



Revisão de alguns conceitos estatísticos

Este apêndice fornece uma introdução bem resumida de alguns dos conceitos estatísticos encontrados no texto. A discussão não é rigorosa e nenhuma prova é fornecida, porque um grande número de livros excelentes sobre estatística faz muito bem esse trabalho. Algumas dessas obras estão listadas no final deste apêndice.

A.1 Operadores somatório e de produto

A letra maiúscula grega $\sum$ (sigma) é utilizada para indicar somatório. Assim,

$$\sum_{i=1}^n x_i = x_1 + x_2 + \cdots + x_n$$

Algumas das propriedades importantes do operador somatório são:

1. $\sum_{i=1}^n k = nk$, em que k é constante. Assim, $\sum_{i=1}^4 3 = 4 \cdot 3 = 12$.
2. $\sum_{i=1}^n kx_i = k \sum_{i=1}^n x_i$, em que k é uma constante.
3. $\sum_{i=1}^n (a + bx_i) = na + b \sum_{i=1}^n x_i$, em que a e b são constantes e aplicam-se as propriedades 1 e 2 anteriores.
4. $\sum_{i=1}^n (x_i + y_i) = \sum_{i=1}^n x_i + \sum_{i=1}^n y_i$.

O operador somatório também pode ser estendido às somas múltiplas. Assim, $\sum\sum$, o operador duplo somatório, é definido como:

$$\begin{aligned} \sum_{i=1}^n \sum_{j=1}^m x_{ij} &= \sum_{i=1}^n (x_{i1} + x_{i2} + \cdots + x_{im}) \\ &= (x_{11} + x_{21} + \cdots + x_{n1}) + (x_{12} + x_{22} + \cdots + x_{n2}) \\ &\quad + \cdots + (x_{1m} + x_{2m} + \cdots + x_{nm}) \end{aligned}$$

Algumas das propriedades de $\sum\sum$ são:

1. $\sum_{i=1}^n \sum_{j=1}^m x_{ij} = \sum_{j=1}^m \sum_{i=1}^n x_{ij}$; a ordem, na qual o duplo somatório é executada, é permutável.
2. $\sum_{i=1}^n \sum_{j=1}^m x_i y_j = \sum_{i=1}^n x_i \sum_{j=1}^m y_j$.

$$3. \sum_{i=1}^n \sum_{j=1}^m (x_{ij} + y_{ij}) = \sum_{i=1}^n \sum_{j=1}^m x_{ij} + \sum_{i=1}^n \sum_{j=1}^m y_{ij}.$$

$$4. \left[\sum_{i=1}^n x_i \right]^2 = \sum_{i=1}^n x_i^2 + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n x_i x_j = \sum_{i=1}^n x_i^2 + 2 \sum_{i < j} x_i x_j.$$

O operador produto é definido como

$$\prod_{i=1}^n x_i = x_1 \cdot x_2 \cdots x_n$$

Assim,

$$\prod_{i=1}^3 x_i = x_1 \cdot x_2 \cdot x_3$$

A.2 Espaço amostral, pontos amostrais e eventos

O conjunto de todos os resultados possíveis de um experimento aleatório, ou ao acaso, é chamado **população**, ou **espaço amostral**, e cada membro desse espaço amostral é chamado de **ponto amostral**. No experimento de lançar duas moedas, o espaço amostral consiste nesses possíveis quatro resultados: *HH*, *HT*, *TH* e *TT*, em que *HH* significa cara no primeiro lançamento e coroa no segundo e assim por diante. Cada uma das ocorrências anteriores constitui um ponto amostral.

Um **evento** é um subconjunto do espaço amostral. Se denotarmos *A* a ocorrência de uma cara e de uma coroa, então, dos possíveis resultados anteriores, apenas dois pertencem a *A*, ou seja, *HT* e *TH*. Nesse caso, *A* constitui um evento. Da mesma maneira, a ocorrência de duas caras em um lançamento de duas moedas é um evento. Diz-se que eventos são **mutuamente exclusivos** se a ocorrência de um eliminar a ocorrência do outro. Se, no exemplo anterior, ocorre *HH*, a ocorrência do evento *HT* ao mesmo tempo não é possível. Diz-se que eventos são (coletivamente) **exaustivos** se exaurem todas os possíveis resultados de um experimento. No exemplo, os eventos (a) duas caras, (b) duas coroas e (c) uma coroa, uma cara exaure todos os resultados; daí eles serem eventos (coletivamente) exaustivos.

A.3 Probabilidade e variáveis aleatórias

Probabilidade

Seja *A* um evento em um espaço amostral. Por $P(A)$, a probabilidade do evento *A*, entendemos a proporção de vezes que o evento *A* ocorrerá em repetidas tentativas de um experimento. Como alternativa, em um total de *n* possíveis resultados igualmente prováveis de um experimento, se *m* deles são favoráveis à ocorrência do evento *A*, definimos a razão m/n como a **frequência relativa** de *A*. Para valores maiores de *n*, essa frequência relativa fornecerá uma aproximação bastante boa da probabilidade de *A*.

Propriedades da probabilidade

$P(A)$ é uma função de valor real¹ e possui essas propriedades:

1. $0 \leq P(A) \leq 1$ para cada *A*.
2. Se *A*, *B*, *C*, ... constituem um conjunto exaustivo de eventos, $P(A + B + C + \cdots) = 1$, em que $A + B + C$ significa *A* ou *B* ou *C* e assim por diante.

¹ Uma função cujo domínio e alcance são subconjuntos de números reais é comumente referida como função de valor real. Para mais detalhes, veja CHIANG, Alpha C. *Fundamental methods of mathematical economics*. 3. ed. Nova York: McGraw-Hill, 1984. cap. 2.

