Capítulo 13

Probabilidade

A lógica é uma ferramenta essencial pois nos permite deduzir o valor lógico de proposições mais complexas a partir dos valores lógicos de suas proposições e predicados elementares. Porém, para usá-la precisamos saber se as proposições e predicados são verdadeiros ou falsos.

Na vida real, é raro sabermos com certeza se uma afirmação é verdadeira ou não. Todas as fontes de informação que temos — notícias, contagens, medidas, evidências, e nossos próprios sentidos e mente — podem ser errôneas ou enganosas; de modo que toda proposição que acreditamos verdadeira pode ser falsa, e vice-versa. Como podemos então usar a lógica, ou tomar qualquer decisão, nessas condições?

Por outro lado, há afirmações sobre as quais temos muito mais confiança do que outras. Podemos tratar a frase "ontem choveu na minha rua" como verdadeira, com confiança quase absoluta, se estávamos lá ontem. Por outro lado, se a previsão do tempo diz que "não vai chover manhã", é prudente pensar na possibilidade que chova.

Para certas afirmações, nossa confiança pode vir do histórico de situações semelhantes que já presenciamos. Podemos tratar como certa a proposição "uma pedra solta no ar cai para baixo" com base em incontáveis experiências que tivemos ao longo da vida. As leis da física, em particular, são "certezas" adquiridas por meio de experimentos cuidadosos e exaustivamente analisados. Mesmo assim sempre é possível que, em situações especiais que nunca encontramos antes, essas afirmações "certamente verdadeiras" venham a ser falsas.

Para algumas proposições, nossa confiança pode se dividir igualmente entre as duas possibilidades. Alguém jogou uma moeda ao ar e ela caiu onde não podemos ver. Será que o resultado foi cara, ou coroa? Nossa experiência com moedas nos diz que às vezes o resultado é um, às vezes é outro. Da mesma forma, quando atiramos um dado, nossa experiência diz apenas que o resultado pode ser qualquer número entre 1 e 6, e que parece não haver diferença entre eles. Por essa experiência, afirmação "o resultado será 3" merece tanta confiança quanto "o resultado será 5". Na verdade, jogos de azar como dados e cara-ou-coroa baseiam-se inteiramente no fato de que todos resultados possíveis são igualmente plausíveis.

Por outro lado, mesmo nesses jogos há afirmações que merecem mais confiança do que outras. Quando atiramos um dado, a afirmação "o resultado será 3" deve nos parecer menos plausível do que "o resultado será diferente de 3". Esta confiança pode vir da experiência, mas também por raciocínio: se todos os 6 resultados tem chances iguais de acontecer, então o resultado 3 deve ter menos chances do que os outros cinco juntos.

A teoria da probabilidade surgiu para formalizar este tipo de raciocínio, que tem o mesmo

objetivo da lógica clássica — ajudar-nos a pensar e decidir — mas lida com graus de confiança, em vez de certezas absolutas.

13.1 Definição

Nesta teoria, cada proposição P tem uma *probabilidade*: um valor real entre 0 e 1, que mede o grau de confiança ou expectativa que temos de que a proposição seja verdadeira. Denotaremos esse número por $\Pr(P)$. Probabilidade 1 significa que temos certeza absoluta de que a afirmação P é verdadeira. Probabilidade 0 significa que temos certeza absoluta que é falsa. O valor $1/2$ significa que não sabemos se P é falsa ou verdadeira, e que qualquer das duas possibilidades nos parece igualmente provável. Assim, por exemplo, quando vamos jogar uma moeda, podemos atribuir probabilidade $1/2$ à afirmação "o resultado será cara". Uma probabilidade mais próxima de 1 significa que não temos certeza, mas acreditamos que é mais provável que a afirmação P seja verdadeira do que ela seja falsa.

Na teoria de da probabilidade, toda proposição P em tese continua tendo um valor lógico "verdadeiro" ou "falso", mas a teoria não exige que esse valor seja conhecido. A probabilidade da afirmação reflete justamente nosso grau de conhecimento. Se conhecemos o valor lógico da afirmação, devemos atribuir a ela probabilidade 0 ou 1; e, nesse caso, como veremos, a teoria da probabilidade se reduz à lágica clássica.

As probabilidades são frequentemente expressas em percentagens. Assim, tanto faz dizer que uma probabilidade é 25% ou $25/100 = 0,25$.

13.1.1 Distribuição uniforme

Em geral, quando temos n alternativas possíveis para uma situação qualquer, e não temos nenhuma informação, experiência ou raciocínio que justifique atribuir probabilidade maior a uma algumas do que outras, é razoável atribuir probabilidade $1/n$ a cada alternativa. Neste caso dizemos que essas alternativas tem uma *distribuição uniforme* de probabilidade.

Um exemplo de distribuição uniforme é o sorteio de um item entre n outros. Para que o sorteio seja justo é importante que ele seja feito de modo que cada item tenha a mesma probabilidade de ser escolhido. Neste caso dizemos que a escolha é *perfeitamente aleatória*. Esse conceito é importante em muitos jogos 'de azar", como cara-ou-coroa, palitinho, par-ou-ímpar, dados, roletas, baralhos, etc.. Esses jogos dependem de dispositivos ou ações que podem dar dois ou mais resultados distintos. Para que o jogo seja justo, é essencial que os jogadores não tenham nenhum conhecimento prévio sobre o resultado, de modo que todos atribuam uma distribuição uniforme de probabilidade ao mesmo.

Por outro lado, é importante observar que a teoria não diz como atribuir as probabilidades de afirmações elementares, mas apenas como combiná-las para obter as probabilidades de afirmações compostas. É importante notar que as probabilidades dependem do observador: se um jogador troca o dado "honesto" por um viciado, ele pode (e deve) atribuir probabilidades diferentes a cada número.

13.1.2 Princípio da exclusão mútua

Intuitivamente, parece pouco razoável termos confiança ao mesmo tempo em duas afirmações contraditórias. Na teoria da probabilidade, essa intuição é formalizada pelo *princípio da exclusão mútua*, ou *aditividade*: se duas proposições P e Q não podem ser verdadeiras ao mesmo tempo (isto é, $P \to \neg Q$ e $Q \to \neg P$), então devemos ter $\Pr(P) + \Pr(Q) \leq 1$.

Por exemplo, considere as afirmações "o Diretor está agora em São Paulo" e "o Diretor está agora no Rio de Janeiro". Quaisquer que sejam as informações que temos a respeito do paradeiro do Diretor, não faz sentido atribuir probabilidade 0,75 para a primeira e 0,80 para a segunda, pois se uma delas for verdadeira, a outra não é.

Essa regra pode ser generalizada para três ou mais proposições $P_1, P_2, \ldots, P_n$. Essas proposições são *mutuamente exclusivas* se sabemos que $P_i \to \neg P_j$, para quaisquer i e j entre 1 e n com $i \neq j$. Nesse caso, o princípio da exclusão mútua exige que $\Pr(P_1) + \Pr(P_2) + \cdots + \Pr(P_n) \leq 1$.

13.1.3 Princípio da exaustão

Por outro lado, se sabemos que pelo menos uma dentre duas afirmações é verdadeira, não é razoável termos pouca confiança nas duas afirmações. Por exemplo, não é razoável não acreditar nem na afirmação "o lucro será maior que R$ 10.000" nem na afirmação "o lucro será menor que R$ 20.000", pois pelo menos uma dessas afirmações com certeza é verdadeira.

Na teoria da probabilidade, essa regra é formalizada pelo *princípio da exaustão*: se sabemos que $P \vee Q$ é verdadeiro, então devemos ter $\Pr(P) + \Pr(Q) \geq 1$. No exemplo acima, podemos atribuir probabilidade $1/2$ ou $3/4$ para ambas, mas não $1/4$; se atribuirmos probabilidade $0,30$ para a primeira, podemos atribuir $0,80$ para a segunda, mas não $0,50$.

