contínua, porque

$$\mathbf{P}(X = 1) = \frac{1}{4}.$$

A parte contínua tem densidade

$$f(x) = \frac{3}{16}\,\mathbb{1}_{[2,6]}(x),$$

pois sua massa total deve ser $3/4$. O valor esperado é, portanto,

$$\begin{aligned}\mathbf{E}[X] &= 1 \cdot \frac{1}{4} + \int_2^6 x\,\frac{3}{16}\,dx \\ &= \frac{1}{4} + 3 = \frac{13}{4}.\end{aligned}$$

A soma registra a contribuição do átomo em 1; a integral registra a contribuição da parte espalhada pelo intervalo. A conta é natural, mas levanta uma questão: por que essas duas operações podem ser reunidas sem conflito?

Uma resposta possível seria aceitar uma fórmula especial para distribuições mistas. Há uma resposta melhor. Soma e integral aparecem juntas porque ambas são casos de uma mesma operação: integrar a variável aleatória contra a probabilidade do espaço amostral.

Para chegar a essa fórmula, começaremos pelo objeto mais simples possível.

Indicadoras e variáveis simples

Se A é um evento, sua função indicadora vale 1 quando A ocorre e 0 caso contrário. Assim,

$$\mathbb{1}_A(\omega) = \begin{cases} 1, & \omega \in A, \\ 0, & \omega \notin A. \end{cases}$$

A esperança de $\mathbb{1}_A$ já foi usada várias vezes neste capítulo:

$$\mathbf{E}[\mathbb{1}_A] = \mathbf{P}(A).$$

Essa identidade contém a ideia essencial. A indicadora transforma um evento em uma variável aleatória; a esperança devolve o peso probabilístico desse evento.

Agora suponha que X assuma apenas os valores $a_1, \ldots, a_n$. Se

$$A_i = \{\omega : X(\omega) = a_i\},$$

então os eventos $A_1, \ldots, A_n$ formam uma partição do espaço amostral e podemos escrever

$$X = \sum_{i=1}^{n} a_i \mathbb{1}_{A_i}.$$

Uma variável dessa forma é chamada *variável aleatória simples*. Para ela, a definição natural é

$$\mathbf{E}[X] = \sum_{i=1}^{n} a_i \mathbf{P}(A_i). \tag{8.27}$$

É exatamente a média ponderada que já conhecemos. O ponto novo é que o peso não está associado, em primeiro lugar, a um ponto da reta: ele está associado a um conjunto de resultados $A_i \subseteq \Omega$.

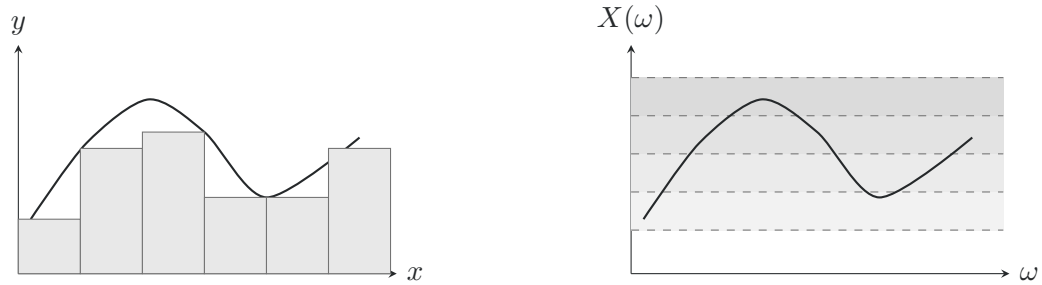
Essa mudança de perspectiva parece pequena, mas é decisiva. Ela permite aproximar variáveis mais complicadas por variáveis simples sem perguntar, a

cada etapa, se a distribuição resultante deve ser tratada por uma soma, por uma integral ou pelas duas coisas.

Riemann e Lebesgue: duas maneiras de organizar a aproximação

A integral de Riemann, familiar do cálculo, aproxima a área sob uma curva dividindo o eixo horizontal. Em cada pequeno intervalo do domínio, escolhemos uma altura e construímos um retângulo. Refinar a partição significa tornar os intervalos cada vez menores.

A construção de Lebesgue pode ser vista de outra maneira. Em vez de organizar a aproximação principalmente por onde estamos no domínio, organizamos os pontos de acordo com os valores assumidos pela função. Escolhemos níveis $y_0 < y_1 < \cdots < y_m$ e agrupamos os pontos para os quais a função cai em cada faixa de alturas. Esses grupos podem ser bastante irregulares no domínio; o que importa é o tamanho que a probabilidade atribui a cada um deles.



Riemann: partição do domínio

Lebesgue: agrupamento por níveis da função

Figura 8.7: Duas maneiras de organizar uma aproximação. Em Riemann, refinamos principalmente o domínio; na construção de Lebesgue, aproximamos a função por níveis e medimos os conjuntos que caem em cada faixa.

No contexto probabilístico, isso se encaixa de forma particularmente simples. Se uma aproximação de X assume os valores $a_1, \ldots, a_n$ nos eventos $A_1, \ldots, A_n$, sua contribuição média é

$$a_1\mathbf{P}(A_1) + \cdots + a_n\mathbf{P}(A_n).$$

Ou seja, a soma de Lebesgue para a esperança já apareceu na Equação (8.27). O que muda ao passar para uma variável geral é que refinamos essas faixas de valores e tomamos um limite.

Há uma diferença conceitual importante em relação ao cálculo usual. O espaço Ω pode nem sequer ser um intervalo da reta. Pode ser um conjunto de sequências, trajetórias, configurações ou qualquer outro espaço amostral. Ainda assim, faz sentido perguntar qual é a probabilidade do conjunto em que X fica entre dois níveis. É essa possibilidade que torna a integral com respeito a $\mathbf{P}$ tão adequada à probabilidade.

De variáveis simples a variáveis gerais

Suponha primeiro que $X \geq 0$. Podemos construir variáveis simples

$$X_1 \leq X_2 \leq \cdots \leq X$$

que se aproximam de X por baixo e cujos degraus ficam progressivamente mais finos. Escrevemos

$$X_n \uparrow X.$$

Cada X_n possui uma esperança definida pela soma (8.27). A ideia de Lebesgue é definir

$$\mathbf{E}[X] = \lim_{n\to\infty} \mathbf{E}[X_n]. \tag{8.28}$$

Para uma variável não negativa, esse limite pode ser finito ou infinito. A teoria garante que o valor obtido não depende da sequência particular de degraus usada para aproximar X, desde que a aproximação crescente seja feita da maneira adequada.

É útil comparar esse procedimento com uma imagem conhecida. Na integral de Riemann, aproximamos uma curva por retângulos cada vez mais estreitos. Aqui, aproximamos uma variável aleatória por funções em degraus cada vez mais precisas. Em ambos os casos, a quantidade procurada surge como limite de objetos elementares; o que muda é a forma de organizar esses objetos.

Para uma variável que pode assumir valores positivos e negativos, separamos as duas partes:

$$X^+ = \max\{X,0\}, \qquad X^- = \max\{-X,0\}.$$

Então

$$X = X^+ - X^-, \qquad |X| = X^+ + X^-.$$

Se $\mathbf{E}[X^+]$ e $\mathbf{E}[X^-]$ são finitas — de modo equivalente, se $\mathbf{E}[|X|] < \infty$ — dizemos que X é integrável e definimos

$$\boxed{\mathbf{E}[X] = \int_\Omega X \, d\mathbf{P} = \mathbf{E}[X^+] - \mathbf{E}[X^-].} \tag{8.29}$$

A notação $\int_\Omega X\, d\mathbf{P}$ enfatiza que estamos integrando os valores de X usando a probabilidade como regra para pesar conjuntos do espaço amostral. É essa a integral de Lebesgue que aparece naturalmente na teoria da probabilidade.

A fórmula é mais abstrata do que uma soma ou uma integral em dx, mas sua função é justamente evitar uma escolha artificial entre essas duas linguagens. Uma vez conhecida a distribuição de X, ela volta a assumir as formas concretas usadas ao longo do capítulo.

Uma fórmula, três regimes

No caso discreto, os eventos

$$A_x = \{\omega : X(\omega) = x\}$$

particionam o espaço amostral. Aplicando a definição para variáveis simples — e depois passando ao limite quando há infinitos valores possíveis — obtemos

$$\int_\Omega X\, d\mathbf{P} = \sum_x x\, \mathbf{P}(X = x).$$

Portanto, a soma usual não é uma definição rival da integral: ela é a forma que a integral assume quando a distribuição está concentrada em pontos.

Se X possui densidade f_X, a probabilidade é distribuída ao longo da reta segundo essa densidade. Nesse caso,

$$\int_\Omega X\, d\mathbf{P} = \int_{-\infty}^{\infty} x f_X(x)\, dx.$$

É a fórmula contínua que usamos desde o início do capítulo.

Finalmente, considere uma distribuição mista regular com átomos x_1,x_2, ..., de massas

$$p_i = \mathbf{P}(X = x_i),$$

e uma parte contínua descrita por uma densidade f. A massa total satisfaz

$$\sum_i p_i + \int_{-\infty}^{\infty} f(x)\, dx = 1,$$

e a mesma integral abstrata fornece

$$\boxed{\mathbf{E}[X] = \sum_i x_i p_i + \int_{-\infty}^{\infty} x f(x)\, dx,} \tag{8.30}$$

sempre que as contribuições envolvidas forem integráveis. A fórmula usada no exemplo do cache aparece, portanto, sem que precisemos criar uma nova noção de esperança para o caso misto.

Podemos resumir a situação assim:

$$\begin{aligned} \text{discreta} &\implies \mathbf{E}[X] = \sum_x x\,\mathbf{P}(X = x), \\ \text{com densidade} &\implies \mathbf{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\,dx, \\ \text{mista regular} &\implies \mathbf{E}[X] = \sum_i x_i p_i + \int_{-\infty}^{\infty} x f(x)\,dx. \end{aligned}$$

As três linhas são manifestações da mesma identidade:

$$\boxed{\mathbf{E}[X] = \int_\Omega X\,d\mathbf{P}.}$$

Essa é a principal mensagem desta seção. Para calcular, continuaremos usando a forma mais conveniente — soma, integral ordinária ou uma combinação das duas. A formulação unificada serve para mostrar que essas receitas não são acidentes separados. Por trás delas há uma única ideia de média: decompor a variável em níveis simples, pesar os conjuntos correspondentes pela probabilidade e passar ao limite.

Ao atravessar este capítulo

Esperança é calculada por soma ou integral, conforme o modelo. Linearidade, indicadores, condicionamento e funções geradoras convertem problemas longos em cálculos organizados. A formulação unificada $\mathbf{E}[X] = \int_\Omega X\,d\mathbf{P}$ mostra que os casos discreto, contínuo e misto pertencem à mesma construção.

8.9 Exercícios

Exercício 8.1 — Cálculo direto

Uma variável X assume os valores $-2{,}0{,}1{,}4$ com probabilidades $0{,}1{,}0{,}3{,}0{,}4{,}0{,}2$. Calcule $\mathbf{E}[X]$, $\mathbf{E}[X^2]$, $\mathrm{Var}(X)$ e $\mathbf{E}[(X-1)^2]$.

Exercício 8.2 — Indicadores

Sete cartas são retiradas sem reposição de um baralho comum. Usando variáveis indicadoras, determine o número esperado de ases e o número esperado de naipes representados na mão.

Exercício 8.3 — Função de uma variável

Se X tem densidade $3x^2$ em $(0{,}1)$, calcule $\mathbf{E}[X]$, $\mathrm{Var}(X)$, $\mathbf{E}[\log X]$ e $\mathbf{E}[1/X]$.