3. Se $A, B, C, \dots$ são eventos mutuamente exclusivos,

$$P(A + B + C + \dots) = P(A) + P(B) + P(C) + \dots$$

EXEMPLO 1

Considere o experimento de lançar um dado numerado de 1 a 6. O espaço amostral consiste nos resultados 1, 2, 3, 4, 5 e 6. Os seis eventos, portanto, exaurem totalmente o espaço amostral. A probabilidade de qualquer um desses números aparecer é de $1/6$, uma vez que há seis resultados igualmente prováveis e qualquer um deles possui uma chance igual de acontecer. Na medida em que 1, 2, 3, 4, 5 e 6 formam um conjunto exaustivo de eventos, $P(1 + 2 + 3 + 4 + 5 + 6) = 1$ em que 1, 2, 3... indica a probabilidade do número 1 ou do número 2 ou do número 3 etc. E, na medida em que 1, 2, ..., 6 são eventos mutuamente exclusivos no sentido de que dois números não podem ocorrer simultaneamente, $P(1 + 2 + 3 + 4 + 5 + 6) = P(1) + P(2) + \dots + P(6) = 1$.

Variáveis aleatórias

Uma variável cujo valor é determinado pelo resultado de um experimento aleatório é chamada de **variável aleatória**. As variáveis aleatórias são normalmente denotadas pelas letras maiúsculas X, Y, Z etc., e os valores assumidos por elas são indicados pelas letras minúsculas x, y, z etc.

Uma variável aleatória pode ser tanto **discreta** como **contínua**. Uma variável aleatória discreta pode assumir apenas um número finito (ou infinito enumerável) de valores.² Por exemplo, ao lançarmos dois dados, cada um com números de 1 a 6, se definirmos a variável aleatória X como a soma dos números mostrados nos dois dados, X terá um desses valores: 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 ou 12. Portanto, é uma variável aleatória discreta. Uma variável aleatória contínua, por outro lado, é aquela que pode assumir qualquer valor em algum intervalo dos valores. A altura de um indivíduo é uma variável contínua — em uma amplitude de, por exemplo, 60 a 65 polegadas, ele pode ter qualquer valor, dependendo da precisão da medição.

A.4 Função de densidade de probabilidade (FDP)**Função de densidade de probabilidade de uma variável aleatória discreta**

Seja X uma variável aleatória discreta que toma valores distintos $x_1, x_2, \dots, x_n, \dots$. Então, a função

$$\begin{aligned} f(x) &= P(X = x_i) \quad \text{para } i = 1, 2, \dots, n, \dots \\ &= 0 \quad \text{para } x \neq x_i \end{aligned}$$

é chamada **função de densidade de probabilidade discreta** (FDP) de X , em que $P(X = x_i)$ significa a probabilidade de que a variável aleatória discreta X tome o valor de x_i .

EXEMPLO 2

Ao lançarem dois dados, a variável aleatória X , a soma dos números apresentados nos dois dados, pode assumir um dos 11 valores exibidos. A FDP dessa variável pode ser representada como se segue (Veja também a Figura A.1):

$$\begin{array}{cccccccccccc} x = & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ f(x) = & \left(\frac{1}{36}\right) & \left(\frac{2}{36}\right) & \left(\frac{3}{36}\right) & \left(\frac{4}{36}\right) & \left(\frac{5}{36}\right) & \left(\frac{6}{36}\right) & \left(\frac{5}{36}\right) & \left(\frac{4}{36}\right) & \left(\frac{3}{36}\right) & \left(\frac{2}{36}\right) & \left(\frac{1}{36}\right) \end{array}$$

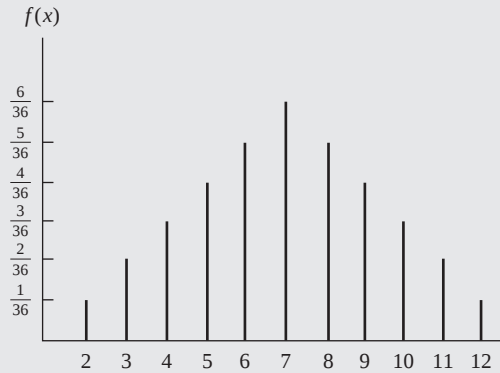
(Continua)

² Para uma discussão simples da noção de conjuntos infinitos enumeráveis, veja ALLEN, R. G. D. *Basic mathematics*. Londres: Macmillan, 1964. p. 104.

EXEMPLO 2 (Continuação)

Essas probabilidades podem ser facilmente verificadas. Em todas, há 36 resultados possíveis, dos quais um é favorável ao número 2, dois são favoráveis ao número 3 (uma vez que a soma 3 pode ocorrer tanto no caso de 1 no primeiro dado, como com 2 no segundo dado ou com 2 no primeiro dado e 1 no segundo dado), e assim por diante.

FIGURA A.1 Função de densidade da variável aleatória discreta do Exemplo 2.



Função de densidade de probabilidade de uma variável aleatória contínua

Seja X uma variável aleatória contínua. Então, $f(x)$ será a função de densidade de probabilidade de X se as seguintes condições forem satisfeitas:

$$f(x) \geq 0$$

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

$$\int_a^b f(x) dx = P(a \leq x \leq b)$$

em que $f(x) dx$ é conhecido como o *elemento da probabilidade* (a probabilidade associada a um pequeno intervalo de uma variável contínua) e $P(a \leq x \leq b)$ indica a probabilidade de que X situe-se no intervalo entre a e b . Geometricamente, temos a Figura A.2.

Para uma variável aleatória contínua, em contraste com uma variável aleatória discreta, a probabilidade de que X assumira um valor específico é zero;³ a probabilidade de tal variável é mensurada apenas para uma dada amplitude, ou intervalo, tal como (a, b) , representado na Figura A.2.