Mais geralmente se sabemos que $P_1 \vee P_2 \vee \cdots \vee P_n$ é verdadeiro, então devemos ter $\Pr(P_1) + \Pr(P_2) + \cdots + \Pr(P_n) \geq 1$.

13.1.4 Princípio da complementaridade

Juntando o princípio da exclusão e da exaustão, podemos concluir que se uma afirmação P é o oposto lógico (negação) da afirmação Q, então a soma das probabilidades deve ser exatamente 1. Ou seja, para qualquer afirmação P, temos

$$\Pr(P) + \Pr(\neg P) = 1 \tag{13.1}$$

ou seja

$$\Pr(\neg P) = 1 - \Pr(P) \tag{13.2}$$

Por exemplo, se a probabilidade de "vai chover amanhã" é $3/4$, a probabilidade de "não vai chover amanhã" tem que ser $1/4$. Esta regra é conhecida como o *princípio da complementaridade*.

Esta regra também pode ser generalizada para três ou mais afirmações. Suponha que sabemos que exatamente uma das afirmações $P_1, P_2, \ldots, P_n$ é verdadeira. Isto é, sabemos que elas são mutuamente exclusivas, mas também que uma delas tem que ser verdadeira. Então devemos ter

$$\Pr(P_1) + \Pr(P_2) + \cdots + \Pr(P_n) = 1 \tag{13.3}$$

Por exemplo, suponha que alguém escolheu e retirou uma carta de um baralho comum. Considere as afirmações "a carta é ouros", "a carta é copas", "a carta é paus", "a carta é espadas", ou "a carta é um coringa". Como a carta só pode ser de um tipo, e tem que ser de um desses cinco tipos, então as probabilidades dessas afirmações devem somar 1.

Observe que este princípio é respeitado quando atribuímos probabilidade $1/n$ para n alternativas igualmente prováveis.

13.1.5 Princípio da exclusão e inclusão

Os princípios acima podem ser vistos como corolários de um princípio mais geral: para quaisquer afirmações P e Q, devemos ter

$$\Pr(P \vee Q) = \Pr(P) + \Pr(Q) - \Pr(P \wedge Q) \tag{13.4}$$

Compare este princípio com a fórmula para cardinalidade de conjuntos

$$|A \cup B| = |A| + |B| - |A \cap B| \tag{13.5}$$

Exercício 13.1: Contagens em uma fábrica mostraram que 5% dos parafusos tem um defeito na rosca, 4% tem um defeito na cabeça, e 2% tem um defeito em ambas as partes. Qual é a probabilidade de que um desses parafusos, escolhido ao acaso, tenha algum defeito?

13.1.6 Princípio da independência

Um dado e uma moeda são atirados ao mesmo tempo. Como discutimos acima, é razoável atribuir probabilidade 1/6 à afirmação "o resultado do dado será 3", e probabilidade 1/2 à afirmação "o resultado da moeda será cara". Que probabilidade devemos atribuir à conjunção dessas duas frases, ou seja "o resultado do dado será 3, e o da moeda será cara"?

Uma maneira de fazer esta escolha é observar que há 12 possíveis resultados para os dois lances. Vamos denotar por $D(x)$ e $M(y)$, respectivamente, os predicados "o resultado do dado será x", e "o resultado da moeda será y". As 12 possibilidades correspondem às afirmações

$$\begin{array}{ll} D(1) \wedge M(\texttt{cara}) & D(1) \wedge M(\texttt{coroa}) \\ D(2) \wedge M(\texttt{cara}) & D(2) \wedge M(\texttt{coroa}) \\ D(3) \wedge M(\texttt{cara}) & D(3) \wedge M(\texttt{coroa}) \\ D(4) \wedge M(\texttt{cara}) & D(4) \wedge M(\texttt{coroa}) \\ D(5) \wedge M(\texttt{cara}) & D(5) \wedge M(\texttt{coroa}) \\ D(6) \wedge M(\texttt{cara}) & D(6) \wedge M(\texttt{coroa}) \end{array} \tag{13.6}$$

Estas afirmações são mutuamente exclusivas e esgotam todas as possibilidades, e portanto a soma de suas probabilidades deve ser 1. Se não temos nenhuma razão para suspeitar que o dado de alguma maneira influencie a moeda, ou vice-versa, então é razoável atribuir a mesma probabilidade (1/12) a estas 12 afirmações.

Note que 1/12 é o produto de $\Pr(D(x)) = 1/2$ e $\Pr(M(y)) = 1/6$. Temos portanto que $\Pr(D(x) \wedge M(y)) = \Pr(D(x)) \Pr(M(y))$ para quaisquer x e y.

Este é um exemplo de uma regra geral, o *princípio da independência*. Por definição, duas afirmações P e Q são ditas *independentes* se e somente se

$$\Pr(P \wedge Q) = \Pr(P)\Pr(q) \tag{13.7}$$

O princípio da independência diz que, se não sabemos de nenhuma ligação ou influência entre o valor lógico de uma afirmação P e o de outra afirmação Q, então é razoável supor que elas são independentes; ou seja, é razoável atribuir à conjunção $P \wedge Q$ o produto das respectivas probabilidades.

Exercício 13.2: Dois dados, um vermelho e um verde, são atirados ao mesmo tempo. Qual é a probabilidade de que o resultado do dado vermelho seja menor que 4, e o do dado verde seja maior que 1?

Exercício 13.3: Se as afirmações P e Q são independentes, quanto vale $\Pr(P \vee Q)$ em função de $\Pr(P)$ e $\Pr(Q)$?

Exercício 13.4: Contagens em uma fábrica mostraram que 20% dos parafusos tem um defeito na rosca, 30% tem um defeito na cabeça. Supondo que os defeitos afetam as duas partes do parafuso de maneira independente, qual é a probabilidade de que um desses parafusos, escolhido ao acaso, tenha algum defeito?

13.1.7 Relação com a lógica clássica

A teoria da probabilidade inclui a lógica clássica como caso particular. Mais precisamente, atribuir probabilidade 0 a uma afirmação equivale a acreditar que a afirmação é falsa; e atribuir probabilidade 1 equivale a acreditar que ela é verdadeira. Se todas as afirmações tem probabilidade 0 ou 1, as regras e conceitos da lógica clássica podem ser traduzidos por regras e conceitos da probabilidade. Por exemplo, o conetivo $P \to Q$ equivale a afirmar que $\Pr(Q|P) = 1$.

13.2 Variável aleatória

Uma *variável aleatória* é uma variável (parâmetro, quantia) X cujo valor é conhecido apenas parcialmente, no sentido probabilístico. Isto é, sabemos que o valor de X é algum elemento de um certo conjunto D, o *domínio* da variável; e, para qualquer v em D, temos uma medida de probabilidade $\Pr(X = v)$ para a afirmação "$X = v$". A função que a cada $v \in D$ associa a probabilidade $\Pr(X = v)$ é chamada de *distribuição de probabilidade* (ou simplesmente *distribuição*) da variável X.

Observe que, se u, v são elementos distintos de D, então as afirmações "$X = u$" e "$X = v$" são mutuamente exclusivas. Além disso, sabemos que existe algum elemento v em D tal que a afirmação "$X = v$" é verdadeira. Pelo princípio de inclusão e exclusão, temos portanto que

$$\sum_{v \in D} \Pr(X = v) = 1$$

Observe também que, nestas condições, temos que atribuir $\Pr(X = v) = 0$ para qualquer valor v que não está no conjunto D.

Exemplo 13.1: Um dado foi lançado, mas o resultado da jogada ainda está oculto. Seja X a variável aleatória cujo valor é esse resultado. Sabemos que o domínio de X é o conjunto $D = \{1, 2, \ldots, 6\}$. Como não temos motivos para distinguir entre esses resultados, é razoável atribuir probabilidades iguais $(1/6)$ para cada valor em D, e probabilidade zero para qualquer outro valor. Em particular, $\Pr(X = 3) = \Pr(X = 5) = 1/6$, e $\Pr(X = 0) = \Pr(X = 7) = \Pr(X = 1/2) = 0$.