Exercício 8.4 — Esperança condicional

Uma urna é escolhida ao acaso: a primeira contém 2 bolas vermelhas e 4 azuis; a segunda, 5 vermelhas e 1 azul. Retiram-se duas bolas sem reposição da urna escolhida. Seja X o número de vermelhas. Calcule $\mathbf{E}[X \mid \text{urna}]$ e use a esperança total para obter $\mathbf{E}[X]$.

Exercício 8.5 — Função geradora de momentos

Se $X \sim \mathrm{Poisson}(\lambda)$, derive sua função geradora de momentos e use derivadas em 0 para recuperar $\mathbf{E}[X]$ e $\mathrm{Var}(X)$.

CAPÍTULO

9

Convergência

Até aqui, estudamos uma variável ou um vetor de cada vez. Agora aparece uma nova pergunta: o que significa dizer que uma sequência de variáveis aleatórias $X_1, X_2, \ldots$ se aproxima de alguma coisa?

A resposta não é única. Podemos pedir que a probabilidade de um erro grande fique pequena; podemos acompanhar cada realização ω e exigir convergência ao longo da trajetória; ou podemos olhar apenas para a forma das distribuições. Essas noções não dizem exatamente a mesma coisa, e entender a diferença entre elas é o fio do capítulo.

Começaremos com ferramentas que ainda não falam em limite — Markov e Chebyshev — mas já convertem informação sobre médias e variâncias em controle de probabilidades. Depois veremos as leis dos grandes números, Borel–Cantelli e a convergência quase certa. Por fim, o Teorema do Limite Central responderá uma pergunta mais fina: depois que uma média se estabiliza, qual é a forma das flutuações que ainda restam?

9.1 Desigualdades de Markov e Chebyshev

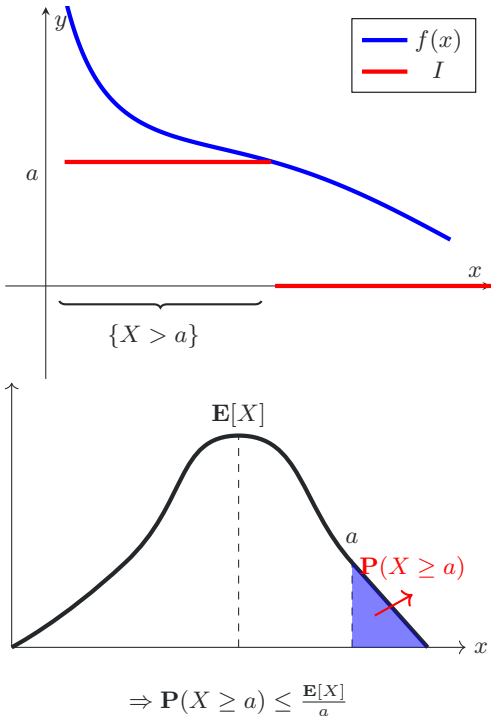
Suponha que não conhecemos a distribuição de X, apenas sua média. Ainda é possível dizer alguma coisa sobre a chance de X assumir um valor muito grande? Surpreendentemente, sim. Markov usa apenas a não negatividade e a esperança. Chebyshev acrescenta a variância e controla desvios em torno

da média. O preço dessa generalidade é que os limites podem ser grosseiros — e os exemplos desta seção mostrarão exatamente isso.

Proposição 9.1 — Desigualdade de Markov

Se X é uma variável aleatória que apresenta apenas valores não negativos então, para qualquer $a > 0$,

$$\mathbf{P}(X \geq a) \leq \frac{\mathbf{E}[X]}{a}$$



Demonstração. Para $a > 0$, suponha

$$I = \begin{cases} 1 & \text{se } X \geq a \\ 0 & \text{caso contrário} \end{cases}$$

e note que, como $X \geq 0$,

$$I \leq \frac{X}{a}$$

Calculando as esperanças da última desigualdade, obtemos

$$\mathbf{E}[I] \leq \frac{\mathbf{E}[X]}{a}$$

o que, como $\mathbf{E}[I] = \mathbf{P}(X \geq a)$, demonstra o resultado. ■

Como um corolário, obtemos

Proposição 9.2 — Desigualdade de Chebyshev

Se X é uma variável aleatória com média finita μ e variância σ^2, então, para qualquer valor $k > 0$,

$$\mathbf{P}(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2}$$

Demonstração. Como $(X-\mu)^2$ é uma variável aleatória não negativa, podemos aplicar a desigualdade de Markov (com $a = k^2$) para obter

$$\mathbf{P}\left((X - \mu)^2 \geq k^2\right) \leq \frac{\mathbf{E}\left[(X - \mu)^2\right]}{k^2}$$

Mas como $(X - \mu)^2 \geq k^2$ se e somente se $|X - \mu| \geq k$, a Equação 9.1 é equivalente a

$$\mathbf{P}(|X - \mu| \geq k) \leq \frac{\mathbf{E}\left[(X - \mu)^2\right]}{k^2} = \frac{\sigma^2}{k^2}$$

e a demonstração está completa. ■

A importância das desigualdades de Markov e Chebyshev reside no fato de que elas nos permitem estabelecer limitantes para probabilidades mesmo quando dispomos apenas da média, ou da média e variância, de uma distribuição de probabilidade. Se a distribuição completa fosse conhecida, poderíamos calcular as probabilidades exatas desejadas sem recorrer a tais limitantes.

O exemplo a seguir demonstra como usar a desigualdade de Markov e o quão imprecisa ela pode ser em alguns casos.

Exemplo 9.1.

Uma moeda é enviesada de tal forma que a probabilidade de sair cara em um único lançamento é de 25%, independentemente dos outros lançamentos. Suponha que a moeda seja lançada 20 vezes. Use a desigualdade de Markov para obter um limite superior para a probabilidade de obter pelo menos 16 caras.

Solução. O número de caras, X, tem distribuição Bin(20,0,25), logo:

$$\mathbf{E}[X] = np = 20 \times 0{,}25 = 5.$$

Aplicando a desigualdade de Markov, que afirma que para qualquer variável aleatória não negativa X e qualquer $a > 0$, temos $\mathbf{P}(X \geq a) \leq \frac{\mathbf{E}[X]}{a}$. Portanto:

$$\mathbf{P}(X \geq 16) \leq \frac{\mathbf{E}[X]}{16} = \frac{5}{16} \approx 0{,}3125.$$

Para comparar esse limite com a probabilidade real, calculamos:

$$\mathbf{P}(X \geq 16) = \sum_{k=16}^{20} \binom{20}{k} (0{,}25)^k (0{,}75)^{20-k} \approx 3{,}8 \times 10^{-6}.$$

Observamos que o limite fornecido pela desigualdade de Markov é muito maior do que a probabilidade real. Isso ocorre porque a desigualdade de Markov utiliza apenas o valor esperado de X e não considera a dispersão ou a forma da distribuição. Em casos onde o evento é altamente improvável em relação ao valor esperado, como neste exemplo, a desigualdade de Markov produz um limite superior pouco preciso.

◁

Exemplo 9.2.

Suponha que se saiba que o número de expressos vendidos por uma cafeteria durante um dia seja uma variável aleatória com média 70.

1. O que se pode dizer sobre a probabilidade de que as vendas de hoje sejam superiores a 100 expressos?
2. Se é conhecido que a variância das vendas diárias é igual a 100, então o que se pode dizer sobre a probabilidade de que as vendas de hoje estejam entre 50 e 90?

Solução. Seja X o número de expressos vendidos em um dia.

[a] Pela desigualdade de Markov,

$$\mathbf{P}(X > 100) \leq \frac{\mathbf{E}[X]}{100} = \frac{70}{100} = \frac{7}{10}$$

[b] Pela desigualdade de Chebyshev,

$$\mathbf{P}(|X - 70| \geq 20) \leq \frac{\sigma^2}{20^2} = \frac{100}{400} = \frac{1}{4}$$

Portanto,

$$\mathbf{P}(|X - 70| < 20) \geq 1 - \frac{1}{4} = \frac{3}{4}$$

Assim, a probabilidade de que as vendas de hoje estejam entre 50 e 90 é de pelo menos 0,75.

◁

Como a desigualdade de Chebyshev é válida para todas as distribuições da variável aleatória X, não podemos esperar que o limite da probabilidade esteja muito próximo da probabilidade real em muitos casos. Por exemplo, considere os Exemplos 9.3 e 9.4.

Exemplo 9.3.

Se X é uniformemente distribuída ao longo do intervalo (0,10), então, como $\mathbf{E}[X] = 5$ e $\text{Var}(X) = \frac{25}{3}$, resulta da desigualdade de Chebyshev que

$$\mathbf{P}(|X - 5| > 4) \leq \frac{25}{3(16)} \approx 0{,}52$$

enquanto o resultado exato é

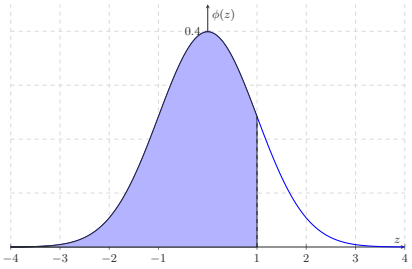
$$\mathbf{P}(|X - 5| > 4) = 0{,}20$$

Assim, embora a desigualdade de Chebyshev esteja correta, o limite superior que ela fornece não está particularmente próximo da probabilidade real.

◁

A função de distribuição cumulativa da normal padrão $\Phi(z)$ é definida por:

$$\Phi(z) = \mathbf{P}(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt$$



Exemplo 9.4.

Se X é uma variável aleatória normal com média μ e variância σ^2, a desigualdade de Chebyshev diz que

$$\mathbf{P}(|X - \mu| > 2\sigma) \leq \frac{1}{4}$$

enquanto a probabilidade real é dada por

$$\mathbf{P}(|X - \mu| > 2\sigma) = \mathbf{P}\left(\left|\frac{X - \mu}{\sigma}\right| > 2\right) = 2[1 - \Phi(2)] \approx 0{,}0456$$

◁

A desigualdade de Chebyshev é frequentemente utilizada como uma ferramenta teórica para demonstrar resultados.

Proposição 9.3

Se $\mathrm{Var}(X) = 0$, então

$$\mathbf{P}(X = \mathbf{E}[X]) = 1$$

Em outras palavras, as únicas variáveis aleatórias com variâncias iguais a 0são aquelas que são constantes com probabilidade 1.

Demonstração. Pela desigualdade de Chebyshev, temos, para qualquer $n \geq 1$,

$$\mathbf{P}\left(|X - \mu| > \frac{1}{n}\right) = 0$$

Fazendo $n \to \infty$ e usando a propriedade da continuidade das probabilidades, obtemos

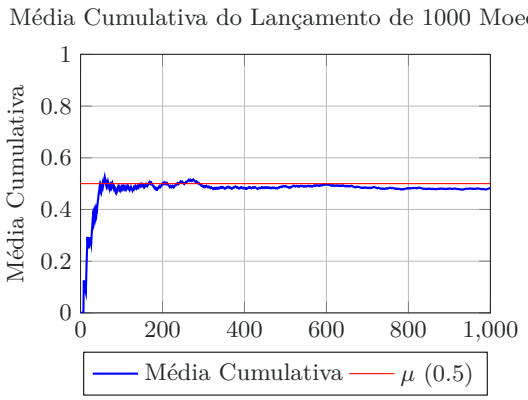
$$\begin{aligned} 0 = \lim_{n\to\infty} \mathbf{P}\left(|X - \mu| > \frac{1}{n}\right) &= \mathbf{P}\left(\lim_{n\to\infty}\left\{|X - \mu| > \frac{1}{n}\right)\right\} \\ &= \mathbf{P}(X \neq \mu) \end{aligned}$$

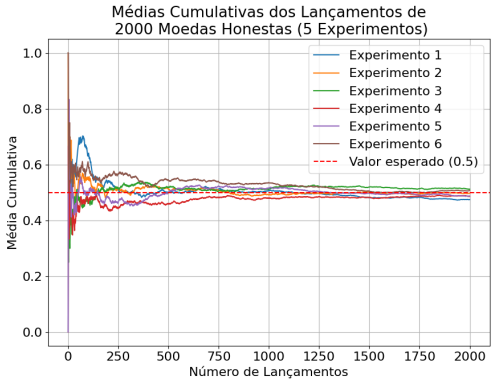
e o resultado está demonstrado. ∎

9.2 Convergência em probabilidade e Lei Fraca dos Grandes Números

Suponha que X_n seja o número de caras em n lançamentos de uma moeda com probabilidade p de cara. Esperamos que X_n/n fique perto de p. Mas o que significa "perto" quando a quantidade é aleatória?