EXEMPLO 3

Considere a seguinte função de densidade:

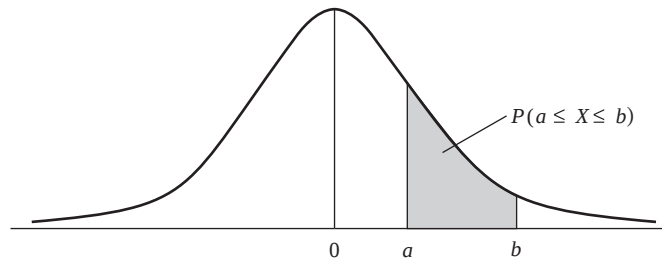
$$f(x) = \frac{1}{9}x^2 \quad 0 \leq x \leq 3$$

Pode ser prontamente verificado que $f(x) \geq 0$ para todos os x no intervalo de 0 a 3 e que $\int_0^3 \frac{1}{9}x^2 dx = 1$. (Nota: a integral é $(\frac{1}{27}x^3|_0^3) = 1$.) Se quisermos avaliar a função de densidade de probabilidade anterior entre, por exemplo, 0 e 1, obtemos $\int_0^1 \frac{1}{9}x^2 dx = (\frac{1}{27}x^3|_0^1) = \frac{1}{27}$; ou seja, a probabilidade de que x situa-se entre 0 e 1 é $1/27$.

³ $\int_a^a f(x) dx = 0$.

FIGURA A.2

Função de densidade
de uma variável
aleatória contínua.



Funções de densidade de probabilidade conjunta

Função de densidade de probabilidade conjunta discreta

Sejam X e Y duas variáveis aleatórias discretas. Então, a função

$$\begin{aligned} f(x, y) &= P(X = x \text{ e } Y = y) \\ &= 0 \quad \text{onde } X \neq x \text{ e } Y \neq y \end{aligned}$$

é conhecida como **função de densidade de probabilidade conjunta discreta** e fornece a probabilidade (conjunta) de que X tome o valor de x e Y tome o valor de y .

EXEMPLO 4

A seguinte tabela oferece a função de densidade de probabilidade conjunta das variáveis discretas X e Y :

		X			
		-2	0	2	3
Y	3	0,27	0,08	0,16	0
	6	0	0,04	0,10	0,35

Essa tabela mostra que a probabilidade de que X tome o valor de -2 enquanto Y simultaneamente assume o valor de 3 é de $0,27$ e que a probabilidade de que X tome o valor de 3 enquanto Y toma o valor de 6 é de $0,35$ e assim por diante.

Função de densidade de probabilidade marginal

Em relação a $f(x, y)$, $f(x)$ e $f(y)$ são chamadas de funções de densidade **individual** ou **marginal**, as funções de densidade de probabilidade. Essas funções de densidade de probabilidade marginais são derivadas como se segue:

$$f(x) = \sum_y f(x, y) \quad \text{FDP marginal de } X$$

$$f(y) = \sum_x f(x, y) \quad \text{FDP marginal de } Y$$

em que, por exemplo, $\sum_y$ significa a soma de todos os valores de Y e $\sum_x$ a soma de todos os valores de X .

EXEMPLO 5

Considere os dados fornecidos no Exemplo 4. A função de densidade de probabilidade marginal é obtida como se segue:

$$\begin{aligned}f(x = -2) &= \sum_y f(x, y) = 0,27 + 0 = 0,27 \\f(x = 0) &= \sum_y f(x, y) = 0,08 + 0,04 = 0,12 \\f(x = 2) &= \sum_y f(x, y) = 0,16 + 0,10 = 0,26 \\f(x = 3) &= \sum_y f(x, y) = 0 + 0,35 = 0,35\end{aligned}$$

De forma semelhante, a função de densidade de probabilidade marginal de Y é obtida como:

$$\begin{aligned}f(y = 3) &= \sum_x f(x, y) = 0,27 + 0,08 + 0,16 + 0 = 0,51 \\f(y = 6) &= \sum_x f(x, y) = 0 + 0,04 + 0,10 + 0,35 = 0,49\end{aligned}$$

Como esse exemplo demonstra, para obter uma função de densidade de probabilidade marginal de X , adicionamos os números da coluna, e, para obter a função de densidade de probabilidade marginal de Y , adicionamos os números das linhas. Perceba que $\sum_x f(x)$ que cobre todos os valores de X é 1, assim como $\sum_y f(y)$ que cobre todos os valores de Y (por quê?).

Função de densidade de probabilidade condicional

Como observado no Capítulo 2, na análise de regressão, frequentemente estamos interessados no estudo do comportamento de uma variável condicional com relação ao(s) valor(es) de outra(s) variável(is). Isso pode ser feito considerando a função de densidade de probabilidade condicional. A função

$$f(x | y) = P(X = x | Y = y)$$

é conhecida como **função de densidade de probabilidade condicional** de X ; ela apresenta a probabilidade de que X assumo o valor de x posto que Y assumiu o valor de y . De forma semelhante,

$$f(y | x) = P(Y = y | X = x)$$

que apresenta a *FDP condicional de Y* .

As funções de densidade de probabilidade condicionais podem ser obtidas como se segue:

$$\begin{aligned}f(x | y) &= \frac{f(x, y)}{f(y)} && \text{FDP condicional de } X \\f(y | x) &= \frac{f(x, y)}{f(x)} && \text{FDP condicional de } Y\end{aligned}$$

Como as expressões anteriores demonstram, a função de densidade de probabilidade condicional de uma variável pode ser expressa como a razão da função de densidade de probabilidade conjunta à função de densidade de probabilidade marginal de outra variável (condicionante).

EXEMPLO 6

Continuando com os Exemplos 4 e 5, calculemos as seguintes probabilidades condicionais:

$$f(X = -2 | Y = 3) = \frac{f(X = -2, Y = 3)}{f(Y = 3)} = 0,27/0,51 = 0,53$$

Perceba que a probabilidade incondicional $f(X = -2)$ é 0,27, mas se Y assumiu o valor de 3, a probabilidade de que X tome o valor de -2 é de 0,53.

$$f(X = 2 | Y = 6) = \frac{f(X = 2, Y = 6)}{f(Y = 6)} = 0,10/0,49 = 0,20$$

Novamente, note que a probabilidade incondicional de que X tome o valor de 2 é de 0,26, o que é diferente de 0,20, que é o seu valor se Y assume o valor de 6.

Independência estatística

As duas variáveis aleatórias X e Y são estatisticamente independentes se, e somente se,

$$f(x, y) = f(x)f(y)$$

ou seja, se a função de densidade de probabilidade conjunta puder ser expressa como o produto das funções de densidade de probabilidade marginais.