Variáveis aleatórias com valores numéricos podem ser combinadas com operações aritméticas e funções matemáticas, resultando em outras variáveis aleatórias. Por exemplo, se α é um número real, a fórmula $\alpha X + \sqrt{Y}$ denota a variável aleatória cujo valor é $\alpha u + \sqrt{v}$, onde u é o valor de X e v o valor de Y. A distribuição dessa nova variável é determinada pelas distribuições de probabilidades de X e de Y.

Exercício 13.5: Sejam X e Y os resultados obtidos atirando-se dois dados de cores diferentes, cada um com distribuição uniforme de probabilidades. Determine a distribuição das seguintes variáveis derivadas de X e Y:

1. X^2
2. $X \bmod 3$
3. $X + Y$
4. $\min\{X, Y\}$

Neste livro só vamos tratar de variáves aleatórias cujos domínios são conjuntos discretos (finitos ou enumeráveis). A teoria pode ser estendida para variáveis com domínios não enumeráveis, como os números reais; mas esse assunto merece uma disciplina à parte.

13.3 Valor esperado

Um uso importante (e o mais antigo) da teoria da probabilidade é avaliar o ganho ou perda que pode decorrer de uma escolha ou acontecimento cujo resultado é desconhecido, como por exemplo uma aposta ou um investimento na bolsa.

Suponha por exemplo que atiramos uma moeda e apostamos R$ 30 contra R$ 10 que o resultado será cara. Temos igual chance de ganhar R$ 10 (se sair cara) e perder R$ 30 (se sair coroa). Ou seja,

$$\Pr(\text{“nosso ganho será R\$ } 10\text{”}) \;=\; \Pr(\text{“nosso ganho será R\$ } -30\text{”}) \;=\; \frac{1}{2}$$

Intuitivamente, se repetirmos essa aposta n vezes, em aproximadamente metade das vezes vamos ganhar 10 e na outra metade perder 30; portanto o ganho por aposta, em média, será aproximadamente

$$\frac{\frac{n}{2}(\text{R\$ } 10) + \frac{n}{2}(\text{R\$ } -30)}{n} = \text{R\$ } -10 \tag{13.8}$$

Para entender melhor este exemplo, suponha que repetimos duas vezes essa aposta. Temos quatro possibilidades: perder nas duas vezes, só na primeira, só na segunda, ou ganhar nas duas. Nosso ganho médio por aposta será respectivamente, $(-30 - 30)/2 = -30$, $(-30 + 10)/2 = -10$, $(10 - 30)/2 = -10$, e $(10 + 10)/2 = +10$. Supondo que o resultado de cada lance seja independente

dos anteriores, e denotando por $G(x)$ o predicado "nosso ganho médio por aposta será x", teremos então

$$\begin{aligned} \Pr(G(-30)) &= 1/4 \\ \Pr(G(-10)) &= 1/4 + 1/4 = 1/2 \\ \Pr(G(+10)) &= 1/4 \end{aligned} \tag{13.9}$$

Ou seja, o ganho médio R\$ -10 é duas vezes mais provável que R\$ -30 ou R\$ $+10$. Para quatro apostas seguidas, podemos ter 0, 1, 2, 3, ou 4 acertos, com ganhos médios por aposta de -30, -20, -10, 0 e $+10$, respectivamente. As probabilidades são

$$\begin{aligned} \Pr(G(-30)) &= \binom{4}{0}/2^4 = 1/16 \\ \Pr(G(-20)) &= \binom{4}{1}/2^4 = 4/16 \\ \Pr(G(-10)) &= \binom{4}{2}/2^4 = 6/16 \\ \Pr(G(0)) &= \binom{4}{3}/2^4 = 4/16 \\ \Pr(G(+10) &= \binom{4}{4}/2^4 = 1/16 \end{aligned} \tag{13.10}$$

Como se pode ver, é muito mais provável que o ganho médio por aposta seja R\$ -10 do que qualquer outro valor. A medida que o número de apostas aumenta, essa tendência permanece: o valor mais provável para o ganho médio por aposta será R\$ -10.

Em geral, suponha que temos uma variável aleatória X que pode assumir qualquer valor de um conjunto de valores numéricos D. O *valor médio esperado* (ou simplesmente o *valor esperado*) de X é, por definição

$$\mathcal{E}X = \sum_{v \in D} = v \Pr(X = v) \tag{13.11}$$

Para entender esta fórmula, suponha que temos uma coleção grande com N variáveis, todas elas semelhantes a X mas tais que o valor de uma delas não tem influência nos valores das outras. Nesse caso, o número de variáveis que tem valor v será aproximadamente $N \Pr(X = v)$.

Observe que se D tem um número finito n valores distintos, e todos os valores de D são igualmente prováveis, então $\Pr(X = v) = 1/n$, e a fórmula do valor esperado (13.11) reduz-se à média aritmética dos elementos de D.

Exercício 13.6: Furar um poço de petróleo em determinada região custa R\$500.000, e tem 30% de chance de encontrar óleo. Se isso acontecer, o poço pode ser vendido por R\$800.000. Caso contrário o investimento é totalmente perdido. Qual o ganho esperado por poço?

Quando o domínio da variável é um conjunto infinito, o valor esperado pode ser infinito, mesmo que todos os seus valores possíveis sejam finitos. Por exemplo, considere a variável X cujo valor é um inteiro positivo, tal que $\Pr(X = k) = (6/\pi^2)/k^2$ para todo $k \in \mathbb{N} \setminus \{0\}$. Esta distribuição de probabilidades é válida, pois verifica-se que a soma de todas as probabilidades é 1. Entretanto, o valor esperado de X deveria ser a somatória

$$\mathcal{E}(X) = \sum_{k=0}^{\infty} k \cdot \frac{A}{k^2} = A \sum_{k=0}^{\infty} \frac{1}{k}$$

que, como sabemos, não tem valor finito (veja seção 8.6).

O valor esperado pode ser definido para qualquer variável cujos valores podem ser somados e multiplicados por um número real. Por exemplo, suponha que o valor de uma variável aleatória X é um par (u, v), onde u é o resultado de lançar uma moeda (0 = cara, 1 = coroa), e v é o resultado de lançar um dado (um inteiro entre 1 e 6); sendo que cada par possível tem a mesma probabilidade $1/12$. Note que esses pares podem ser considerados vetores do espaço $\mathbb{R}^2$. Portanto podemos calcular o valor esperado de X

$$\mathcal{E}(X) \;=\; \frac{1}{12}\left((0,1) + (0,2) + \cdots + (1,5) + (1,6)\right) \;=\; \left(\frac{1}{6}, \frac{7}{2}\right)$$

13.3.1 Propriedades do valor esperado

Seja X uma variável aleatória com domínio numérico, sejam α e β dois números reais quaisquer. Nesse caso, pode-se provar que

$$\mathcal{E}(\alpha X + \beta) = \alpha\, \mathcal{E}(X) + \beta \tag{13.12}$$

Porém, se uma variável aleatória Z depende de X de maneira não linear (por exemplo, se Z é o quadrado de X), não existe uma fórmula geral que relacionem $\mathcal{E}(Z)$ a $\mathcal{E}(X)$ (Veja o exercício 13.8.)

Sejam X e Y duas variáveis aleatórias com valores numéricos, e seja Z a variável aleatória, denotada por $X + Y$, cujo valor é a soma dos valores de X e de Y. Verifica-se que

$$\mathcal{E}(Z) = \mathcal{E}(X) + \mathcal{E}(Y) \tag{13.13}$$

Estas fórmulas valem mesmo que as variáveis X e Y tenham alguma dependência entre si. Note que não há fórmulas análogas para outras operações (como produto, divisão, etc.).

Exercício 13.7: Um dado vai ser lançado, e a seguinte aposta é oferecida: o cliente paga R$7,00 ao banqueiro, e recebe em reais o dobro do valor que sair no dado. Por exemplo, se sair um 4, o cliente recebe R$8,00, obtendo um ganho líquido de R$1,00. Qual é o ganho esperado do cliente?