Fixamos uma tolerância $\varepsilon > 0$ e olhamos para a chance de sair da faixa. A convergência em probabilidade exige que essa chance vá a zero. O conjunto de realizações ruins pode mudar com n; o que precisa desaparecer é a sua probabilidade.





Definição 9.1

Dizemos que a sequência de variáveis aleatórias $X_1, X_2, \ldots, X_n$ **converge em probabilidade** para a constante c se, para todo número positivo ε,

$$\lim_{n\to\infty} \mathbf{P}\left(|X_n - c| < \varepsilon\right) = 1$$

A média amostral é o exemplo mais importante desse mecanismo. Se repetimos uma medição sob as mesmas condições, cada observação ainda flutua, mas a média de muitas observações tende a ser mais estável. A Lei Fraca dos Grandes Números transforma essa ideia em uma afirmação precisa: para uma amostra i.i.d., a probabilidade de a média se afastar da média populacional por mais de uma tolerância fixa tende a zero.

Esse resultado responde a uma primeira pergunta — *a média se aproxima do valor correto?* — mas ainda não descreve a forma nem a escala das

flutuações que permanecem para um tamanho de amostra grande e finito. Essa segunda pergunta será respondida pelo Teorema do Limite Central.

Teorema 9.4 — Lei Fraca dos Grandes Números

Seja $X_1, X_2, \ldots$ uma sequência de variáveis aleatórias independentes e identicamente distribuídas, com média μ e variância finita σ^2. Então, para qualquer $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbf{P}\left(\left|\overline{X}_n - \mu\right| \geq \varepsilon\right) = 0$$

ou de modo equivalente

$$\lim_{n\to\infty} \mathbf{P}\left(\left|\overline{X}_n - \mu\right| < \varepsilon\right) = 1$$

Assim, $\overline{X}_n$ converge em probabilidade para μ.

Demonstração. Como

$$\mathbf{E}\left[\frac{X_1 + \cdots + X_n}{n}\right] = \mu \quad \text{e} \quad \operatorname{Var}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{\sigma^2}{n}$$

resulta da desigualdade de Chebyshev que

$$\mathbf{P}\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}$$

e o resultado está demonstrado. ∎

A Lei Fraca fornece, portanto, uma noção de consistência: médias de observações independentes tendem ao valor esperado. Comecemos pelo exemplo que motivou a seção.

Exemplo 9.5.

Seja X_n uma variável aleatória binomial com probabilidade de sucesso p e número de tentativas n. Mostre que X_n/n converge em probabilidade para p.
Solução. Vimos que podemos escrever X_n como $\sum_{i=1}^{n} Y_i$, onde $Y_i = 1$ se a i-ésima tentativa resulta em sucesso, e $Y_i = 0$ caso contrário. Então

$$\frac{X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Além disso, $\mathbf{E}(Y_i) = p$ e $\mathrm{Var}(Y_i) = p(1-p)$. As condições do Teorema 9.4 são então cumpridas com $\mu = p$ e concluímos que, para qualquer $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbf{P}\left(\left|\frac{X_n}{n} - p\right| \geq \varepsilon\right) = 0$$

◁

Exemplo 9.6 — Monte Carlo: aproximando π.

Uma aplicação especialmente importante da Lei dos Grandes Números é o método de Monte Carlo. Sorteie, de forma independente, pontos (U_i,V_i) uniformemente no quadrado unitário $(0,1)^2$ e defina

$$I_i = \mathbb{1}_{\{U_i^2+V_i^2\leq 1\}}.$$

Como a área do quarto de disco unitário é $\pi/4$, temos $\mathbf{E}[I_i] = \pi/4$. Logo

$$\widehat{\pi}_n = 4\bar{I}_n = \frac{4}{n}\sum_{i=1}^{n} I_i$$

é um estimador de π. Pela Lei Fraca,

$$\widehat{\pi}_n \xrightarrow{P} \pi.$$

Assim, aumentar o número de pontos torna a aproximação consistente. Mas a Lei dos Grandes Números, sozinha, não diz com precisão quanto erro devemos esperar para um dado n. Voltaremos a este mesmo exemplo depois do Teorema do Limite Central.

◁

> A demonstração apresentada é válida para variáveis aleatórias X_k não correlacionadas. Isso por que o único lugar onde usamos a independência foi para afirmar que as somas das variâncias é a variância da soma, e esse resultado ainda é válido para variáveis não correlacionadas duas a duas.

A demonstração anterior generaliza de maneira óbvia para vetores aleatórios. Além disso temos as seguintes propriedades:

Teorema 9.5

Suponha que X_n converge em probabilidade para μ_1 e Y_n converge em probabilidade para μ_2. Então, as seguintes afirmações também são verdadeiras.

1. $X_n + Y_n$ converge em probabilidade para $\mu_1 + \mu_2$.
2. $X_n Y_n$ converge em probabilidade para $\mu_1 \mu_2$.
3. X_n / Y_n converge em probabilidade para μ_1/μ_2, desde que $\mu_2 \neq 0$.
4. $\sqrt{X_n}$ converge em probabilidade para $\sqrt{\mu_1}$, desde que $\mathbf{P}(X_n \geq 0) = 1$.

Exemplo 9.7.

Suponha que $X_1, X_2, \ldots$ sejam variáveis aleatórias independentes e identicamente distribuídas, com $\mathbf{E}[X_i] = \mu$ e $\mathbf{E}[X_i^4] < \infty$. Seja

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$$

a variância amostral. Mostre que S_n^2 converge em probabilidade para $\mathrm{Var}(X_1)$.
Solução. A identidade

$$S_n^2 = \frac{n}{n-1} \left(\frac{1}{n} \sum_{i=1}^{n} X_i^2 - \overline{X}_n^2 \right)$$

isola duas médias amostrais. Pela Lei Fraca,

$$\overline{X}_n \xrightarrow{P} \mu \qquad \text{e} \qquad \frac{1}{n} \sum_{i=1}^{n} X_i^2 \xrightarrow{P} \mathbf{E}[X_1^2].$$

Na segunda convergência usamos $\mathbf{E}[X_1^4] < \infty$, que garante variância finita para X_1^2. Como $n/(n-1) \to 1$, as propriedades da convergência em probabilidade dão

$$S_n^2 \xrightarrow{P} \mathbf{E}[X_1^2] - \mu^2 = \mathrm{Var}(X_1).$$

Assim, a variância amostral também é um estimador consistente.

◁

9.3 Lema de Borel-Cantelli

Convergência em probabilidade responde a uma pergunta por vez: para um dado n, quão provável é um erro grande? Mas uma trajetória pode cometer erros raros e ainda assim cometê-los infinitas vezes. Como distinguir essas duas situações? Os lemas de Borel–Cantelli transformam somas de probabilidades em informação sobre eventos que se repetem infinitamente.

Definição 9.2 — Infinitas vezes e eventualmente sempre

Para eventos E_1,E_2,...,

$$\limsup E_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} E_k = \{\omega : \omega \text{ pertence a infinitos } E_n\}.$$

Já $\liminf E_n$ reúne os resultados que pertencem a todos os E_n a partir de algum índice.

Se E_n é o evento "o sistema falha no dia n", então $\limsup E_n$ significa falhar em infinitos dias. É esse tipo de repetição que os lemas de Borel–Cantelli controlam.

Teorema 9.6 — Lema de Borel-Cantelli

Sejam $(\Omega, \mathcal{F}, \mathbf{P})$ um espaço de probabilidade e $(A_n)_n$ uma sequência de eventos.

1. Se $\sum_{n=1}^{\infty} \mathbf{P}(A_n) < \infty$, então $\mathbf{P}(A_n \text{ i.v.}) = 0$.
2. Se $\sum_{n=1}^{\infty} \mathbf{P}(A_n) = \infty$ e os eventos $(A_n)_n$ são independentes, então $\mathbf{P}(A_n \text{ i.v.}) = 1$.

Demonstração. Para a primeira parte, observe inicialmente que $\bigcup_{k=n}^{\infty} A_k \downarrow \{A_n \text{ i.v.}\}$. Portanto,

$$\mathbf{P}(\{A_n \text{ i.v.}\}) = \mathbf{P}\left(\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k\right).$$

Assim,

$$\mathbf{P}(\{A_n \text{ i.v.}\}) = \lim_{n\to\infty} \mathbf{P}\left(\bigcup_{k=n}^{\infty} A_k\right).$$

Pela desigualdade da soma:

$$\mathbf{P}\left(\bigcup_{k=n}^{\infty} A_k\right) \leq \sum_{k=n}^{\infty} \mathbf{P}(A_k).$$

Como $\sum_{n=1}^{\infty} \mathbf{P}(A_n) < \infty$, segue que o limite é zero. Assim, $\mathbf{P}(\{A_n \text{ i.v.}\}) = 0$.

Para a segunda parte, consideremos o evento $\{A_n \text{ i.v.}\}^c$. Pela Lei de De Morgan, temos:

$$\{A_n \text{ i.v.}\}^c = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k^c.$$

Note que $\bigcap_{k=n}^{\infty} A_k^c \uparrow \{A_n \text{ i.v.}\}^c$ conforme $n \to \infty$. Logo:

$$\mathbf{P}(\{A_n \text{ i.v.}\}^c) = \lim_{n\to\infty} \mathbf{P}\left(\bigcap_{k=n}^{\infty} A_k^c\right).$$

Devido à independência dos eventos $(A_n)_n$:

$$\mathbf{P}\left(\bigcap_{k=n}^{\infty} A_k^c\right) = \prod_{k=n}^{\infty} (1 - \mathbf{P}(A_k)).$$

Utilizando a cota superior $1 - x \leq e^{-x}$ para todo $x \in \mathbb{R}$, temos:

$$\prod_{k=n}^{\infty} (1 - \mathbf{P}(A_k)) \leq \exp\left(-\sum_{k=n}^{\infty} \mathbf{P}(A_k)\right).$$

Como $\sum_{n=1}^{\infty} \mathbf{P}(A_n) = +\infty$, o termo $\exp(-\sum_{k=n}^{\infty} \mathbf{P}(A_k)) \to 0$. Portanto, $\mathbf{P}(\{A_n \text{ i.v.}\}^c) = 0$, o que implica que $\mathbf{P}(\{A_n \text{ i.v.}\}) = 1$. ∎

Se pensarmos em n como tempo ou algo análogo, o que o resultado acima nos diz é que se uma sequência de eventos tem probabilidade decaindo muito rápido, então eventualmente deixaremos de observar essa sequência de eventos. De fato, note que

$$\left(\limsup_n A_n\right)^c = \left(\bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} A_n\right)^c = \left(\bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} A_n^c\right) = \liminf_n A_n^c,$$

de modo que se $\mathbf{P}(A_n \text{ i.v.}) = 0$, então $\mathbf{P}(A_n^c \text{ q.s.}) = 1$.