EXEMPLO 7

Uma bolsa contém três bolas numeradas 1, 2 e 3. Duas bolas são retiradas aleatoriamente, com reposição, dessa bolsa (a primeira bola retirada é recolocada antes que a segunda seja retirada). Seja X o número da primeira bola retirada e Y o número da segunda bola retirada. A seguinte tabela apresenta a FDP conjunta de X e Y .

		X		
		1	2	3
Y	1	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$
	2	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$
	3	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$

Agora, $f(X = 1, Y = 1) = \frac{1}{9}$, $f(X = 1) = \frac{1}{3}$ (obtido pela soma da primeira coluna) e $f(Y = 1) = \frac{1}{3}$ (obtido pela soma da primeira linha). Uma vez que $f(X, Y) = f(X)f(Y)$ neste exemplo, podemos dizer que as duas variáveis são estatisticamente independentes. Pode ser facilmente verificado que, para qualquer outra combinação de valores X e Y dados nessa tabela, a função de densidade de probabilidade conjunta pode ser representada como o produto das funções de densidade de probabilidade individuais.

Pode-se demonstrar que as variáveis X e Y do Exemplo 4 não são estatisticamente independentes, na medida em que o produto das duas funções de densidade de probabilidade marginal não é igual à função de densidade de probabilidade conjunta. (Nota: $f(X, Y) = f(X)f(Y)$ deve ser verdadeiro para todas as combinações de X e Y para que as duas variáveis sejam estatisticamente independentes).

Função de densidade de probabilidade conjunta contínua

A função de densidade de probabilidade $f(x, y)$ de duas variáveis contínuas X e Y é tal que

$$\begin{aligned} f(x, y) &\geq 0 \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy &= 1 \\ \int_c^d \int_a^b f(x, y) dx dy &= P(a \leq x \leq b, c \leq y \leq d) \end{aligned}$$

EXEMPLO 8

Considere a seguinte função de densidade de probabilidade

$$f(x, y) = 2 - x - y \quad 0 \leq x \leq 1; 0 \leq y \leq 1$$

É óbvio que $f(x, y) \geq 0$. Além do mais⁴,

$$\int_0^1 \int_0^1 (2 - x - y) dx dy = 1$$

A função de densidade de probabilidade marginal de X e Y pode ser obtida como

$$f(x) = \int_{-\infty}^{\infty} f(x, y) dy \quad \text{FDP marginal de } X$$

$$f(y) = \int_{-\infty}^{\infty} f(x, y) dx \quad \text{FDP marginal de } Y$$

EXEMPLO 9

As duas funções de densidade de probabilidade marginais da função de densidade de probabilidade conjunta dadas no Exemplo 8 são as seguintes:

$$\begin{aligned} f(x) &= \int_0^1 f(x, y) dy = \int_0^1 (2 - x - y) dy \\ &= \left(2y - xy - \frac{y^2}{2} \right) \Big|_0^1 = \frac{3}{2} - x \quad 0 \leq x \leq 1 \end{aligned}$$

(Continua)

⁴

$$\begin{aligned} \int_0^1 \left[\int_0^1 (2 - x - y) dx \right] dy &= \int_0^1 \left[\left(2x - \frac{x^2}{2} - xy \right) \Big|_0^1 \right] dy \\ &= \int_0^1 \left(\frac{3}{2} - y \right) dy \\ &= \left(\frac{3}{2}y - \frac{y^2}{2} \right) \Big|_0^1 = 1 \end{aligned}$$

A expressão $\left(\frac{3}{2}y - \frac{y^2}{2} \right) \Big|_0^1$ significa que a expressão entre parênteses deve ser avaliada com o limite superior de 1 e o limite inferior de 0; o último valor é subtraído pelo primeiro para obter o valor da integral. No exemplo anterior, os limites são $\left(\frac{3}{2} - \frac{1}{2} \right)$ em $y = 1$ e 0 perfazendo o valor da integral igual a 1.

EXEMPLO 9
(Continuação)

$$f(y) = \int_0^1 (2 - x - y) dx$$

$$\left(2x - xy - \frac{x^2}{2} \right) \Big|_0^1 = \frac{3}{2} - y \quad 0 \leq y \leq 1$$

Para verificarmos se as duas variáveis do Exemplo 8 são estatisticamente independentes, precisamos descobrir se $f(x, y) = f(x)f(y)$. Uma vez que $(2 - x - y) \neq (\frac{3}{2} - x)(\frac{3}{2} - y)$, podemos dizer que as duas variáveis não são estatisticamente independentes.

A.5 As características das distribuições de probabilidade

Uma distribuição de probabilidade pode, com frequência, ser resumida em termos de algumas poucas características, conhecidas como **momentos** da distribuição. Dois dos momentos mais amplamente utilizados são a **média**, ou **valor esperado**, e a **variância**.

Valor esperado

O valor esperado de uma variável aleatória discreta X , denotado por $E(X)$, é definido como:

$$E(X) = \sum_x x f(x)$$

em que $\sum_x$ significa a soma que inclui todos os valores de X e $f(x)$ é a função de densidade de probabilidade discreta de X .

EXEMPLO 10

Considere a distribuição da probabilidade da soma de dois números no lançamento dos dois dados apresentados no Exemplo 2. (Veja a Figura A.1.) Multiplicando os vários valores X lá apresentados por suas probabilidades e fazendo a soma geral de todas as observações, obtemos:

$$E(X) = 2\left(\frac{1}{36}\right) + 3\left(\frac{2}{36}\right) + 4\left(\frac{3}{36}\right) + \cdots + 12\left(\frac{1}{36}\right)$$

$$= 7$$

que é o valor médio da soma dos números observados no lançamento dos dois dados.