Exercício 13.8: Na mesma situação do exercício 13.7, uma outra aposta é oferecida: cliente paga R$49,00 ao banqueiro, e recebe em reais o dobro do quadrado do valor que sair no dado. Por exemplo, se sair um 6, o cliente recebe 2×6^2 = R$72,00, obtendo um ganho líquido de R$23,00. Qual é o ganho esperado do cliente?

13.4 Mediana

O valor esperado de uma variável aleatória X pode em muitos casos ser considerado o "valor típico" de X. Por exemplo, se X é a altura (em metros) de uma pessoa que não vimos ainda, o valor esperado de X para a população brasileira é próximo a 1,70 m. Podemos então imaginar o "brasileiro típico" como tendo essa altura.

Porém este raciocínio nem sempre é apropriado. Por exemplo, suponha uma vila com 99 casas térreas e um prédio de 101 andares, e considere a variável aleatória X que é o número de andares de um edifício arbitrário dessa vila, escolhido com probabilidade uniforme. O valor esperado da

variável X será 2, mas obviamente não é correto dizer que o "edifício típico" dessa vila tem dois andares.

Devido a exemplos como esse, foram propostas outras maneiras de obter o "valor típico" de uma variável aleatória. O mais comum é a *mediana*. Idealmente, este é um valor v tal que $\Pr(X \leq v) \geq 1/2$ e $\Pr(X \geq v) \geq 1/2$.

Por exemplo, suponha que a variável aleatória X pode ter qualquer valor inteiro entre 1 e 6, com as seguintes probabilidades

k	1	2	3	4	5	6
$\Pr(X = k)$	$\frac{6}{20}$	$\frac{2}{20}$	$\frac{1}{20}$	$\frac{3}{20}$	$\frac{7}{20}$	$\frac{1}{20}$

Neste caso podemos tomar a mediana de X como sendo 4, pois

$$\begin{array}{rcccl} \Pr(X \leq 4) & = & \frac{6}{20} + \frac{2}{20} + \frac{1}{20} + \frac{3}{20} & = & \frac{12}{20} \geq \frac{1}{2} \\ \Pr(X \geq 4) & = & \frac{3}{20} + \frac{7}{20} + \frac{1}{20} & = & \frac{11}{20} \geq \frac{1}{2} \end{array}$$

Note que o valor esperado de X é

$$1 \cdot \frac{6}{20} + 2 \cdot \frac{2}{20} + 3 \cdot \frac{1}{20} + 4 \cdot \frac{3}{20} + 5 \cdot \frac{7}{20} + 6 \cdot \frac{1}{20} = \frac{66}{20} = 3,3$$

Note porém que pode haver diversos valores v que satisfazem a condição $\Pr(X < v) = \Pr(X > v)$. Por exemplo, se a distribuição de probabilidades de X for

k	1	2	3	4	5	6
$\Pr(X = k)$	$\frac{6}{20}$	$\frac{2}{20}$	$\frac{2}{20}$	$\frac{1}{20}$	$\frac{8}{20}$	$\frac{1}{20}$

então, para qualquer valor v tal que $3 < v < 4$, teremos $\Pr(X \leq v) = (6 + 2 + 2)/20 = 1/2$ e $\Pr(X \geq v) = (1 + 8 + 1)/20 = 1/2$.

Quando isso acontece, pode-se provar que os valores de v que satisfazem a definição formam um intervalo finito dos números reais. Nesses casos, alguns autores definem a mediana como sendo o ponto médio desse intervalo; no exemplo acima, seria $v = (3 + 4)/2 = 3,5$.

Exercício 13.9: Seja X o quadrado de um número entre 1 e 6 que será obtido pelo lançamento de um dado. Note que o valor de X pode ser 1, 4, 9, 16, 25, ou 36. Qual é o valor esperado da variável X? E sua mediana?

Exercício 13.10: Seja X o *produto* dos dois números entre 1 e 6 que serão obtidos pelo lançamento de dois dados. Qual é a distribuição de probabilidades da variável X? Qual é seu valor esperado? E sua mediana?

Exercício 13.11: Prove que qualquer variável aleatória com valores inteiros tem uma mediana.

13.5 Moda

Outra maneira de definir o "valor típico" de uma variável aleatória é tomar o *valor mais provável*, também chamado de *moda* da variável. Por exemplo, se a distribuição for

k	1	2	3	4	5	6
$\Pr(X = k)$	$\frac{6}{20}$	$\frac{2}{20}$	$\frac{1}{20}$	$\frac{3}{20}$	$\frac{7}{20}$	$\frac{1}{20}$

diremos que a moda de X é 5. Por outro lado, se as probabilidades forem um pouco diferentes

k	1	2	3	4	5	6
$\Pr(X = k)$	$\frac{7}{20}$	$\frac{2}{20}$	$\frac{1}{20}$	$\frac{3}{20}$	$\frac{6}{20}$	$\frac{1}{20}$

A moda será 1.

13.6 Variância e desvio padrão

Em muitas situações, não basta saber o valor esperado $\mathcal{E}(X)$ de uma variável aleatória; é preciso também saber até que ponto o valor da variável pode diferir desse valor esperado.

Considere por exemplo as variáveis aleatórias X e Y, que podem assumir valores entre 1 e 5 com as seguintes probabilidades:

k	1	2	3	4	5
$\Pr(X = k)$	$\frac{1}{20}$	$\frac{7}{20}$	$\frac{4}{20}$	$\frac{7}{20}$	$\frac{1}{20}$
$\Pr(X = k)$	$\frac{7}{20}$	$\frac{2}{20}$	$\frac{2}{20}$	$\frac{2}{20}$	$\frac{7}{20}$

As duas variáveis tem o mesmo valor esperado $v = 3$, mas intuitivamente podemos ver que Y varia mais do que X. Como podemos transformar essa intuição em números?

A maneira mais comum é calcular a *variância* $\mathcal{V}(X)$ da variável, definida pela fórmula

$$\mathcal{V}(X) = \sum_{v \in D} (v - \mathcal{E}(X))^2 \Pr(X = v) \tag{13.14}$$

Pode-se verificar que este é o valor esperado da variável $Y = (X - \mathcal{E}(X))^2$.

No exemplo acima, temos

$$\begin{aligned} \mathcal{V}(X) &= (1-3)^2 \cdot \tfrac{1}{20} + (2-3)^2 \cdot \tfrac{7}{20} + (3-3)^2 \tfrac{4}{20} + (4-3)^2 \cdot \tfrac{7}{20} + (5-3)^2 \cdot \tfrac{1}{20} = \tfrac{26}{20} = 1,3 \\ \mathcal{V}(Y) &= (1-3)^2 \cdot \tfrac{7}{20} + (2-3)^2 \cdot \tfrac{2}{20} + (3-3)^2 \tfrac{2}{20} + (4-3)^2 \cdot \tfrac{2}{20} + (5-3)^2 \cdot \tfrac{7}{20} = \tfrac{60}{20} = 3,0 \end{aligned}$$

evidenciando assim que os valores de Y tendem a estar mais longe de sua média do que os valores de X.

Observe que as parcelas $(v - \mathcal{E}(X))^2$ da somatória (13.14) nunca são negativas, portanto a variância também não pode ser negativa. Além disso, a variância só pode ser zero se todas as parcelas forem zero, ou seja se a variável X só pode ter um valor — que é portanto seu valor esperado $\mathcal{E}(X)$. Se ela pode assumir dois ou mais valores distintos, com probabilidades diferentes de zero, então a variância será estritamente positiva.