Exemplo 9.8.

Podemos aplicar o Lema de Borel-Cantelli a uma situação interessante em que se pode lucrar com uma série de jogos que, individualmente, têm valor esperado 0. Esse fato é contra-intuitivo

Considere uma sequência infinita de jogos, onde no n-ésimo jogo se perde 2^n dólares com probabilidade $\frac{1}{2^n+1}$ e se ganha um dólar com probabilidade $\frac{2^n}{2^n+1}$. Logo pode-se perder uma grande quantia de dinheiro ás vezes, e ganhar um pouco de dinheiro a maior parte do tempo. Podemos ver que a esperança de qualquer jogo individual é

$$\frac{-2^n}{2^n+1} + \frac{2^n}{2^n+1} = 0$$

Aqui é onde o Lema de Borel-Cantelli torna as coisas interessantes. Seja E_n o evento de perder no n-ésimo jogo.

Temos que

$$\sum_{n=1}^{\infty} \mathbf{P}(E_n) = \sum_{n=1}^{\infty} \frac{1}{2^n+1} \leq \sum_{n=1}^{\infty} \frac{1}{2^n} = 1 < \infty$$

e, consequentemente, é finita.

Como a soma das probabilidades é finita, pelo Lema de Borel-Cantelli. temos que com probabilidade 1 a pessoa só perderá um número finito de vezes. Portanto, a quantidade de dinheiro ganha pelo jogador vai para o infinito, já que ele ganha infinitas vezes e perde apenas um número finito de vezes!

◁

Exemplo 9.9 — Passeio Aleatório Assimétrico.

Considere um processo aleatório em $\mathbb{Z}$ onde uma partícula inicia em $x = 0$ e a cada instante n salta uma unidade para a direita com probabilidade $p \in (0,1)$ ou uma unidade para a esquerda com probabilidade $1-p$. Denote por $S_n \in \mathbb{Z}$ a posição da partícula no instante $n \geq 0$.

O processo S_n é conhecido na literatura como passeio aleatório simples, e no caso em que $p \neq 1/2$, dizemos que o passeio é assimétrico.

Sejam $Y_1, Y_2, \ldots$ incrementos independentes, com $\mathbf{P}(Y_k = 1) = p$ e $\mathbf{P}(Y_k = -1) = 1-p$. Definimos $S_0 = 0$ e $S_n = Y_1 + \cdots + Y_n$.

Assim para cada $n \geq 1$

$$\mathbf{P}(S_n = S_{n-1} + 1) = p,$$

enquanto

$$\mathbf{P}(S_n = S_{n-1} - 1) = 1 - p,$$

como gostaríamos.

Estamos interessados agora em estudar o retorno de tal processo à origem do espaço. Mais especificamente, queremos saber com que frequência o processo retorna para o ponto onde começou o passeio. Em outras palavras, estamos interessados na sequência de eventos $\{S_n = 0\} := \{\omega : S_n(\omega) = 0\}$.

Para que tenhamos $S_n = 0$ é necessário que o total de passos para a direita até o instante n seja igual ao total de passos para a esquerda. Com isso concluímos, antes de mais nada, que n deve ser par.

Fica como exercício para o leitor justificar agora que

$$\mathbf{P}(S_{2n} = 0) = \binom{2n}{n} p^n (1-p)^n.$$

Usando a fórmula de Stirling,[1] encontramos que

$$\mathbf{P}(S_{2n} = 0) \approx \frac{\sqrt{4\pi n}\, 4^n\, n^{2n}\, e^{-2n}}{(\sqrt{2\pi n}\, n^n\, e^{-n})^2} p^n (1-p)^n = \frac{[4p(1-p)]^n}{\sqrt{\pi n}}.$$

Note agora que se $p \neq 1/2$ então $r := 4p(1-p) < 1$ e portanto

$$\sum_{n=1}^{\infty} \mathbf{P}(S_{2n} = 0) < \infty,$$

pois, para n suficientemente grande, esses termos são limitados por uma série geométrica de razão estritamente menor que 1.

E portanto $\mathbf{P}(S_{2n} = 0 \text{ i.v.}) = 0$, mostrando que eventualmente o processo visita a origem do espaço uma última vez, para nunca mais retornar.

◁

Exemplo 9.10.

Considere o exemplo de infinitos lançamentos independentes de uma moeda não-viciada.

Intuitivamente, esperamos que o total de caras observados, assim como o total de coroas, sejam infinitos. Para mostrar isso, tome o evento

$$A_n = \{\text{o } n\text{-ésimo lançamento foi cara}\} = \{\omega \in \Omega : \omega_n = 1\}.$$

[1] A fórmula de Stirling é uma aproximação assintótica para o fatorial de um número n, muito útil em análise combinatória e teoria de números. Ela é dada por:

$$n! \sim \sqrt{2\pi n} \left(\frac{n}{e}\right)^n.$$

Já sabemos que os eventos $A_n, n \geq 1$ são independentes e que $\mathbf{P}(A_n) = 1/2$. Segue então que

$$\sum_{n=1}^{\infty} \mathbf{P}(A_n) = \infty,$$

e pelo Lema de Borel-Cantelli, $\mathbf{P}(A_n \text{ i.v.}) = 1$. Ou seja, observaremos infinitas caras com probabilidade 1.

De modo análogo podemos mostrar que $\mathbf{P}(A_n^c \text{ i.v.}) = 1$.

◁

Exemplo 9.11.

O **teorema do macaco infinito** afirma que um macaco pressionando teclas aleatoriamente em uma máquina de escrever por uma quantidade infinita de tempo **quase certamente** digitará qualquer texto dado, incluindo as obras completas de William Shakespeare. Na verdade, o macaco quase certamente digitaria qualquer texto finito possível um número infinito de vezes.

Solução. Considere uma sequência infinita de caracteres produzida a partir de um alfabeto finito de n letras, onde cada letra é escolhida de forma independente e uniformemente distribuída. Seja S uma cadeia fixa de comprimento m pertencente ao mesmo alfabeto. Defina E_k como o evento em que a substring de comprimento m que começa na posição k é exatamente a cadeia S. Então, com probabilidade 1, ocorrerão infinitos eventos E_k ao longo da sequência.

A probabilidade de que a substring iniciando na posição k seja exatamente S é:

$$\mathbf{P}(E_k) = \left(\frac{1}{n}\right)^m,$$

pois cada uma das m posições deve corresponder à letra específica de S.

Os eventos E_k são independentes porque a escolha de cada letra na sequência é feita de forma independente. Para aplicar o Segundo Lema de Borel-Cantelli, consideremos os eventos E_{mj+1} para $j = 0, 1, 2, \ldots$. Esses eventos são independentes entre si, já que estão separados por m posições e não compartilham letras em comum.

Calculamos a soma das probabilidades desses eventos:

$$\sum_{j=0}^{\infty} \mathbf{P}(E_{mj+1}) = \sum_{j=0}^{\infty} \left(\frac{1}{n}\right)^m = \infty,$$

pois estamos somando infinitas vezes um valor positivo constante.

Pelo Segundo Lema de Borel-Cantelli, se a soma das probabilidades de eventos independentes diverge, então a probabilidade de que ocorram infinitamente muitos desses eventos é 1.

Isso implica que ocorrerão infinitos eventos E_{mj+1} com probabilidade 1. Como os eventos E_k incluem todos os E_{mj+1}, concluímos que infinitos E_k ocorrerão quase certamente.

◁

9.4 Convergência quase certa e Lei Forte dos Grandes Números

A convergência em probabilidade permite que o conjunto de trajetórias ruins mude com n. E se quisermos algo mais forte: que, para quase toda realização, a sequência inteira acabe se aproximando do limite e permaneça próxima dele? Essa é a convergência quase certa.

Definição 9.3 — Convergência Quase Certa

Uma sequência de variáveis aleatórias $X_1, X_2, X_3, \cdots$ converge quase certamente para uma variável aleatória X, indicada por $X_n \xrightarrow{\text{q.c.}} X$, se

$$\mathbf{P}\left(\left\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}\right) = 1.$$

Em outras palavras, o conjunto de todos os resultados possíveis nos quais a sequência (X_n) não converge para X tem probabilidade zero. Essa forma de convergência é muito forte, pois exige que a convergência ocorra quase sempre, exceto em um conjunto de casos excepcionalmente raro (de probabilidade zero).

Esse tipo de convergência é mais forte do que a *convergência em probabilidade*. Enquanto a convergência em probabilidade exige que a probabilidade de $|X_n - X| > \varepsilon$ tenda a zero para qualquer $\varepsilon > 0$, a convergência quase certa implica que, para quase todos os resultados possíveis, a sequência efetivamente converge ponto a ponto.

É direto ver que a convergência quase certamente implica convergência em probabilidade.

O Lema de Borel-Cantelli é útil para avaliar a convergência quase certa. Para uma sequência de variáveis aleatórias $\{X_n\}$ e uma variável aleatória limite X, suponha que, para $\epsilon > 0$, o evento $A_n(\epsilon)$ seja dado por:

$$A_n(\epsilon) \equiv \{\omega : |X_n(\omega) - X(\omega)| > \epsilon\}.$$

O Lema de Borel-Cantelli afirma:

1. Se, para todo $m = 1,2,\ldots,$

$$\sum_{n=1}^{\infty} \mathbf{P}\left(|X_n - X| > \frac{1}{m}\right) < \infty,$$

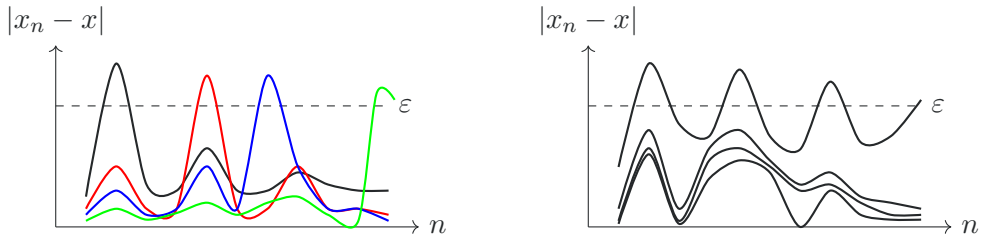
então $X_n \xrightarrow{\text{q.c.}} X$.

2. Reciprocamente, para um $\epsilon > 0$, se

$$\sum_{n=1}^{\infty} \mathbf{P}(A_n(\epsilon)) = \sum_{n=1}^{\infty} \mathbf{P}(|X_n - X| > \epsilon) = \infty,$$

e os eventos $A_n(\epsilon)$ forem independentes, então esses desvios ocorrem infinitas vezes com probabilidade 1. Logo X_n não converge quase certamente para X.

Na Fig. 9.1, plotamos a diferença $|x_n - x|$ como uma função de n, onde, para simplicidade, as sequências são representadas como curvas. Cada curva representa, assim, uma sequência particular $|x_n(t) - x(t)|$. A convergência em probabilidade significa que, para um $n > n_0$ específico, apenas uma pequena porcentagem dessas curvas terá ordenadas que excedem ε (Fig 9.1a). É claro que é possível que nem uma dessas curvas permaneça menor que ε para cada $n > n_0$. A convergência quase certamente, por outro lado, exige que quase todas as curvas estejam abaixo de ε para todo $n > n_0$ (Fig. 9.1b).



(a) Convergência em probabilidade (b) Convergência quase-certa

Figura 9.1: Representações das oscilações de $|x_n - x|$ em relação ao número de iterações n.