EXEMPLO 11

Estime $E(X)$ e $E(Y)$ para os dados apresentados no Exemplo 4. Vimos que

x	-2	0	2	3
$f(x)$	0,27	0,12	0,26	0,35

Portanto,

$$E(X) = \sum_x x f(x)$$

$$= (-2)(0,27) + (0)(0,12) + (2)(0,26) + (3)(0,35)$$

$$= 1,03$$

(Continua)

EXEMPLO 11 De forma semelhante,
(Continuação)

$$\begin{aligned} & \begin{array}{ccc} y & 3 & 6 \\ f(y) & 0,51 & 0,49 \end{array} \\ E(Y) &= \sum_y y f(y) \\ &= (3)(0,51) + (6)(0,49) \\ &= 4,47 \end{aligned}$$

O valor esperado de uma variável aleatória contínua é definido como:

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

A única diferença entre esse caso e o valor esperado de uma variável aleatória discreta é que substituímos o símbolo do somatório pelo símbolo da integral.

EXEMPLO 12 Vamos descobrir o valor esperado da função de densidade de probabilidade contínua apresentada no Exemplo 3.

$$\begin{aligned} E(X) &= \int_0^3 x \left(\frac{x^2}{9} \right) dx \\ &= \frac{1}{9} \left[\left(\frac{x^4}{4} \right) \right]_0^3 \\ &= \frac{9}{4} \\ &= 2,25 \end{aligned}$$

Propriedades dos valores esperados

1. O valor esperado de uma constante é a própria constante. Se b é uma constante, $E(b) = b$;
2. Se a e b são constantes,

$$E(aX + b) = aE(X) + b$$

Isso pode ser generalizado. Se $X_1, X_2, \dots, X_N$ são N variáveis aleatórias e $a_1, a_2, \dots, a_N$ e b são constantes, então

$$E(a_1X_1 + a_2X_2 + \dots + a_NX_N + b) = a_1E(X_1) + a_2E(X_2) + \dots + a_NE(X_N) + b$$

3. Se X e Y são variáveis aleatórias independentes, então

$$E(XY) = E(X)E(Y)$$

Ou seja, a expectativa do produto XY é o produto das expectativas (individuais) de X e Y . Entretanto, observe que

$$E\left(\frac{X}{Y}\right) \neq \frac{E(X)}{E(Y)}$$

mesmo se X e Y forem independentes;

4. Se X é uma variável aleatória com função de densidade de probabilidade $f(x)$ e se $g(x)$ é qualquer função de X , então

$$\begin{aligned} E[g(X)] &= \sum_x g(x)f(x) && \text{se } X \text{ for discreta} \\ &= \int_{-\infty}^{\infty} g(x)f(x) dx && \text{se } X \text{ for contínua} \end{aligned}$$

Assim, se $g(X) = X^2$,

$$\begin{aligned} E(X^2) &= \sum_x x^2 f(x) && \text{se } X \text{ for discreta} \\ &= \int_{-\infty}^{\infty} x^2 f(x) dx && \text{se } X \text{ for contínua} \end{aligned}$$

EXEMPLO 13

Considere a seguinte função de densidade de probabilidade (FDP):

x	-2	1	2
$f(x)$	$\frac{5}{8}$	$\frac{1}{8}$	$\frac{2}{8}$

Então,

$$\begin{aligned} E(X) &= -2\left(\frac{5}{8}\right) + 1\left(\frac{1}{8}\right) + 2\left(\frac{2}{8}\right) \\ &= -\frac{5}{8} \end{aligned}$$

e

$$\begin{aligned} E(X^2) &= 4\left(\frac{5}{8}\right) + 1\left(\frac{1}{8}\right) + 4\left(\frac{2}{8}\right) \\ &= \frac{29}{8} \end{aligned}$$

Variância

Seja X uma variável aleatória e seja $E(X) = \mu$. A distribuição, ou dispersão, dos valores de X em torno do valor esperado pode ser mensurada pela variância, definida como

$$\text{var}(X) = \sigma_X^2 = E(X - \mu)^2$$

A raiz quadrada positiva de σ_X^2 , σ_X é definida como **desvio padrão** de X . A variância, ou desvio padrão, indica quão próximos ou distantes os valores individuais de X estão distribuídos em torno de seu valor médio.

A variância definida previamente é calculada como se segue:

$$\begin{aligned} \text{var}(X) &= \sum_x (X - \mu)^2 f(x) && \text{se } X \text{ for uma variável aleatória discreta} \\ &= \int_{-\infty}^{\infty} (X - \mu)^2 f(x) dx && \text{se } X \text{ for uma variável aleatória contínua} \end{aligned}$$

Para conveniência de cálculo, a fórmula da variância apresentada pode ser expressa como

$$\begin{aligned} \text{var}(X) &= \sigma_X^2 = E(X - \mu)^2 \\ &= E(X^2) - \mu^2 \\ &= E(X^2) - [E(X)]^2 \end{aligned}$$

Aplicando essa fórmula, podemos verificar que a variância da variável aleatória apresentada no Exemplo 13 é $\frac{29}{8} - \left(-\frac{5}{8}\right)^2 = \frac{207}{64} = 3,23$.

EXEMPLO 14

Vamos descobrir a variância da variável aleatória apresentada no Exemplo 3.

$$\text{var}(X) = E(X^2) - [E(X)]^2$$

Agora

$$\begin{aligned} E(X^2) &= \int_0^3 x^2 \left(\frac{x^2}{9}\right) dx \\ &= \int_0^3 \frac{x^4}{9} dx \\ &= \frac{1}{9} \left[\frac{x^5}{5}\right]_0^3 \\ &= 243/45 \\ &= 27/5 \end{aligned}$$

Uma vez que $E(X) = \frac{9}{4}$ (veja o Exemplo 12), finalmente temos

$$\begin{aligned} \text{var}(X) &= 243/45 - \left(\frac{9}{4}\right)^2 \\ &= 243/720 = 0,34 \end{aligned}$$

Propriedades da variância

1. $E(X - \mu)^2 = E(X^2) - \mu^2$, como observado anteriormente.
2. A variância de uma constante é zero.
3. Se a e b são constantes, então

$$\text{var}(aX + b) = a^2 \text{var}(X)$$

4. Se X e Y são variáveis aleatórias independentes, então

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$$

$$\text{var}(X - Y) = \text{var}(X) + \text{var}(Y)$$

Isso pode ser generalizado para mais do que duas variáveis independentes.