Observe que, se o domínio D da variável X é um conjunto infinito, a variância pode ser infinita (mesmo que o valor esperado exista e seja finito). Por exemplo, seja $D = \mathbb{Z} \setminus \{0\}$, e $\Pr(X = v) = B/|v|^3$, onde B é uma constante tal que a soma das probabilidades seja 1. O valor esperado existe ($\mathcal{E}(X) = 0$). Porém, temos

$$\sum_{v \in D} \Pr(X = v)(v - \mathcal{E}(X))^2 = 2 \sum k = 1^{+\infty} \frac{B}{v^3} v^1 = 2B \sum k = 1^{+\infty} \frac{1}{v}$$

que, como sabemos, é infinita.

13.6.1 Propriedades da variância

Seja X uma variável aleatória com valores numéricos. Sejam α e β dois valores reais arbitrários. Verifica-se então que

$$\mathcal{V}(\alpha X + \beta) = \alpha^2 \mathcal{V}(X) \tag{13.15}$$

Note que somar uma constante β a uma variável não altera sua variância.

Se X e Y são duas variáveis aleatórias *independentes*, verifica-se que

$$\mathcal{V}(X + Y) = \mathcal{V}(X) + \mathcal{V}(Y) \tag{13.16}$$

Esta fórmula não vale se soubermos de alguma dependência entre as variáveis X e Y (isto é, se atribuímos a alguma afirmação do tipo "$(x = u) \wedge (Y = v)$" uma probabilidade diferente de $\Pr(X = u)\Pr(Y = v)$). Nesse caso, a variância de $X + Y$ pode ser maior ou menor que $\mathcal{V}(X) + \mathcal{V}(Y)$.

13.6.2 Desvio padrão

Pode-se dizer que, quanto maior a variância, mais "espalhada" é a distribuição de probabilidade da variável. Entretanto, não é fácil interpretar o valor numérico da variância. Por exemplo, se o valor de X é uma medida em metros, a variância é medida em metros quadrados. Uma medida de "espalhamento" que é mais fácil de interpretar é o *desvio padrão*, definido como a raiz quadrada da variância:

$$\mathcal{D}(X) \;=\; \sqrt{\mathcal{V}(X)} \;=\; \sqrt{\sum_{v \in D} (v - \mathcal{E}(X))^2 \Pr(X = v)}$$

O desvio padrão é medido com as mesmas unidades da variável. Informalmente, pode ser interpretado como o valor "típico" da diferença entre o valor da variável e seu valor esperado.

> **Exemplo 13.2:** Suponha um lote de parafusos que deveriam ser todos iguais, e Seja X o comprimento real de um desses parafusos, escolhido ao acaso. Se dissermos que o valor esperado de X é 150 mm e o desvio padrão é 1 mm, estamos dizendo que o comprimento do parafuso dificilmente será muito maior que 151 mm ou muito menor que 149 mm.

Esta interpretação informal do desvio padrão tem por base o seguinte resultado, devido ao matemático russo Pafnuti Chebyshev ou Tchebychev (1821–1894):

> **Teorema 13.1:** Para qualquer variável aleatória X, e qualquer número real $\alpha \geq 1$,
>
> $$\Pr(|X - \mathcal{E}(X)| \geq \alpha\, \mathcal{D}(X)) \leq \frac{1}{\alpha^2} \tag{13.17}$$

A demonstração deste resultado foge do escopo deste livro. Em outras palavras, se $\mathcal{E}(X) = \mu$ e $\mathcal{D}(X) = \sigma$, então o valor de X estará dentro do intervalo $[\mu - \alpha\sigma, \mu + \alpha\sigma]$ com probabilidade $1 - 1/\alpha^2$. Para a variável X do exemplo 13.2, o teorema de Tchebychev diz que o comprimento do parafuso (em milímetros) está

- no intervalo $[150 - 2 \cdot 1, 150 + 2 \cdot 1] = [148, 152]$ com probabilidade maior ou igual a $1 - 1/2^2 = 75\%$;

- no intervalo $[150 - 3 \cdot 1, 150 + 3 \cdot 1] = [147, 153]$ com probabilidade maior ou igual a $1 - 1/3^2 \approx 88\%$;
- no intervalo $[150 - 4 \cdot 1, 150 + 4 \cdot 1] = [146, 154]$ com probabilidade maior ou igual a $1 - 1/4^2 \approx 93\%$;

e assim por diante.

Observe que o resultado de Tchebychev vale qualquer que seja a distribuição de probabilidade da variável X.

Exercício 13.12: Seja X uma variável aleatória que pode assumir qualquer valor entre 0 e 100, com igual probabilidade. Calcule o valor esperado, a variância e o desvio padrão de X. Calcule a probabilidade de X estar entre 40 e 60 (inclusive ambos). Compare esse resultado com a probabilidade obtida pelo teorema de Tchebychev.

13.6.3 Covariância

Se X e Y são variáveis aleatórias numéricas, a *covariãncia* entre as duas é definda pela fórmula

$$\mathcal{C}(X, Y) = \sum_{u,v} \Pr((X = u) \wedge (Y = v))(u - \mathcal{E}(X))(v - \mathcal{E}(Y))$$

A covariância é uma medida da dependência entre X e Y. A grosso modo, ela tende a ser positiva quando é muito provável que os valores de X e Y sejam ambos maiores ou ambos menores que suas médias (caso em que o produto $(u - \mathcal{E}(X))(v - \mathcal{E}(Y))$ é positivo). Ela tende a ser negativa quando X e Y tendem a variar em direções opostas em relação a suas médias — quando um está acima da média, o outro provavelmente está abaixo. Observe que $\mathcal{V}(X)$ é a mesma coisa que $\mathcal{C}(X, X)$.

É fácil provar que, se X e Y são independentes, então sua covariância é zero. Prova-se também que, para quaiquer variáveis aleatórias numéricas X e Y,

$$\mathcal{V}(X + Y) = \mathcal{V}(X) + \mathcal{V}(Y) + 2\,\mathcal{C}(X, Y)$$

Note que esta fórmula implica na fórmula (13.16) quando X e Y são independentes.

Exercício 13.13: Encontre duas variáveis aleatórias X e Y que possuem covariância nula mas *não* são independentes.

13.6.4 Coeficiente de correlação

O sinal de $\mathcal{C}(X, Y)$ revela o sentido geral da dependência entre X e Y, mas seu valor numérico é difícil de interpretar. Por essa razão é interessante definir o *coeficiente de correlação*

$$\kappa(X, Y) = \frac{C(X, Y)}{\sqrt{\mathcal{V}(X)\,\mathcal{V}(Y)}} = \frac{C(X, Y)}{\mathcal{D}(X)\,\mathcal{D}(Y)}$$

Prova-se que este número está sempre entre -1 e $+1$. Ele é zero se X e Y são independentes, $+1$ se cada variável é função linear crescente da outra (isto é, se $Y = \alpha X + \beta$ com $\alpha > 0$) e -1 se cada variável é função linear descrecente da outra ($Y = \alpha X + \beta$ com $\alpha < 0$). Um valor intermediário, por exemplo $0,50$, significa que o valor de cada variável é parcialmente função da outra, mas inclui um termo que não depende dela. Neste caso diz-se que *há correlação entre X e Y* (*positiva* ou *negativa*, conforme o sinal do coeficiente).

13.7 Probabilidade condicional

Seja X a variável aleatória cujo valor é o resultado do lançando um dado, e considere as duas afirmações "X é par" e "X é ímpar". Se não temos nenhuma outra informação sobre X, como vimos, é razoável atribuir a probabilidade $1/6$ a cada um dos possíveis valores 1, 2, ..., 6, e portanto

$$\begin{array}{lcl} \Pr(X \text{ é par}) & = & \Pr(X=2) + \Pr(X=4) + \Pr(X=6) \;=\; 1/2 \\ \Pr(X \text{ é ímpar}) & = & \Pr(X=1) + \Pr(X=3) + \Pr(X=5) \;=\; 1/2 \end{array}$$

Suponha agora que sabemos que o valor de X não é 3. Que probabilidade devemos atribuir a essas duas afirmações? Não podemos simplesmente eliminar o termo $\Pr(X = 3)$ na segunda fórmula, pois a soma não seria 1. Como a probabilidade do valor ser 3 é zero, temos que corrigor a probabilidade dos demais valores para que elas tenham soma 1. Ou seja, temos que supor $\Pr(X = 3) = 0$ e $\Pr(X = v) = 1/5$ para os demais valores. Então teremos

$$\begin{array}{lcl} \Pr(X \text{ é par}) & = & \Pr(X=2) + \Pr(X=4) + \Pr(X=6) \;=\; 3/5 \\ \Pr(X \text{ é ímpar}) & = & \Pr(X=1) + \Pr(X=5) \;=\; 2/5 \end{array}$$

Observe que a informação adicional "$X \neq 3$" afetou não apenas a probabilidade de X ser ímpar, mas também a probabilidade de ele ser par.