Considere $X_1, X_2, \ldots$ variáveis aleatória independentes e identicamente distribuídas com esperança finita μ. Seja $S_n := X_1 + \cdots + X_n$. A Lei Fraca dos Grandes Números afirma que

$$\frac{S_n}{n} \to \mu \text{ em probabilidade.}$$

Por outro lado, a Lei Forte dos Grandes Números que provaremos a seguir afirma que

$$\frac{S_n}{n} \to \mu \text{ q.c.}$$

Apresentaremos uma demonstração simples da Lei Forte dos Grandes Números sob o pressuposto de que o quarto momento é finito.

Teorema 9.7 — Lei Forte dos Grandes Números

Sejam $X_1, X_2, \ldots$ variáveis aleatórias i.i.d. com esperança μ e assuma $\mathbf{E}\left[X_k^4\right] < \infty$. Então, $S_n = X_1 + X_2 + \cdots + X_n$ satisfaz

$$\frac{S_n}{n} \to \mu \text{ quase certamente.}$$

$$\begin{cases} \text{Lei fraca : } \lim\limits_{n\to\infty} \mathbf{P}\left(|\bar{X}_n - \mathbf{E}[X]| < \varepsilon\right) = 1 \\ \text{Lei forte : } \mathbf{P}\left(\lim\limits_{n\to\infty} |\bar{X}_n - \mathbf{E}[X]| = 0\right) = 1 \end{cases}$$

Demonstração. Primeiro assumimos que a esperança $\mu = 0$.

De fato, fazendo $X'_k = X_k - \mu$, temos

$$\frac{\sum_{k=1}^n X'_k}{n} = \frac{\sum_{k=1}^n (X_k - \mu)}{n} = \frac{S_n}{n} - \mu,$$

e assim o resultado geral segue do caso $\mu = 0$.

Lembrando, nosso objetivo agora é mostrar que para qualquer $\varepsilon > 0$,

$$\mathbf{P}\left(\left|\frac{S_n}{n}\right| > \varepsilon \text{ i.v.}\right) = 0. \tag{9.1}$$

Isso será feito a partir da desigualdade de Markov aplicada à quarta potência, usando o fato de que as variáveis X_k possuem quarto momento finito.

Para começar, calculemos primeiro o quarto momento de S_n.

$$\mathbf{E}\Big[S_n^4\Big] = \mathbf{E}\left[(\sum_{k=1}^n X_k)^4\right] = \sum_{1\leq i,j,k,\ell\leq n} \mathbf{E}[(X_i X_j X_k X_\ell)]$$

Observe que se i, j são distintos, então, pela independência, $\mathbf{E}\big[X_i^3 X_j\big] = \mathbf{E}\big[X_i^3\big]\,\mathbf{E}[X_j] = 0$ (lembre que $\mu = 0$) . Da mesma forma, temos $\mathbf{E}\big[X_i^2 X_j\big] = 0$ e $\mathbf{E}[X_i X_j X_k X_\ell] = 0$, se i, j, k, ℓ forem diferentes.

Desta forma os únicos termos que não são zero são da forma:

$$\mathbf{E}\Big[X_k^4\Big] \quad \text{e} \quad \mathbf{E}\Big[X_j^2 X_k^2\Big].$$

Agora uma contagem simples nos mostra que existem n termos da forma $\mathbf{E}[X_k^4]$, e outros $\binom{n}{2}\binom{4}{2} = 3n(n-1)$ termos da forma $\mathbf{E}\Big[X_j^2 X_k^2\Big]$.

Como as variáveis envolvidas são identicamente distribuídas, então

$$\mathbf{E}[X_k^4] = \mathbf{E}[X_1^4] \qquad \text{e} \qquad \mathbf{E}\Big[X_j^2 X_k^2\Big] = \mathbf{E}\Big[X_1^2 X_2^2\Big].$$

Assim, se $C = \max\{\mathbf{E}\big[X_1^2 X_2^2\big], \mathbf{E}\big[X_1^4\big]\}$, então

$$\mathbf{E}\left[S_n^4\right] = n\mathbf{E}[X_1^4] + (3n^2 - 3n)\mathbf{E}\Big[X_1^2 X_2^2\Big] \leq C \cdot (3n^2 - 2n) \leq 3Cn^2.$$

Pela desigualdade de Markov, temos então que

$$\begin{aligned}\mathbf{P}\left(\left|\frac{S_n}{n}\right| > \varepsilon\right) &= \mathbf{P}(S_n^4 > \varepsilon^4 n^4) \\ &\leq \frac{\mathbf{E}[S_n^4]}{\varepsilon^4 n^4}\end{aligned}$$

$$\leq \frac{3C}{\varepsilon^4 n^2}$$

E como a série $\sum_{n=1}^{\infty} \frac{3C}{\varepsilon^4 n^2}$ converge, segue do Lema de Borel-Cantelli que

$$\mathbf{P}\left(\left|\frac{S_n}{n}\right| > \varepsilon \text{ i.v.}\right) = 0.$$

E logo temos a convergência quase certa. ∎

9.5 Convergência em distribuição e Teorema do Limite Central

As noções anteriores comparam variáveis definidas num mesmo experimento. Para descrever as flutuações de uma média, porém, queremos olhar para a forma de sua distribuição depois de uma reescala. A Lei dos Grandes Números diz que $\overline{X}_n - \mu$ tende a zero; se multiplicarmos esse erro por $\sqrt{n}$, a flutuação deixa de colapsar e pode ter uma distribuição limite não trivial.

Isso nos leva a uma noção mais fraca de convergência: em vez de comparar as variáveis realização por realização, comparamos suas distribuições. Perguntamos se as funções de distribuição de X_n se aproximam da função de distribuição de alguma variável X. Essa é a convergência em distribuição.

Definição 9.4

Seja X_n uma variável aleatória com função de distribuição $F_n(x)$. Seja X uma variável aleatória com função de distribuição $F(x)$. Se

$$\lim_{n\to\infty} F_n(x) = F(x)$$

em todo ponto x para o qual $F(x)$ é contínua, então dizemos que X_n converge em distribuição para X, e $F(x)$ é chamada de função de distribuição limite de X_n.

O conceito de convergência em distribuição baseia-se na seguinte intuição: duas variáveis aleatórias são "próximas uma da outra" se suas funções de distribuição forem "próximas uma da outra".

É importante notar que a **convergência em distribuição** envolve apenas as funções de distribuição das variáveis aleatórias da sequência, e essas variáveis

não precisam ser definidas no mesmo espaço amostral. Em contraste, os modos de convergência que discutimos anteriormente—**convergência quase certa** e **convergência em probabilidade**—exigem que todas as variáveis da sequência sejam definidas no mesmo espaço amostral.

Exemplo 9.12.

O requisito de continuidade, mencionado na definição acima, se justifica para evitar algumas anomalias. Por exemplo, para $n \geq 1$, seja $X_n = \frac{1}{n}$ e $X = 0$, para todo Ω. Parece aceitável que deveríamos ter convergência de X_n para X, qualquer que fosse o modo de convergência. Observe que

$$F_n(x) = \begin{cases} 0, & x < \frac{1}{n}; \\ 1, & x \geq \frac{1}{n}, \end{cases}$$

e

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

Portanto, como $\lim_{n\to\infty} F_n(0) = 0 \neq F(0) = 1$, não temos $\lim_{n\to\infty} F_n(x) = F(x)$ para todo $x \in \mathbb{R}$. Dessa forma, se houvesse a exigência de convergência em todos os pontos, não teríamos convergência em distribuição. Entretanto, note que para $x \neq 0$, temos $\lim_{n\to\infty} F_n(x) = F(x)$ e, como o ponto 0 não é de continuidade de F, concluímos que $X_n \xrightarrow{d} X$.

◁

Muitas vezes, é mais fácil encontrar distribuições limite trabalhando com as funções geradoras de momentos. O Teorema 9.8 fornece a relação entre a convergência das funções de distribuição e a convergência das funções geradoras de momentos.

Teorema 9.8

Sejam X_n e X variáveis aleatórias cujas funções geradoras de momentos $M_n(t)$ e $M(t)$ são finitas em um mesmo intervalo aberto ao redor de 0. Se

$$\lim_{n\to\infty} \mathrm{M}_n(t) = M(t)$$

para todo t nesse intervalo, então X_n converge em distribuição para X.

A prova do Teorema 9.8 está além do escopo deste texto.

Exemplo 9.13.

Seja X_n uma variável aleatória binomial com n ensaios e probabilidade p de sucesso em cada ensaio. Se n tende ao infinito e p tende a zero com np permanecendo fixo, mostre que X_n converge em distribuição para uma variável aleatória de Poisson.

Solução. Resolveremos usando funções geradoras de momentos e o Teorema 9.8.

Sabemos que a função geradora de momentos de X_n - ou seja, $\mathrm{M}_n(t)$ - é dada por

$$\mathrm{M}_n(t) = \left(q + pe^t\right)^n$$

onde $q = 1 - p$. Isso pode ser reescrito como

$$\mathrm{M}_n(t) = \left[1 + p\left(e^t - 1\right)\right]^n.$$

Fazendo $np = \lambda$ e substituindo em $\mathrm{M}_n(t)$, obtemos

$$\mathrm{M}_n(t) = \left[1 + \frac{\lambda}{n}\left(e^t - 1\right)\right]^n.$$

Agora, vamos tomar o limite dessa expressão conforme n tende ao infinito.

$$\lim_{n\to\infty}\left(1 + \frac{k}{n}\right)^n = e^k$$

Finalmente fazendo $k = \lambda\left(e^t - 1\right)$, temos

$$\lim_{n\to\infty} \mathrm{M}_n(t) = \exp\left[\lambda\left(e^t - 1\right)\right]$$

Reconhecemos a expressão do lado direito como a função geradora de momentos para a variável aleatória de Poisson. Portanto, segue do Teorema 9.8 que X_n converge em distribuição para uma variável aleatória de Poisson.

◁

Exemplo 9.14.

No monitoramento da poluição, um experimentador coleta um pequeno volume de água e conta o número de bactérias na amostra. Diferentemente dos casos anteriores, aqui dispomos de apenas uma única observação. Para aproximar a distribuição de probabilidade das contagens, podemos considerar o volume como uma grandeza crescente, o que nos permite tratar situações em que a média do número de bactérias (a intensidade do processo) é elevada.

Seja X a variável aleatória que representa a contagem de bactérias por centímetro cúbico de água, supondo que X siga uma distribuição de Poisson com média λ. Nosso objetivo é aproximar a distribuição de X quando λ é grande, mostrando que

$$Y = \frac{X - \lambda}{\sqrt{\lambda}}$$

converge em distribuição para uma variável aleatória normal padrão conforme $\lambda \to \infty$.

Especificamente, suponha que a contagem máxima aceitável de bactérias por centímetro cúbico em um determinado suprimento de água seja 110, e que $\lambda = 100$. Queremos aproximar $\mathbf{P}(X \leq 110)$.

Solução. Consideramos inicialmente a função geradora de momentos de Y e estudamos seu comportamento à medida que $\lambda \to \infty$, aplicando o Teorema 9.8. A função geradora de momentos de X, denotada por $\mathrm{M}_X(t)$, é

$$\mathrm{M}_X(t) = e^{\lambda(e^t - 1)}.$$

Desta forma, a função geradora de momentos de Y é

$$\begin{aligned} \mathrm{M}_Y(t) &= e^{-t\sqrt{\lambda}} \mathrm{M}_X\left(\frac{t}{\sqrt{\lambda}}\right) \\ &= e^{-t\sqrt{\lambda}} \exp\left[\lambda\left(e^{t/\sqrt{\lambda}} - 1\right)\right]. \end{aligned}$$

Expandindo $e^{t/\sqrt{\lambda}}$, obtemos

$$e^{t/\sqrt{\lambda}} - 1 = \frac{t}{\sqrt{\lambda}} + \frac{t^2}{2\lambda} + \frac{t^3}{6\lambda^{3/2}} + \cdots$$

Substituindo na função geradora de momentos de Y:

$$\begin{aligned} \mathrm{M}_Y(t) &= \exp\left[-t\sqrt{\lambda} + \lambda\left(\frac{t}{\sqrt{\lambda}} + \frac{t^2}{2\lambda} + \frac{t^3}{6\lambda^{3/2}} + \cdots\right)\right] \\ &= \exp\left(\frac{t^2}{2} + \frac{t^3}{6\sqrt{\lambda}} + \cdots\right). \end{aligned}$$

À medida que $\lambda \to \infty$, os termos após $\frac{t^2}{2}$ tendem a zero, de modo que

$$\lim_{\lambda \to \infty} \mathrm{M}_Y(t) = e^{t^2/2},$$

a função geradora de momentos de uma variável aleatória normal padrão.