5. Se X e Y são variáveis aleatórias independentes, e a e b são constantes,

$$\text{var}(aX + bY) = a^2 \text{var}(X) + b^2 \text{var}(Y)$$

Covariância

Seja X e Y duas variáveis aleatórias com médias μ_x e μ_y , respectivamente. Então, a **covariância** entre as duas variáveis é definida como:

$$\text{cov}(X, Y) = E\{(X - \mu_x)(Y - \mu_y)\} = E(XY) - \mu_x \mu_y$$

É prontamente verificado que a variância de uma variável é a covariância daquela variável com ela mesma.

A covariância é calculada como se segue:

$$\begin{aligned}\operatorname{cov}(X, Y) &= \sum_y \sum_x (X - \mu_x)(Y - \mu_y)f(x, y) \\ &= \sum_y \sum_x XYf(x, y) - \mu_x \mu_y\end{aligned}$$

se X e Y são variáveis aleatórias discretas, e

$$\begin{aligned}\operatorname{cov}(X, Y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (X - \mu_x)(Y - \mu_y)f(x, y) dx dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} XYf(x, y) dx dy - \mu_x \mu_y\end{aligned}$$

se X e Y são variáveis aleatórias contínuas.

Propriedades da covariância

1. Se X e Y são independentes, a sua covariância é zero, pois

$$\begin{aligned}\operatorname{cov}(X, Y) &= E(XY) - \mu_x \mu_y \\ &= \mu_x \mu_y - \mu_x \mu_y \quad \text{uma vez que } E(XY) = E(X)E(Y) = \mu_x \mu_y \\ &= 0 \quad \text{quando } X \text{ e } Y \text{ são independentes}\end{aligned}$$

- 2.

$$\operatorname{cov}(a + bX, c + dY) = bd \operatorname{cov}(X, Y)$$

em que a, b, c e d são constantes.

EXEMPLO 15

Vamos descobrir a covariância entre as variáveis aleatórias discretas X e Y cuja função de densidade de probabilidade conjunta é como demonstrado no Exemplo 4. Com base no Exemplo 11, já sabemos que $\mu_x = E(X) = 1,03$ e que $\mu_y = E(Y) = 4,47$.

$$\begin{aligned}E(XY) &= \sum_y \sum_x XYf(x, y) \\ &= (-2)(3)(0,27) + (0)(3)(0,08) + (2)(3)(0,16) + (3)(3)(0) \\ &\quad + (-2)(6)(0) + (0)(6)(0,04) + (2)(6)(0,10) + (3)(6)(0,35) \\ &= 6,84\end{aligned}$$

Portanto,

$$\begin{aligned}\operatorname{cov}(X, Y) &= E(XY) - \mu_x \mu_y \\ &= 6,84 - (1,03)(4,47) \\ &= 2,24\end{aligned}$$

Coefficiente de correlação

O coeficiente de correlação (população) ρ (rho) é definido como:

$$\rho = \frac{\operatorname{cov}(X, Y)}{\sqrt{\{\operatorname{var}(X) \operatorname{var}(Y)\}}} = \frac{\operatorname{cov}(X, Y)}{\sigma_x \sigma_y}$$

Assim definido, ρ é uma medida de associação *linear* entre duas variáveis e situa-se entre -1 e $+1$, -1 indicando associação negativa perfeita e indicando associação positiva perfeita.

Por meio da fórmula anterior, pode-se verificar que

$$\text{cov}(X, Y) = \rho\sigma_x\sigma_y$$

EXEMPLO 16

Vamos estimar o coeficiente da correlação para os dados do Exemplo 4. Com base nas funções de densidade de probabilidade apresentadas no Exemplo 11, pode-se facilmente demonstrar que $\sigma_x = 2,05$ e $\sigma_y = 1,50$. Já mostramos que $\text{cov}(X, Y) = 2,24$. Portanto, aplicando a fórmula anterior, estimamos que ρ é $2,24/(2,05)(1,50) = 0,73$.

Variâncias de variáveis correlacionadas

Sejam X e Y as duas variáveis aleatórias. Então,

$$\begin{aligned}\text{var}(X + Y) &= \text{var}(X) + \text{var}(Y) + 2\text{cov}(X, Y) \\ &= \text{var}(X) + \text{var}(Y) + 2\rho\sigma_x\sigma_y \\ \text{var}(X - Y) &= \text{var}(X) + \text{var}(Y) - 2\text{cov}(X, Y) \\ &= \text{var}(X) + \text{var}(Y) - 2\rho\sigma_x\sigma_y\end{aligned}$$

Entretanto, se X e Y forem independentes, a $\text{cov}(X, Y)$ é zero, neste caso a $\text{var}(X + Y)$ e a $\text{var}(X - Y)$ são ambas iguais a $\text{var}(X) + \text{var}(Y)$, como anteriormente observado.

Os resultados anteriores podem ser generalizados como se segue: $\sum_{i=1}^n X_i = X_1 + X_2 + \cdots + X_n$, então a variância da combinação linear $\sum X_i$ é

$$\begin{aligned}\text{var}\left(\sum_{i=1}^n x_i\right) &= \sum_{i=1}^n \text{var} X_i + 2\sum_{i<j} \text{cov}(X_i, X_j) \\ &= \sum_{i=1}^n \text{var} X_i + 2\sum_{i<j} \rho_{ij}\sigma_i\sigma_j\end{aligned}$$

em que ρ_{ij} é o coeficiente de correlação entre X_i e X_j e σ_i e σ_j são os desvios padrão de X_i e X_j . Assim,

$$\begin{aligned}\text{var}(X_1 + X_2 + X_3) &= \text{var} X_1 + \text{var} X_2 + \text{var} X_3 + 2\text{cov}(X_1, X_2) \\ &\quad + 2\text{cov}(X_1, X_3) + 2\text{cov}(X_2, X_3) \\ &= \text{var} X_1 + \text{var} X_2 + \text{var} X_3 + 2\rho_{12}\sigma_1\sigma_2 \\ &\quad + 2\rho_{13}\sigma_1\sigma_3 + 2\rho_{23}\sigma_2\sigma_3\end{aligned}$$

em que σ_1 , σ_2 e σ_3 são, respectivamente, os desvios padrão de X_1 , X_2 e X_3 e ρ_{12} é o coeficiente de correlação entre X_1 e X_2 , ρ_{13} que entre X_1 e X_3 e ρ_{23} que entre X_2 e X_3 .