Em casos como este, costuma-se usar a notação $\Pr(P|Q)$ para denotar a *probabilidade condicional* da afirmação P, *sabendo-se que* (ou *dado que*) a afirmação Q é verdadeira. Verifica-se que essa probabilidade pode ser calculada pela fórmula

$$\Pr(P|Q) = \frac{\Pr(P \wedge Q)}{\Pr(Q)} \tag{13.18}$$

Aplicando esta fórmula ao exemplo acima, a afirmação P seria "X é ímpar" e Q a afirmação "$X \neq 3$". Temos então que

$$\begin{array}{lcll} \Pr(P \wedge Q) & = & \Pr(X=1) + \Pr(X=5) & = \; 2/6 \\ \Pr(Q) & = & \Pr(X=1) + \Pr(X=2) + \Pr(X=4) + \Pr(X=5) + \Pr(X=6) & = \; 5/6 \\ \Pr(P|Q) & = & \dfrac{2/6}{5/6} & = \; 2/5 \end{array}$$

Exercício 13.14: Seja X o valor obtido lançando um dado. Calcule, pela fórmula (13.18)

1. $\Pr(X \text{ é par} | X \neq 3)$
2. $\Pr(X \text{ é par} | X \text{ é quadrado perfeito})$
3. $\Pr(X \text{ é primo} | X \text{ é maior que } 2)$

Exercício 13.15: Seja X a soma dos valores obtidos no lançamento de dois dados. Calcule, pela fórmula (13.18)

1. $\Pr(X \text{ é par} | \text{os dois dados deram o mesmo resultado})$
2. $\Pr(X \text{ é par} | \text{os dois dados deram resultados diferentes})$
3. $\Pr(X = 6 | \text{os dois valores não são primos entre si})$

A fórmula da probabilidade condicional é também muito usada na forma inversa:

$$\Pr(P \wedge Q) = \frac{\Pr(P|Q)}{\Pr(Q)} \tag{13.19}$$

Ou seja, uma vez definida a probabilidade de P dado Q, e também a probabilidade de Q, a probabilidade da afirmação "P e Q" é simplesmente o produto das duas.

Exercício 13.16: Suponha que a probabilidade de algum hacker tentar violar seu computador no próximo minuto é 10%, e que a probabilidade de tal tentativa ter sucesso é 80%. Qual é a probabilidade de seu computador ser violado por algum hacker no próximo minuto? (Ignore a possibilidade de haver mais de um ataque por minuto.)

Exercício 13.17: Suponha que atiramos dois dados, um verde e um vermelho. Qual a probabilidade de que o dado verde mostre o valor 2, e o dado vermelho mostre o valor 3? E qual é a probabilidade de que um deles mostre o valor 2, e o outro 3? Agora suponha que os dois dados são idênticos, a tal ponto que não podemos dizer qual é um e qual é o outro. Qual é a probabilidade de que um deles mostre 2, e o outro 3?

13.8 Inferência bayesiana

Combinando as fórmulas (13.18) e (13.19), obtemos a equação

$$\Pr(P|Q) = \frac{\Pr(Q|P)\ \Pr(P)}{\Pr(Q)} \tag{13.20}$$

Esta fórmula é conhecida como *regra de Bayes* ou *teorema de Bayes*, desenvolvida pelo matemático inglês Thomas Bayes (≈1702–1761) e, independentemente, pelo matemático francês Pierre-Simon Laplace (1749–1827). Ela é geralmente usada quando se quer obter a probabilidade $\Pr(P|Q)$ de uma possível causa P, sabendo-se que uma consequência Q ocorreu, a partir da probabilidade condicional inversa $\Pr(Q|P)$ (de que essa consequência produza essa causa). Este raciocínio probabilístico é conhecido como *inferência bayesiana* ou *dedução bayesiana*.

Por exemplo, considere uma coleção de caixas quadradas e redondas, cada uma contendo uma bola que pode ser azul ou branca. Suponha que há igual número de caixas de cada formato, sendo que há bolas azuis em metade das caixas quadradas, mas em apenas 10% das caixas redondas. Imagine que alguém escolheu uma caixa ao acaso, e encontrou nela uma bola azul. Qual a probabilidade de que ele tenha escolhido uma caixa quadrada? E se a bola for branca?

Se não tivéssemos a informação sobre a bola, seria razoável supor que a caixa era quadrada com probabilidade 1/2. Porém, como bolas brancas são mais comuns nas caixas redondas, intuitivamente, a informação de que a bola era branca aumenta a probabilidade de que a caixa seja redonda.

Para calcular essas probabilidades, vamos denotar por Q, R, A e B as afirmações "a caixa era quadrada", "a caixa era redonda", "a bola era azul" e "a bola era branca", respectivamente. Pelo

enunciado do problema, temos

$$\begin{array}{llll} \Pr(Q) = \frac{1}{2} & & \Pr(R) = \frac{1}{2} \\ \Pr(A|Q) = \frac{1}{2} & & \Pr(B|Q) = \frac{1}{2} \\ \Pr(A|R) = \frac{1}{10} & & \Pr(B|R) = \frac{9}{10} \end{array}$$

O que se pede são as probabilidade condicionais $\Pr(Q|A)$ e $\Pr(Q|B)$. Para aplicar a fórmula (13.18), precisamos determinar $\Pr(B)$ e $\Pr(Q \wedge B)$. Para chegar lá, temos que calcular as probabilidades de todas as combinações válidas dessas afirmações. Aplicando a fórmula (13.19) temos

$$\begin{array}{rcccccl} \Pr(Q \wedge A) & = & \Pr(A \wedge Q) & = & \Pr(A|Q)\Pr(Q) & = & \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} \\ \Pr(Q \wedge B) & = & \Pr(B \wedge Q) & = & \Pr(B|Q)\Pr(Q) & = & \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} \\ \Pr(R \wedge A) & = & \Pr(A \wedge R) & = & \Pr(A|R)\Pr(R) & = & \frac{9}{10} \cdot \frac{1}{2} = \frac{9}{20} \\ \Pr(R \wedge B) & = & \Pr(B \wedge R) & = & \Pr(B|R)\Pr(R) & = & \frac{1}{10} \cdot \frac{1}{2} = \frac{1}{20} \end{array}$$

Daí tiramos

$$\begin{array}{rcccl} \Pr(A) & = & \Pr(B \wedge Q) + \Pr(B \wedge R) & = & \frac{1}{4} + \frac{1}{20} = \frac{3}{10} \\ \Pr(B) & = & \Pr(A \wedge Q) + \Pr(A \wedge R) & = & \frac{1}{4} + \frac{9}{20} = \frac{7}{10} \end{array}$$

portanto

$$\begin{array}{rcccccl} \Pr(Q|A) & = & \frac{\Pr(Q \wedge A)}{\Pr(A)} & = & \frac{\Pr(A|Q)\Pr(Q)}{\Pr(A)} & = & \frac{1/4}{3/10} = \frac{5}{6} \approx 0,833 \\ \Pr(Q|B) & = & \frac{\Pr(Q \wedge B)}{\Pr(B)} & = & \frac{\Pr(B|Q)\Pr(Q)}{\Pr(B)} & = & \frac{1/4}{7/10} = \frac{5}{14} \approx 0,357 \end{array}$$