Dessa forma, para $\lambda = 100$:

$$\mathbf{P}(X \leq 110) \approx \mathbf{P}\left(Z \leq \frac{110{,}5 - 100}{10}\right) = \Phi(1{,}05),$$

onde 110,5 é a correção de continuidade. Assim, $\mathbf{P}(X \leq 110) \approx 0{,}8531$.

◁

O exemplo anterior já exibe uma aproximação normal. O Teorema do Limite Central mostra que esse fenômeno é muito mais geral e, em particular, completa a história iniciada pela Lei dos Grandes Números. Para uma amostra i.i.d. com média μ e variância σ^2,

$$\overline{X}_n \xrightarrow{P} \mu$$

diz que a média se estabiliza. O TLC acrescenta a informação que faltava: depois de ampliar o erro pela escala natural $\sqrt{n}$,

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$$

tem distribuição aproximadamente normal padrão para n grande.

Assim, a Lei dos Grandes Números responde *onde* a média se concentra; o Teorema do Limite Central descreve *como* ela flutua em torno desse valor. Em particular, o erro típico de uma média é da ordem de $1/\sqrt{n}$. Essa é a ponte que leva das leis-limite à aproximação de probabilidades, ao método de Monte Carlo e, mais adiante, aos intervalos de confiança.

Dizemos também que a média padronizada é assintoticamente normal. A versão abaixo é formulada para variáveis i.i.d. com variância finita; existem extensões para situações mais gerais.

Teorema 9.9 — Teorema do Limite Central

Seja $X_1, X_2, \ldots$ uma sequência de variáveis aleatórias independentes e identicamente distribuídas, cada uma com média μ e variância σ^2. Então, a distribuição de

$$\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

tende à distribuição normal padrão quando $n \to \infty$. Isto é, para

$-\infty < a < \infty$,

$$\mathbf{P}\left(\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \leq a\right) \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2} dx \text{ quando } n \to \infty$$

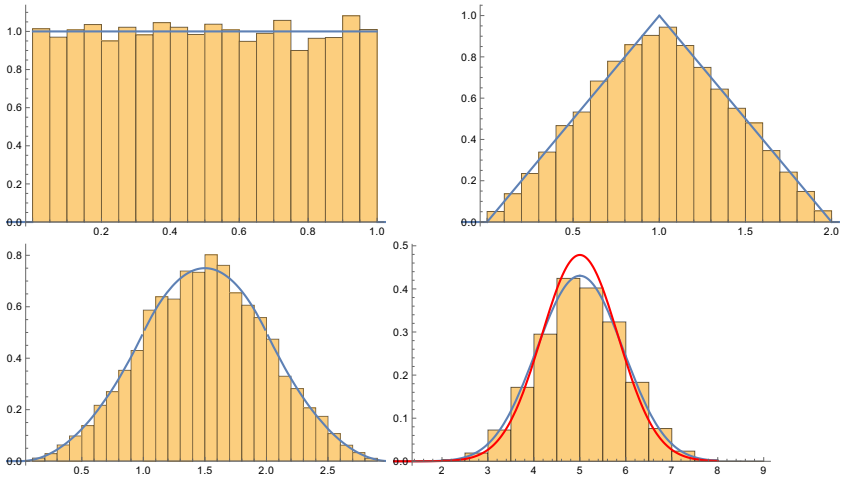


Figura 9.2: Soma de variáveis uniformes para $n = 1,2,3$ e 12 respectivamente. Na última figura comparamos a densidade da soma com a normal correspondente.

Demonstração. Esboçamos a demonstração no caso em que a função geradora de momentos de X_1 existe em uma vizinhança de 0. Defina

$$Z_i = \frac{X_i - \mu}{\sigma}.$$

Então $\mathbf{E}[Z_i] = 0$ e $\text{Var}(Z_i) = 1$. Se M_Z é a função geradora de momentos de Z_i, sua expansão em torno de 0 fornece

$$M_Z(u) = 1 + \frac{u^2}{2} + o(u^2), \qquad u \to 0.$$

Para

$$Y_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma},$$

a independência implica

$$M_{Y_n}(t) = \left[M_Z\left(\frac{t}{\sqrt{n}}\right)\right]^n = \left(1 + \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right)^n.$$

Logo,

$$M_{Y_n}(t) \longrightarrow e^{t^2/2},$$

que é a função geradora de momentos da normal padrão. Pelo Teorema 9.8, Y_n converge em distribuição para $N(0,1)$. ■

Voltemos ao estimador de Monte Carlo do Exemplo 9.6. Se $p = \pi/4$, então I_i é Bernoulli com variância $p(1-p)$ e

$$\widehat{\pi}_n = 4\overline{I}_n.$$

A Lei dos Grandes Números garantiu que $\widehat{\pi}_n$ converge para π. O Teorema do Limite Central diz mais:

$$\frac{\sqrt{n}(\widehat{\pi}_n - \pi)}{4\sqrt{p(1-p)}} \xrightarrow{d} N(0,1).$$

Portanto, para n grande, o erro de Monte Carlo é aproximadamente normal e tem desvio padrão

$$4\sqrt{\frac{p(1-p)}{n}}.$$

A escala $n^{-1/2}$ tem uma consequência prática importante: para reduzir pela metade o erro típico, precisamos aproximadamente quadruplicar o número de simulações.

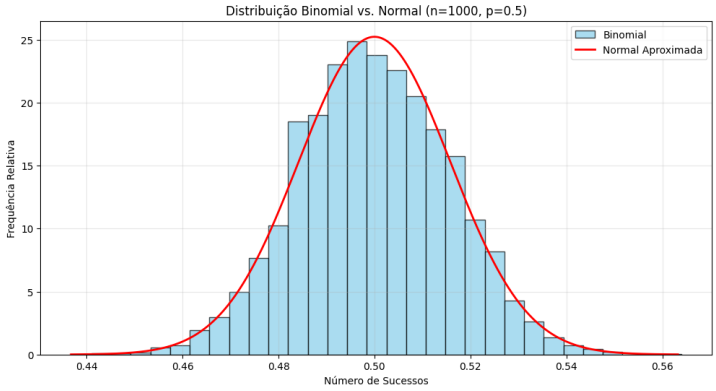
Um corolário do teorema anterior é:

Teorema 9.10 — Aproximação Normal à Distribuição Binomial

Se S_n é uma variável binomial com parâmetros n e p, então:

$$\mathbf{P}\left(a \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq b\right) \xrightarrow{n\to\infty} \mathbf{P}(a \leq Z \leq b),$$

onde $Z \sim \mathcal{N}(0,1)$.



Exemplo 9.15.

Considere a variável aleatória X_i que representa se um indivíduo selecionado aleatoriamente aprova ($X_i = 1$) ou não aprova ($X_i = 0$) o trabalho que o presidente está fazendo. Sabendo disso, calcule a probabilidade de que uma pesquisa eleitoral com $n = 1000$ pessoas encontre uma taxa de aprovação entre 45% e 55% dado que a aprovação do presidente é de 50

Solução. A variável $X = X_1 + \cdots + X_{1000}$ é uma binomial de parâmetros 1000 e 0.5. Para resolver este problema, utilizamos a aproximação normal à distribuição binomial. Primeiro, calculamos a média e o desvio padrão da distribuição binomial:

$$\mu = np = 1000 \times 0{,}5 = 500,$$

$$\sigma = \sqrt{np(1-p)} = \sqrt{1000 \times 0{,}5 \times 0{,}5} = \sqrt{250} \approx 15{,}81.$$

Aplicando a correção de continuidade, consideramos o intervalo $449{,}5 \leq X \leq 550{,}5$. A probabilidade desejada é então calculada através da função de distribuição cumulativa da normal padrão $\Phi(z)$:

$$\mathbf{P}(449{,}5 \leq X \leq 550{,}5) = \mathbf{P}\left(\frac{449{,}5 - \mu}{\sigma} \leq Z \leq \frac{550{,}5 - \mu}{\sigma}\right).$$

Calculando os valores de Z:

$$Z_1 = \frac{449{,}5 - 500}{15{,}81} \approx -3{,}1957,$$

$$Z_2 = \frac{550{,}5 - 500}{15{,}81} \approx 3{,}1957.$$

A probabilidade é então:

$$\begin{aligned}\mathbf{P}(-3{,}1957 \leq Z \leq 3{,}1957) &= \Phi(3{,}1957) - \Phi(-3{,}1957) \\ &\approx 0{,}9993 - 0{,}0007 = 0{,}9986.\end{aligned}$$

Portanto, há aproximadamente 99,86% de probabilidade de que uma pesquisa eleitoral com 1000 pessoas encontre uma taxa de aprovação entre 45% e 55%.

◁

Exemplo 9.16.

Se 100 dados honestos são rolados, determine a probabilidade aproximada de que a soma obtida esteja entre 300 e 400, inclusive.

Solução.

Suponha que X_i represente o valor do i-ésimo dado, $i = 1,2,\ldots,100$. Já que

$$\mathbf{E}\left(X_i\right) = \frac{7}{2}, \quad \operatorname{Var}\left(X_i\right) = \mathbf{E}\left[X_i^2\right] - \left(\mathbf{E}\left[X_i\right]\right)^2 = \frac{35}{12}$$

então

$$\mathbf{E}(X) = 100 \times \mathbf{E}\left(X_i\right) = 350, \quad \operatorname{Var}(X) = 100 \times \operatorname{Var}\left(X_i\right) = \frac{3500}{12}$$

o Teorema do Limite Central resulta em

$$\begin{aligned}\mathbf{P}(299{,}5 \leq X \leq 400{,}5) &= \mathbf{P}\left(\frac{299{,}5 - 350}{\sqrt{\frac{3500}{12}}} \leq \frac{X - 350}{\sqrt{\frac{3500}{12}}} \leq \frac{400{,}5 - 350}{\sqrt{\frac{3500}{12}}}\right) \\ &= \mathbf{P}\left(\frac{-50{,}5}{\sqrt{\frac{3500}{12}}} \leq Z \leq \frac{50{,}5}{\sqrt{\frac{3500}{12}}}\right) \\ &\approx 2\Phi(2{,}957) - 1 \\ &\approx 0{,}9969.\end{aligned}$$

◁

Exemplo 9.17.

O número de atendimentos realizados por um centro de suporte psicológico em um dia é uma variável aleatória de Poisson com média 60. A coordenação do centro decidiu que, se o número de atendimentos exceder 70, será necessário alocar uma equipe extra para lidar com a demanda. Por outro lado, se esse número for menor ou igual a 70, a equipe padrão será suficiente para atender

a todos os casos. Qual é a probabilidade de que a coordenação precise alocar uma equipe extra?