Expectativa condicional e variância condicional

Seja $f(x, y)$ a FDP conjunta das variáveis aleatórias X e Y . A expectativa condicional de X , dado $Y = y$, é definida como

$$\begin{aligned}E(X | Y = y) &= \sum_x x f(x | Y = y) && \text{se } X \text{ for discreta} \\ &= \int_{-\infty}^{\infty} x f(x | Y = y) dx && \text{se } X \text{ for contínua}\end{aligned}$$

em que $E(X = Y = y)$ representa a expectativa condicional de X dado $Y = y$ em que $f(x | Y = y)$ é a FDP condicional de X . A expectativa condicional de Y , $E(Y | X = x)$, é definida de forma semelhante.

Expectativa condicional

Observe que $E(X | Y)$ é uma variável aleatória, porque ela é uma função da variável condicionante Y . Contudo, $E(X | Y = y)$, em que y é um valor específico de Y , é uma constante.

Variância condicional

A variância condicional de X dado $Y = y$ é definida como:

$$\begin{aligned}\text{var}(X | Y = y) &= E\{[X - E(X | Y = y)]^2 | Y = y\} \\ &= \sum_x [X - E(X | Y = y)]^2 f(x | Y = y) \quad \text{se } X \text{ for discreta} \\ &= \int_{-\infty}^{\infty} [X - E(X | Y = y)]^2 f(x | Y = y) dx \quad \text{se } X \text{ for contínua}\end{aligned}$$

EXEMPLO 17

Calcule $E(Y | X = 2)$ e a $\text{var}(Y | X = 2)$ para os dados do Exemplo 4.

$$\begin{aligned}E(Y | X = 2) &= \sum_y y f(Y = y | X = 2) \\ &= 3f(Y = 3 | X = 2) + 6f(Y = 6 | X = 2) \\ &= 3(0,16/0,26) + 6(0,10/0,26) \\ &= 4,15\end{aligned}$$

Nota: $f(Y = 3 | X = 2) = f(Y = 3, X = 2)/f(X = 2) = 0,16/0,26$ e $f(Y = 6 | X = 2) = f(Y = 6, X = 2)/f(X = 2) = 0,10/0,26$, então

$$\begin{aligned}\text{var}(Y | X = 2) &= \sum_y [Y - E(Y | X = 2)]^2 f(Y | X = 2) \\ &= (3 - 4,15)^2(0,16/0,26) + (6 - 4,15)^2(0,10/0,26) \\ &= 2,13\end{aligned}$$

Propriedades da expectativa condicional e da variância condicional

1. Se $f(X)$ for uma função de X , então $E(f(X) | X) = f(X)$, isto é, a função de X comporta-se como uma constante no cálculo de sua expectativa condicional sobre X . Assim, $[E(X^3 | X)] = E(X^3)$; se X for conhecido, X^3 também será.
2. Se $f(X)$ e $g(X)$ são funções de X , então

$$E[f(X)Y + g(X) | X] = f(X)E(Y | X) + g(X)$$

Por exemplo, $E[XY + cX^2 | X] = XE(Y | X) + cX^2$, em que c é uma constante.

3. Se X e Y forem independentes, $E(Y | X) = E(Y)$. Ou seja, se X e Y são variáveis aleatórias independentes, a expectativa condicional de Y , dado X , é a mesma que a expectativa incondicional de Y .

4. **Lei das expectativas iteradas.** É interessante notar a seguinte relação entre a expectativa incondicional de uma variável aleatória Y , $E(Y)$, e sua expectativa condicional baseada em outra variável aleatória X , $E(Y | X)$:

$$E(Y) = E_X[E(Y | X)]$$

Essa relação é conhecida como lei das expectativas iteradas, que, neste contexto, estabelece que a expectativa marginal, ou incondicional, de Y é igual à expectativa de sua expectativa condicional, na qual o símbolo E_X denota que a expectativa está cobrindo os valores de X . Simplificando, essa lei estabelece que, se, primeiramente, obtemos $E(Y | X)$ como uma função de X e tomamos seu valor esperado para a distribuição de valores X , terminamos obtendo $E(Y)$, a expectativa incondicional de Y . O leitor pode verificar isso, utilizando os dados fornecidos no Exemplo 4.

Uma implicação da lei de expectativas iteradas é que, se a média condicional de Y dado X ($E(Y | X)$) for zero, a média (incondicional) de Y também será zero. Isso acontece, porque, neste caso,

$$E[E(Y|X)] = E[0] = 0$$

5. Se X e Y são independentes, $\text{var}(Y | X) = \text{var}(Y)$;
6. $\text{var}(Y) = E[\text{var}(Y | X)] + \text{var}[E(Y | X)]$; isto é, a variância (incondicional) de Y é igual à expectativa da variância condicional de Y mais a variância da expectativa condicional de Y .

Momentos de ordem superior das distribuições de probabilidade

Embora a média, a variância e a covariância sejam as medidas-resumo mais frequentemente utilizadas nas FDP univariadas e multivariadas, por vezes precisamos considerar os momentos de ordem superior das FDP, como os momentos de terceira e de quarta ordem. Os momentos de terceira e quarta ordem de uma FDP univariada $f(x)$ em torno de seu valor médio (μ) são definidos como

$$\text{Terceiro momento: } E(X - \mu)^3$$

$$\text{Quarto momento: } E(X - \mu)^4$$

Em geral, o momento de ordem r em torno da média é definido como

$$\text{Momento de ordem } r: E(X - \mu)^r$$

O terceiro e quarto momentos de uma distribuição são normalmente utilizados no estudo da “forma” de uma probabilidade, em particular, da sua **assimetria**, S (falta de simetria) e **curtose**, K (elevação ou achatamento), como apresentado na Figura A.3.