Observe que a informação adicional "a bola sorteada é azul" aumenta a probabilidade de que a caixa escolhda seja quadrada, de $0,5$ a $0,833$

Generalizando este exemplo, suponha que temos m afirmações $A_1, A_2, \ldots A_m$, os *antecedentes*, exaustivas e mutuamente exclusivas, cujo valor lógico pode influir na probabilidade de outras n afirmações $B_1, B_2, \ldots B_n$, os *consequentes*, também exaustivas e mutuamente exclusivas. As afirmações A_i podem ser as alternativas possíveis para um evento-causa (no exemplo acima, a escolha caixa, quadrada ou redonda), e as afirmações B_j a possíveis consequências do mesmo (a cor da bola). Suponha que atribuímos probabilidades $\Pr(A_i)$ para cada antecedente A_i, sem levar em conta as afirmações B_j; e temos também a probabilidade condicional $\Pr(B_j|A_i)$ de cada consequente, dado o antecedente. Uma vez sabido que um determinado B_j é verdadeiro, a probabilidade de cada A_i passa a ser

$$\Pr(A_i|B_j) = \frac{\Pr(A_i \wedge B_j)}{\Pr(B_j)} = \frac{\Pr(A_i \wedge B_j)}{\sum_k = 1^m \Pr(B_j \wedge A_k)} = \frac{\Pr(B_j|A_i)\Pr(A_i)}{\sum_k^m \Pr(B_j|A_k)\Pr(A_k)} \tag{13.21}$$

Note que para aplicar a fórmula (13.21) precisamos atribuir uma probabilidade $\Pr(A_i)$ a cada antecedente, independente de qual consequente é verdadeiro. O fator $\Pr(A_i)$ nesta fórmula é chamado de *probabilidade a priori* do antecedente A_i, enquanto que o resultado $\Pr(A_i|B_j)$ é sua *probabilidade a posteriori*.

A influência das probabilidades *a priori* $\Pr(A_i)$ é uma característica essencial da inferência bayesiana. Elas podem ser vistas como "preconceitos" que temos a respeito das afirmações A_i, antes de olharmos para as evidências B_j. A fórmula portanto explicita quantitativamente a constatação comum, de que nossos preconceitos sempre afetam nossa interpretação dos fatos.

Exercício 13.18: Suponha que há duas gavetas em uma mesa de jogo. Uma delas contém um dado "honesto", que dá cada valor de 1 a 6 com igual probabilidade $1/6$; a outra contém um dado "viciado", que dá o valor 6 com probabilidade $1/2$, e os valores de 1 a 5 com probabilidade $1/10$ cada.

1. Uma pessoa escolhe (sem você ver) um desses dois dados. Na falta de informações, você atribui a probabilidade *a priori* $1/2$ de que esse dado seja viciado. O dado é então lançado e o resultado é 6. Como fica a probabilidade de que o dado seja viciado?
2. Suponha agora que a pessoa seja um notório vigarista, de modo que, mesmo antes de lançar, você dá 90% de chance de que ele tenha escolhido o dado viciado. Como fica essa probabilidade depois que o dado foi lançado, com resultado 6?
3. Finalmente suponha que você confia na pessoa e portanto acredita que ela escolheu o dado honesto, com 90% de probabilidade. Como fica sua confiança nessa hipótese depois que o dado deu 6?

Exercício 13.19: Uma moeda é lançada 10 vezes seguidas, e o resultado é sempre cara. Talvez a moeda seja normal, e esse resultado seja coincidência; ou talvez ela seja uma moeda anormal, com cara dos dois lados. Suponha que a probabilidade *a priori* da moeda ser anormal é p. Qual é a probabilidade *a posteriori*, depois desses 10 lances? Faça um gráfico dessa probabilidade em função de p.

.

13.9 Teoria da informação

Hoje em dia todos conhecem o conceito de *bit* e outras unidades derivadas, como *byte* (8 bits), *megabyte* (10^6 ou 2^20 bytes, conforme o contexto), *gigabyte* (10^9 ou 2^30 bytes) etc. Em geral esses conceitos são usados para descrever tamanhos de arquivos, capacidade de memória, taxas de transmissão, etc. Porém é necessário distinguir entre a *capacidade de armazenamento de informação* de tais sistemas, e a *quantidade de informação* contida neles em determinado momento. Este segundo conceito é o centro da *teoria da informação*, desenvolvida principalmente em meados do século 20 pelo matemático e engenheiro americano Claude Shannon (1916–2001).

13.9.1 Capacidade de informação

Considere um sistema físico (real ou imaginário) que em qualquer momento pode assumir um único estado dentre uma coleção finita de estados possíveis; sendo que esse estado pode ser identificado com precisão por algum tipo de teste ou medida. Por exemplo, uma moeda sobre uma mesa, que pode estar na posição 'cara' ou 'coroa'; um dado de jogar, que pode estar virado com qualquer face para cima, de 1 a 6; uma chave elétrica, que pode estar 'desligada' ou 'ligada'; um fio elétrico, que pode estar a zero volts ou a +5 volts; uma barra de ferro, que pode estar magnetizada em dois sentidos diferentes; e assim por diante. Tal objeto é dito um *sistema discreto*.

Suponha que o sistema tem apenas dois estados possíveis (ou seja, é um *sistema binário*). Por definição, a capacidade de informação de tal sistema é 1 bit. Se o sistema tem 2^b estados possíveis, sua capacidade é b bits. Observe que podemos numerar os estados de tal sistema em base 2 usando

b algarismos, cada qual 0 ou 1: — $0\cdots00 = 0$, $0\cdots01 = 1$, $0\cdots10 = 2$, $0\cdots11 = 3$, ..., $1\cdots11 = 2^b - 1$. Daí o nome "bit", que é abreviação do inglês *binary digit*.

Mais geralmente, se o número de estados possíveis n, a capacidade de informação é definida como $\log_2 n = (\ln n)/(\ln 2)$, o logaritmo de n na base 2. Assim, por exemplo, a capacidade de informação de um dado de jogar, em repouso sobre a mesa, é $\log_2 6 = 2,5849625007\ldots$ bits. Note que, se n não é uma potência de 2, a capacidade em bits não é um número inteiro (e, na verdade, é um número irracional). Note também que se o sistema tem apenas um estado possível, sua capacidade de armazenar informação é (como se pode esperar) zero bits.

Esta definição implica na seguinte propriedade:

Teorema 13.2: Se um sistema S consiste de dois sub-sistemas discretos A e B independentes (no sentido de que cada estado possível de A pode co-existir com qualquer estado possível de B, e vice-versa), então a capacidade de S é a soma das capacidades de A e de B.

Exercício 13.20: Determine a capacidade de informação dos seguintes sistemas:

1. Um odômetro (mostrador de quilometragem) de automóvel com 6 algarismos decimais.
2. Um dado em forma de octaedro, com faces numeradas de 1 a 8, em respouso sobre a mesa.
3. Uma cadeia de DNA com 100 elementos (*nucleotídeos*), cada qual podendo ter quatro estruturas químicas possíveis — adenosina (A), timina (T), guanina (G), ou citosina (C).

Exercício 13.21: Determine a capacidade de informação dos seguintes sistemas, constituídos de 4 moedas, cada qual podendo ser de 5, 10, 25, ou 50 centavos, que somente podem ser distinguidas pelo seu valor:

1. Uma pilha, em qualquer ordem.
2. Uma pilha, em ordem crescente de valor.
3. Uma coleção em um saco.
4. Uma pilha onde todas as moedas tem o mesmo valor.

Exercício 13.22: Refaça o exercício 13.21, supondo que todas as moedas de mesmo valor estão marcadas com letras distintas entre 'A' e 'D'. Assim, por exemplo, na alternativa 1, as moedas poderiam ser, na ordem, $(10, D), (25, C), (10, B), (10, C)$ mas não poderiam ser $(10, D), (25, C), (10, B), (10, D)$.