Solução. A solução exata

$$e^{-60} \sum_{i=71}^{\infty} \frac{(60)^i}{i!}$$

não fornece diretamente uma resposta numérica. Entretanto, lembrando que uma variável aleatória de Poisson com média 60 é a soma de 60 variáveis aleatórias independentes, cada uma com média 1, podemos usar o Teorema do Limite Central para obter uma solução aproximada. Se X representa o número de atendimentos realizados no centro, temos

$$\begin{aligned}
\mathbf{P}(X > 70) &= \mathbf{P}(X \geq 70{,}5) \quad \text{(a correção de continuidade)} \\
&= \mathbf{P}\left(\frac{X - 60}{\sqrt{60}} \geq \frac{70{,}5 - 60}{\sqrt{60}}\right) \\
&\approx 1 - \Phi(1{,}35) \\
&\approx 0{,}0885
\end{aligned}$$

onde usamos o fato de que a variância de uma variável aleatória de Poisson é igual à sua média.

◁

9.6 Intervalo de confiança

Até aqui usamos os teoremas-limite principalmente no sentido direto: conhecendo a distribuição de uma observação, aproximamos a distribuição de uma soma ou de uma média. Em inferência estatística, a direção se inverte. O parâmetro é desconhecido, observamos a amostra e usamos a média ou outra estatística para aprender sobre ele.

A Lei dos Grandes Números fornece a primeira garantia: o estimador se estabiliza no parâmetro. O Teorema do Limite Central acrescenta a escala e a forma aproximada do erro. Para a média amostral, por exemplo,

$$\overline{X}_n \xrightarrow{P} \mu, \qquad \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0,1).$$

A primeira relação é uma afirmação de consistência; a segunda permite quantificar a incerteza para amostras grandes. Como transformar essa informação em um intervalo para o parâmetro fixo? A ideia é construir, a partir da

amostra, um intervalo aleatório cuja probabilidade de cobrir o parâmetro possa ser controlada.

Suponha que $X_1,\dots,X_n$ seja uma amostra cuja distribuição depende de um parâmetro desconhecido θ. Um estimador é uma estatística

$$\widehat{\Theta} = g(X_1,\dots,X_n).$$

Antes de observar a amostra, $\widehat{\Theta}$ é uma variável aleatória; depois de observar os dados, obtemos um valor numérico, a estimativa. Por exemplo, a média amostral $\overline{X}$ é um estimador da média populacional.

Uma estimativa pontual não informa, sozinha, o tamanho de sua incerteza. Um **intervalo de confiança** é uma regra que usa a amostra para produzir limites aleatórios

$$I = [L(X_1,\dots,X_n), U(X_1,\dots,X_n)].$$

O parâmetro θ é fixo; o intervalo é que é aleatório. Dizemos que o procedimento tem nível de confiança pelo menos $1-\alpha$ quando, para todo valor admissível de θ,

$$\mathbf{P}_\theta(\theta \in I) \geq 1-\alpha.$$

Essa probabilidade é chamada de *probabilidade de cobertura*. No caso normal com variância conhecida, estudado abaixo, ela será exatamente $1-\alpha$.

Em um intervalo simétrico

$$I = [\widehat{\Theta} - \varepsilon, \widehat{\Theta} + \varepsilon],$$

o evento de cobertura é

$$\{\theta \in I\} = \{|\widehat{\Theta} - \theta| \leq \varepsilon\}.$$

O que é aleatório?

Antes de observar a amostra, θ é fixo e o intervalo I é aleatório; por isso o evento $\{\theta \in I\}$ tem uma probabilidade. Depois de observar os dados, obtemos um intervalo numérico: ele contém θ ou não. O nível de confiança descreve a regra que produz o intervalo, não uma probabilidade atribuída ao parâmetro depois da coleta.

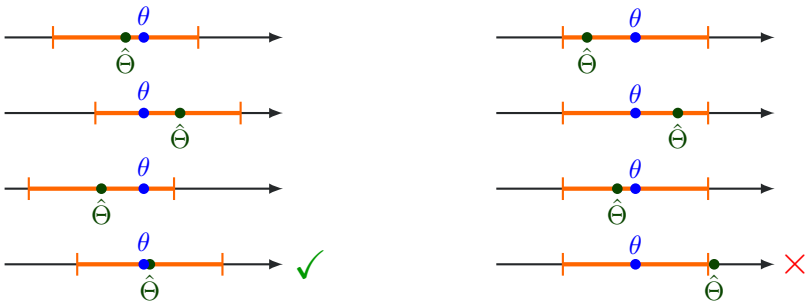


Figura 9.3: À esquerda, θ permanece fixo enquanto o intervalo muda com a amostra; alguns intervalos cobrem θ e outros não. À direita, o intervalo foi artificialmente fixado em torno do parâmetro desconhecido, o que não corresponde a um procedimento de confiança.

A figura enfatiza essa distinção: o parâmetro permanece fixo, enquanto os intervalos mudam de amostra para amostra.

Definição 9.5

Um procedimento de intervalo de confiança de nível pelo menos $1-\alpha$ para θ é um par de estatísticas $L = L(X_1, \ldots, X_n)$ e $U = U(X_1, \ldots, X_n)$ tal que, para todo valor admissível de θ,

$$\mathbf{P}_\theta(L \leq \theta \leq U) \geq 1 - \alpha.$$

Se a igualdade vale para todo θ, dizemos que o procedimento é exato no nível $1-\alpha$.

Métodos assintóticos não garantem essa cobertura em amostras finitas; neles, o alvo $1-\alpha$ é atingido no limite, sob as hipóteses do método.

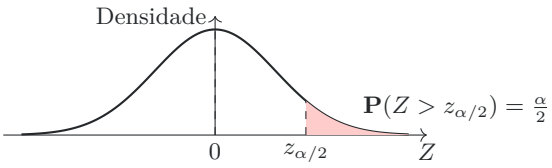
Para construir um intervalo concreto, precisamos conhecer ou aproximar a distribuição do estimador. Começamos pelo caso em que a média amostral tem distribuição normal exata.

Intervalo de confiança para a média de uma população normal com variância conhecida

A seguir, obteremos o intervalo de confiança para a média de uma população normal com variância conhecida. O ponto de partida é o resultado

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

Usaremos a seguinte notação: dado o número α entre 0 e 1, $z_{\alpha/2}$ denota o número real que deixa uma probabilidade de $\alpha/2$ à sua direita em uma distribuição normal padrão, ou seja, $z_{\alpha/2}$ é o número real que resolve a equação: $\mathbf{P}\left(Z > z_{\alpha/2}\right) = \alpha/2$, onde $Z \sim N(0,1)$.



Teorema 9.11

Seja $X_1, X_2, \ldots, X_n$ uma amostra aleatória simples de uma variável aleatória X com distribuição normal $N(\mu, \sigma^2)$, onde σ^2 é conhecida. Seja $\overline{X}$ a média amostral. Então, um intervalo de confiança exato de nível $100(1-\alpha)\%$ para a média populacional μ é dado por

$$\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right).$$

Demonstração. Considere a variável padronizada

$$\frac{\overline{X}-\mu}{\sqrt{\sigma^2/n}}.$$

Seja $X \sim N(\mu,\sigma^2)$. Então, a média amostral $\overline{X}$ é tal que $\overline{X} \sim N(\mu, \sigma^2/n)$, o que implica

$$\frac{\overline{X}-\mu}{\sqrt{\sigma^2/n}} \sim N(0,1).$$

Portanto,

$$\mathbf{P}\left(-z_{\alpha/2} < \frac{\overline{X}-\mu}{\sqrt{\sigma^2/n}} < z_{\alpha/2}\right) = 1-\alpha.$$

Multiplicando cada termo da desigualdade por $\frac{\sigma}{\sqrt{n}}$, obtemos

$$\mathbf{P}\left(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \overline{X}-\mu < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha.$$

Rearranjando em termos de μ, temos

$$\mathbf{P}\left(\overline{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha.$$

Assim, o intervalo de confiança de nível $100(1-\alpha)\%$ para μ é

$$\left(\overline{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \overline{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right).$$

■

Por exemplo, se $\alpha = 0{,}05$, então $z_{\alpha/2} = z_{0{,}025} \approx 1{,}96$ e o intervalo de confiança de 95% é

$$\left(\overline{X}-1{,}96\frac{\sigma}{\sqrt{n}},\ \overline{X}+1{,}96\frac{\sigma}{\sqrt{n}}\right).$$

Dado $\left(\overline{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$, em um intervalo de confiança de nível $100(1-\alpha)\%$ para a média populacional μ, a largura do intervalo é $2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ e metade dessa largura é denominada margem de erro, que é $\epsilon = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

Observe que:

- quanto menor a largura do intervalo, mais informativo ele é.
- os fatores que determinam a largura do intervalo são σ, n e nível de confiança
- a largura do intervalo diminui à medida que n aumenta. Intuição: à medida que o número de observações aumenta, o intervalo se torna mais informativo.
- a largura do intervalo aumenta com o nível de confiança: exigir maior probabilidade de cobertura aumenta o valor crítico e, portanto, alarga o intervalo.

Exemplo 9.18.

O tempo (em minutos) que leva para consertar um dispositivo específico é uma variável aleatória com distribuição $N(\mu, \sigma^2)$, com $\sigma = 1{,}5$. Queremos obter um intervalo de confiança para μ com nível de confiança de 95% a partir de uma amostra com 25 observações de X.

Solução. Neste caso, temos que $z_{0{,}025} = 1{,}96$, então a expressão para o intervalo de confiança é:

$$\left(\overline{X} - 1{,}96\frac{1{,}5}{\sqrt{25}}, \overline{X} + 1{,}96\frac{1{,}5}{\sqrt{25}}\right)$$

Observe que, como σ é assumido como conhecido, a única informação da amostra necessária para obter este intervalo é $\overline{X}$. Se $\overline{X} = 20$ minutos, o intervalo que obtemos é:

$$(20 - 0{,}588,\ 20 + 0{,}588) = (19{,}412,\ 20{,}588)$$

A largura deste intervalo é o dobro da margem de erro, que é $2 \times 0{,}588 = 1{,}176$. Se quisermos obter o intervalo de confiança para μ com um nível de confiança de 99% com a mesma amostra, desde que $z_{0{,}005} = 2{,}58$, o intervalo que obtemos é:

$$\left(\overline{X} - 2{,}58\frac{1{,}5}{\sqrt{25}}, \overline{X} + 2{,}58\frac{1{,}5}{\sqrt{25}}\right)$$

Portanto, como a média da amostra é de 20 minutos, obtemos o seguinte intervalo de confiança:

$$(20 - 0{,}773, 20 + 0{,}773) = (19{,}227, 20{,}773)$$

A largura deste intervalo, $2 \times 0{,}773 = 1{,}546$, é maior do que no caso anterior; esse é o custo de aumentar o nível de confiança mantendo o mesmo tamanho de amostra.

Por fim, suponha que queremos fixar o nível de confiança em 99% mas obtemos uma amostra de X com mais observações, por exemplo 100 observações. Neste caso, a expressão para o intervalo de confiança de nível 99% é:

$$\left(\overline{X} - 2{,}58\frac{1{,}5}{\sqrt{100}},\ \overline{X} + 2{,}58\frac{1{,}5}{\sqrt{100}}\right)$$

Se a média da amostra também fosse de 20 minutos, o intervalo de confiança seria:

$$(20 - 0{,}387,\ 20 + 0{,}387) = (19{,}613,\ 20{,}387)$$

A largura deste intervalo é $2 \times 0{,}387 = 0{,}774$; logicamente, este intervalo de nível 99% construído com 100 observações tem uma largura menor (e, portanto, é mais informativo) do que o intervalo de nível 99% construído com 25 observações.

◁

Intervalo de confiança para uma proporção

Considere variáveis independentes $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$, com $0 < p < 1$, e seja

$$\widehat{p} = \frac{X_1 + \cdots + X_n}{n}.$$

Então

$$\mathbf{E}[\widehat{p}] = p, \qquad \text{SD}(\widehat{p}) = \sqrt{\frac{p(1-p)}{n}}.$$

Pelo Teorema do Limite Central,

$$\frac{\widehat{p} - p}{\sqrt{p(1-p)/n}} \xrightarrow{\mathcal{D}} N(0,1).$$

Como $\widehat{p} \to p$ em probabilidade, o erro-padrão estimado

$$\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

é assintoticamente equivalente ao erro-padrão verdadeiro. Substituí-lo produz o intervalo de Wald. A aproximação pode ser ruim em amostras pequenas ou quando p está muito perto de 0 ou 1.

Teorema 9.12 — Intervalo de Wald

Se $X_1, \ldots, X_n$ são independentes com distribuição Bernoulli(p), para cada $p \in (0,1)$ fixo a cobertura do intervalo

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

converge para $1 - \alpha$ quando $n \to \infty$.

Demonstração. Pelo Teorema do Limite Central,

$$\frac{\widehat{p}-p}{\sqrt{p(1-p)/n}} \xrightarrow{\mathcal{D}} N(0,1).$$

Além disso, $\widehat{p} \to p$ em probabilidade, de modo que

$$\frac{\widehat{p}(1-\widehat{p})}{p(1-p)} \to 1 \qquad \text{em probabilidade.}$$

Pelo teorema de Slutsky, substituir o erro-padrão verdadeiro pelo estimado não altera o limite normal. Portanto, a probabilidade de cobertura do intervalo acima converge para $1-\alpha$. ∎

A largura desse intervalo é

$$2z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

e a metade dessa largura, conhecida como margem de erro, é

$$\epsilon = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Exemplo 9.19.

De uma amostra aleatória de 100 estudantes em uma universidade, 82 afirmaram participar de atividades extracurriculares. Com base nisso, construa o intervalo de Wald de nível nominal 99% para p, a proporção de todos os estudantes na universidade que participam de atividades extracurriculares.
Solução. Para calcular o intervalo de confiança de 99%, primeiro identificamos os parâmetros necessários:

- Nível de significância: $\alpha = 1 - 0{,}99 = 0{,}01$.
- Valor crítico: $z_{\alpha/2} = z_{0{,}005} = 2{,}576$ (obtido da tabela da distribuição normal padrão).

A estimativa pontual da proporção é:

$$\hat{p} = \frac{\text{Número de estudantes que participam de atividades}}{\text{Tamanho da amostra}} = \frac{82}{100} = 0{,}82.$$

O erro-padrão estimado da proporção amostral é:

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0{,}82 \times 0{,}18}{100}} = \sqrt{\frac{0{,}1476}{100}} = 0{,}0384.$$

A margem de erro é:

$$\epsilon = z_{\alpha/2} \times 0{,}0384 = 2{,}576 \times 0{,}0384 \approx 0{,}099.$$

Portanto, o intervalo de Wald de nível nominal 99% para p é:

$$\hat{p} \pm \epsilon = 0{,}82 \pm 0{,}099.$$

Em termos percentuais, entre 72,1% e 91,9%.

◁

A margem de erro do intervalo aproximado acima é

$$z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

onde $\hat{p}$ é a proporção da amostra que possui a característica. Como $\hat{p}(1-\hat{p}) \leq 1/4$, obtemos o limite conservador

$$\text{margem de erro} \leq \frac{z_{\alpha/2}}{2\sqrt{n}}.$$

O limite precedente pode ser usado para escolher o tamanho da amostra de modo que a meia-largura do intervalo de Wald não ultrapasse um valor fixado b, qualquer que seja o valor observado de $\widehat{p}$. Para isso, basta escolher n de modo que

$$\frac{z_{\alpha/2}}{2\sqrt{n}} \leq b$$

será suficiente. Ou seja, n deve ser escolhido de forma que

$$\sqrt{n} \geq \frac{z_{\alpha/2}}{2b}$$

Ao elevar ambos os lados ao quadrado, vemos que n deve ser tal que

$$n \geq \left(\frac{z_{\alpha/2}}{2b}\right)^2$$

Exemplo 9.20.
Qual tamanho de amostra é suficiente para que a meia-largura conservadora do intervalo de Wald de nível nominal 95% não ultrapasse 2%?
Solução. Precisamos escolher n de modo que o limite conservador para a margem de erro seja no máximo 0,02; isto é,

$$n \geq \left(\frac{z_{0,025}}{2 \cdot 0{,}02} \right)^2$$

Como $z_{0,025} = 1{,}96$, isso resulta em

$$n \geq 49^2 = 2401.$$

Ou seja, $n = 2401$ é suficiente para que a meia-largura conservadora do intervalo de Wald de nível nominal 95% seja, no máximo, 2%. Isso não transforma a cobertura assintótica do método de Wald em uma garantia exata de 95% para amostras finitas.

◁

Duas conexões antes de encerrar

O Teorema 9.5 foi usado acima para manipular limites em probabilidade. Vale registrar sua demonstração, porque ela depende apenas de duas ideias que reaparecem muitas vezes: desigualdade triangular e união de eventos.

Demonstração do Teorema 9.5. Para a soma, fixe $\varepsilon > 0$. Pela desigualdade triangular,

$$|(X_n + Y_n) - (\mu_1 + \mu_2)| \leq |X_n - \mu_1| + |Y_n - \mu_2|.$$

Logo,

$$\mathbf{P}(|(X_n + Y_n) - (\mu_1 + \mu_2)| > \varepsilon) \leq \mathbf{P}\left(|X_n - \mu_1| > \frac{\varepsilon}{2}\right) + \mathbf{P}\left(|Y_n - \mu_2| > \frac{\varepsilon}{2}\right),$$

e o lado direito tende a zero.

Para o produto,

$$X_n Y_n - \mu_1 \mu_2 = (X_n - \mu_1)(Y_n - \mu_2) + \mu_1 (Y_n - \mu_2) + \mu_2 (X_n - \mu_1).$$

Os dois últimos termos convergem em probabilidade para zero. Para o primeiro, se $\delta > 0$,

$$\mathbf{P}(|X_n - \mu_1|\,|Y_n - \mu_2| > \delta) \leq \mathbf{P}(|X_n - \mu_1| > \sqrt{\delta}) + \mathbf{P}(|Y_n - \mu_2| > \sqrt{\delta}),$$

que também tende a zero. Outra aplicação da desigualdade triangular conclui que $X_n Y_n \to \mu_1 \mu_2$ em probabilidade. ■

9.7 Exemplos integradores

Os resultados assintóticos ficam mais concretos quando uma mesma quantidade pode ser interpretada tanto como média de muitas observações quanto como aproximação numérica de um objeto determinístico.

Exemplo 9.21 — A densidade de arestas de $G(n,p)$.

Seja M_n o número de arestas de $G(n,p)$ e $N_n = \binom{n}{2}$ o número de arestas possíveis. Como $M_n \sim \mathrm{Bin}(N_n,p)$,

$$\mathbf{E}\left[\frac{M_n}{N_n}\right] = p, \qquad \mathrm{Var}\left(\frac{M_n}{N_n}\right) = \frac{p(1-p)}{N_n}.$$

Chebyshev fornece, para todo $\varepsilon > 0$,

$$\mathbf{P}\left(\left|\frac{M_n}{N_n} - p\right| \geq \varepsilon\right) \leq \frac{p(1-p)}{N_n \varepsilon^2} \longrightarrow 0.$$

Portanto, a fração de arestas presentes converge em probabilidade para p. O parâmetro usado para gerar o grafo reaparece, no limite, como a densidade observada de arestas.

◁

Exemplo 9.22 — Integração por Monte Carlo.

Considere uma função $g : [0,1] \to \mathbb{R}$ com segundo momento finito e a integral

$$I = \int_0^1 g(x)\,dx.$$

Se $U_1, U_2, \ldots$ são independentes e uniformes em $(0,1)$, então

$$\mathbf{E}[g(U_i)] = I.$$

Isso sugere o estimador

$$\widehat{I}_n = \frac{1}{n}\sum_{i=1}^{n} g(U_i).$$

Pela Lei dos Grandes Números,

$$\widehat{I}_n \longrightarrow I,$$

e, se

$$\sigma^2 = \mathrm{Var}(g(U_1)),$$

então

$$\mathrm{Var}(\widehat{I}_n) = \frac{\sigma^2}{n}.$$

O Teorema do Limite Central fornece ainda

$$\frac{\sqrt{n}(\widehat{I}_n - I)}{\sigma} \xrightarrow{d} N(0,1).$$

Por exemplo, podemos estimar

$$\int_0^1 e^{-x^2}\, dx,$$

que não possui primitiva elementar, gerando uniformes e calculando a média de $e^{-U_i^2}$. O método troca uma integral por uma média aleatória: a Lei dos Grandes Números explica a convergência e o TLC quantifica a escala do erro.

Como o erro típico decai como $n^{-1/2}$, reduzi-lo por um fator 10 exige aproximadamente 100 vezes mais amostras. A taxa é lenta, mas não se deteriora diretamente com a dimensão, uma das razões para a utilidade de Monte Carlo em problemas de dimensão alta.

◁

Ao atravessar este capítulo

As desigualdades controlam probabilidades de desvio; as leis dos grandes números justificam médias; Borel–Cantelli separa ocorrências finitas de recorrências infinitas; o teorema central do limite descreve as flutuações e fundamenta aproximações, Monte Carlo e intervalos de confiança.

9.8 Exercícios

Exercício 9.1 — Markov e Chebyshev

Uma variável não negativa tem média 12. Dê o limite de Markov para $\mathbf{P}(X \geq 30)$. Se, além disso, $\mathrm{Var}(X) = 16$, use Chebyshev para limitar $\mathbf{P}(|X - 12| \geq 8)$. Compare o tipo de informação usado.

Exercício 9.2 — Lei dos grandes números

Uma moeda tem probabilidade 0,57 de cara. Seja $\overline{X}_n$ a proporção de caras em n lançamentos. Use Chebyshev para encontrar um n que garanta $\mathbf{P}(|\overline{X}_n - 0{,}57| < 0{,}04) \geq 0{,}95$.

Exercício 9.3 — Modos de convergência

Se $X_n \sim \text{Bernoulli}(1/n)$, mostre que $X_n \to 0$ em probabilidade. Discuta, a partir das hipóteses de independência e do lema de Borel–Cantelli, o que pode ocorrer com a convergência quase certa.

Exercício 9.4 — Aproximação normal

Se $X \sim \text{Bin}(180{,}0{,}4)$, aproxime $\mathbf{P}(65 \leq X \leq 82)$ pela normal, usando correção de continuidade.

Exercício 9.5 — Intervalo e tamanho amostral

O desvio padrão de uma medição é conhecido e vale 3,2. Uma amostra de 64 observações tem média 18,7. Construa um intervalo de confiança de 95% para a média. Depois determine o tamanho mínimo de amostra para margem de erro no máximo 0,5, mantendo 95% de confiança.

Referências

Durrett, R. (2010). *Probability: theory and examples*. Cambridge university press.

Grimmett, G. e D. Welsh (2014). *Probability: an introduction*. Oxford University Press.

Gut, A. (2012). *Probability: a graduate course*. Springer Science & Business Media.

Karr, A. (1992). *Probability*. Springer Science & Business Media.

Klenke, A. (2013). *Probability theory: a comprehensive course*. Springer Science & Business Media.

Resnick, S. I. (2013). *A probability path*. Springer Science & Business Media.

Rosenthal, J. S. (2006). *A first look at rigorous probability theory*. World Scientific Publishing Co Inc.

Shiryaev, A. N. (1996). *Probability*. Springer-Verlag, New York.

Índice Remissivo