Exercício 13.23: Qual é a capacidade de informação de uma carta retirada de um baralho com 13 cartas? E de um baralho com 52 cartas? Se acrescentarmos um coringa ao baralho, de quanto aumenta a capacidade, em cada caso?

13.9.2 Quantidade de informação

A capacidade de informação de um sistema discreto diz apenas o limite máximo de informação que pode ser armazenada nele. Porém, dependendo de como o sistema é usado, nem toda a capacidade pode ser utilizada.

Por exemplo, considere uma lâmpada que, ao meio-dia, pode estar acesa ou apagada conforme o sol tenha nascido ou não naquele dia. Embora a capacidade de informação desse sistema seja 1 bit, intuitivamente a notícia de que essa lâmpada está acesa não traz muita informação. Por outro lado, uma lâmpada que indica se está chovendo ou não fora do prédio parece fornecer mais informação — muito embora sua *capacidade* de informação seja exatamente a mesma.

A diferença estes dois exemplos está na probabilidade que atribuímos aos dois estados do sistema. No primeiro caso, é natural atribuir probabilidade bem próxima a 1 à afirmação "a lâmpada está acesa". (A menos que sejamos extremamente pessimistas!) Por isso, a notícia de que essa informação é verdadeira não muda muito nosso estado de conhecimento. Já, no segundo exemplo, faz sentido atribuir probabilidade bem menor que 1 a essa afirmação. (A menos que estejamos na Bolívia, onde nunca chove!)

Para tornar esta intuição mais precisa, suponha que X é uma variável aleatória que pode assumir um certo valor v. A *quantidade de informação* trazida pela notícia "o valor de X é v" é, por definição,

$$Q(X = v) = \log_2 \frac{1}{\Pr(X = v)} = -\log_2 \Pr(X = v)$$

Este valor, como a capacidade de informação, é medido em bits, e nunca é negativo. Em particular, se X pode assumir n valores distintos com igual probabilidade $\Pr(X = v) = 1/n$, a quantidade de informação que recebemos quando ficamos sabendo o valor de X (qualquer valor de X) é exatamente $Q(X = v) = \log_2 n$ bits — ou seja, a capacidade da variável X.

Porém, se as probabilidades dos valores de X não são iguais, a quantidade de informação pode ser menor ou maior, dependendo do valor. Por exemplo:

> **Exemplo 13.3:** Suponha que um dado está para ser lançado, e X é uma variável que vale 100 se o resultado do dado é 1, e 200 caso contrário. Então as notícias "$X = 100$" e "$X = 200$" carregam as seguintes quantidades de informação:
>
> $$\begin{aligned} Q(X = 100) &= -\log_2 \Pr(X = 100) = -log_2 \frac{1}{6} \approx 2,5849625\ldots \\ Q(X = 200) &= -\log_2 \Pr(X = 200) = -log_2 \frac{5}{6} \approx 0,2630344\ldots \end{aligned}$$

Neste exemplo, observe que a notícia "$X = 200$" traz muito menos informação do que a notícia "$X = 100$", porque tem probabilidade maior — $5/6$ em vez de $1/6$.

13.9.3 Quantidade esperada de informação

No exemplo 13.3, observe também que a notícia "$X = 100$" traz mais que 1 bit de informação — muito embora a variável X tenha apenas dois valores possíveis, e portanto tenha apenas 1 bit de capacidade.

Este paradoxo é resolvido se considerarmos a *quantidade esperada de informação*, ou *entropia*, da variável X. Ou seja, a quantia

$$\mathcal{H}(X) = \sum_v \Pr(X = v)\, \mathcal{Q}(X = v) = \sum_v -\Pr(X = v) \log_2 \Pr(X = v) \tag{13.22}$$

Nesta fórmula, o índice v do somatório assume todos os valores possíveis da variável X. Observe que, como na fórmula (13.11), cada termo desta soma é a quantidade de informação trazida pela notícia "$X = v$", vezes a probabilidade de recebermos essa notícia. Pode-se verificar que $\mathcal{H}(X)$, assim como cada termo $\mathcal{Q}(X = v)$, é um valor real não negativo.

No exemplo 13.3, a quantidade esperada de informação que recebemos ao conhecer o valor de X é

$$\begin{aligned} \mathcal{H}(X) &= \Pr(X = 100)\, \mathcal{Q}(X = 100) + \Pr(X = 200)\, \mathcal{Q}(X = 200) \\ &= \tfrac{1}{6} \log_2 \tfrac{6}{1} + \tfrac{5}{6} \log_2 \tfrac{6}{5} \\ &\approx \tfrac{1}{6} 2,5849625\ldots + \tfrac{5}{6} 0,2630344\ldots \\ &\approx 0,65002241\ldots \end{aligned}$$

Observe que, embora a notícia "$X = 100$" forneça mais de 2,5 bits de informação, ela é muito menos provável que a notícia "$X = 200$", que fornece menos que 0,27 bits de informação. Assim, a quantidade esperada de informação que ganhamos ao saber o valor de X é cerca de 0,65 bits, ou seja abaixo da capacidade de X (1 bit). Esta última observação é um resultado importante:

> **Teorema 13.3:** Se uma variável aleatória X pode assumir n valores distintos, então a quantidade esperada de informação que ganhamos conhecendo o valor de X é no máximo a capacidade de X, $\log_2 n$; e é exatamente $\log_2 n$ apenas quando todos esses valores podem ocorrer com igual probabilidade $1/n$.

Devido a este teorema, a fórmula (13.22) é muito usada para medir a "uniformidade" da distribuição de probabilidades de uma variável aleatória X. O valor de $\mathcal{H}(X)$ varia entre 0 e $\log_2 n$, onde n é o número de valores possíveis de X. Quanto maior $\mathcal{H}(X)$, mais uniforme a distribuição. Na verdade, a fórmula (13.22) pode ser usada com qualquer lista de n valores reais $p_0, p_1, \ldots p_{n-1}$ não negativos cuja soma é 1.

Observe que se X tem uma distribuição degenerada — com $\Pr(X = v) = 1$ para um único valor v, e zero para os demais valores — então $\mathcal{H}(X)$ é zero. Ou seja, se temos certeza de qual vai ser o valor de X, nossa expectativa é que a revelação desse valor não vai nos trazer nenhuma informação.

Referências Bibliográficas

[1] Béla Bollobás. *Modern Graph Theory*. Springer, 1998.

[2] J. A. Bondy and U. S. R. Murty. *Graph Theory with Applications*. MacMillan, London, 1976.

[3] J. A. Bondy and U. S. R. Murty. *Graph Theory*. Springer, 2008.

[4] Thomas H. Cormen, Charles E. Leiserson, and Ronald L. Rivest. *Introduction to Algorithms*. MIT Press, 1989.

[5] Ronald L. Graham, Donald E. Knuth, and Oren Patashnik. *Matemática Concreta: Fundamentos para Ciência da Computação*. LTC, 1995. Segunda edição.

[6] Paul R. Halmos. *Teoria Ingênua dos Conjuntos*. Editora da USP, 1960.

[7] Frank Harary. *Graph Theory*. Addison Wesley, 1972.

[8] John M. Harris, Jeffry L. Hirst, and Michael J. Mossinghoff. *Combinatorics and Graph Theory*. Springer, 2000.

[9] Thomas L. Heath. *The Thirteen Books of Euclid's Elements*. Dover, 1956. Segunda edição.

[10] David C. Kurtz. *Foudations of Abstract Mathematics*. McGraw-Hill, 1992.

[11] Luiz Henrique Jacy Monteiro. *Elementos de Álgebra*. Ao Livro Técnico, 1969.

[12] Kenneth H. Rosen. *Discrete Mathematics and Its Applications*. McGraw-Hill, 2003. Quinta edição.

[13] J. Plínio O. Santos, Margarida P. Mello, and Idani T. C. Murari. *Introdução à Análise Combinatória*. Editora da UNICAMP, 1995.

[14] Daniel J. Velleman. *How to Prove It: A Structured Approach*. Cambridge University Press, 2006. Segunda edição.