Uma medida de **assimetria** é definida como:

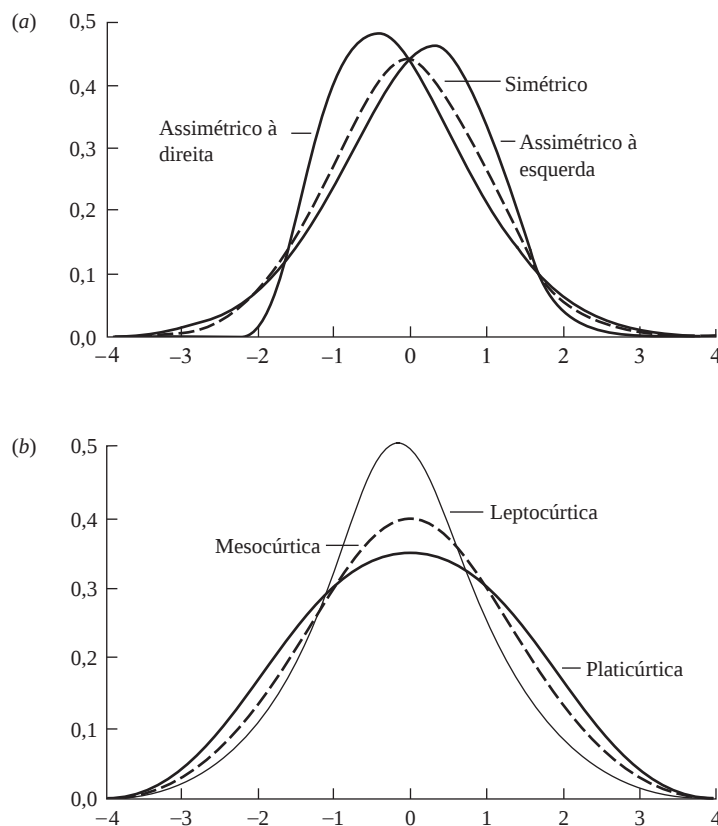
$$\begin{aligned} S &= \frac{E(X - \mu)^3}{\sigma^3} \\ &= \frac{\text{terceiro momento em torno da média}}{\text{cubo do desvio padrão}} \end{aligned}$$

Uma medida comumente utilizada de curtose é dada por:

$$\begin{aligned} K &= \frac{E(X - \mu)^4}{[E(X - \mu)^2]^2} \\ &= \frac{\text{quarto momento em torno da média}}{\text{quadrado do segundo momento}} \end{aligned}$$

FIGURA A.3

(a) Assimetria;
(b) Curtose.



As FDP com valor de K menores de 3 são chamadas **platicúrticas** (gordas ou de caudas curtas) e aquelas com valores maiores de 3 são chamadas **leptocúrticas** (magras ou de caudas longas). Veja a Figura A.3. Uma FDP com um valor curtose de 3 é conhecida como **mesocúrtica**, e desta a distribuição normal é o principal exemplo. (Veja a discussão da distribuição normal na Seção A.6.)

Mostraremos, de forma sucinta, como as medidas de assimetria e curtose podem ser combinadas para determinar se uma variável aleatória segue uma distribuição normal. Lembremos que o procedimento de teste da hipótese, como nos testes t e F , é baseado na hipótese (ao menos para as amostras pequenas e finitas) de que a distribuição subjacente da variável (ou estatística da amostra) é normal. É, portanto, muito importante descobrir nas aplicações concretas se essa hipótese é cumprida.

A.6 Algumas distribuições de probabilidade teóricas importantes

No livro, é feito amplo uso das seguintes distribuições de probabilidade.

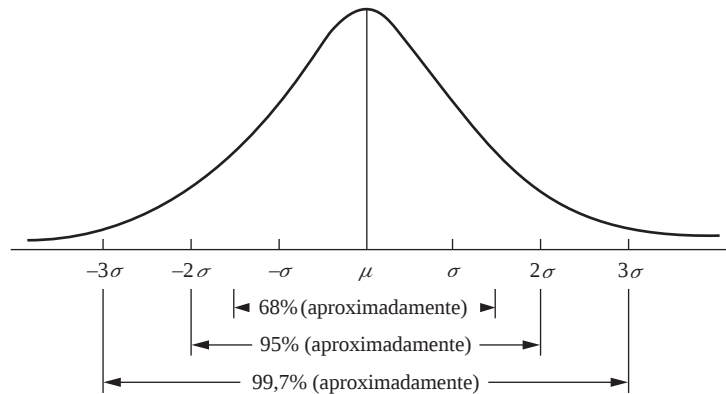
Distribuição normal

A mais conhecida de todas as distribuições de probabilidade teóricas é a distribuição normal, cuja figura em forma de sino é familiar a qualquer um com conhecimento estatístico mínimo.

Uma variável aleatória (contínua) X é considerada normalmente distribuída se a sua FDP tem a seguinte forma:

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

FIGURA A.4
Áreas sob a curva normal.



em que μ e σ^2 , conhecidas como *parâmetros da distribuição*, são, respectivamente, a média e a variância da distribuição. As propriedades dessa distribuição são as seguintes:

1. Ela é simétrica em torno do seu valor médio.
2. Aproximadamente 68% da área sob a curva normal situa-se entre os valores de $\mu \pm \sigma$, cerca de 95% da área situa-se entre $\mu \pm 2\sigma$, e cerca de 99,7% situa-se entre $\mu \pm 3\sigma$, como mostra a Figura A.4.
3. A distribuição normal depende de dois parâmetros μ e σ^2 ; como estes são especificados, pode-se encontrar a probabilidade de que X se situará dentro de um certo intervalo ao utilizar a FDP da distribuição normal. Mas essa tarefa pode ser facilitada consideravelmente ao consultarmos Tabela D.1 do **Apêndice D**. Para utilizarmos a tabela, convertamos a conhecida variável X de distribuição normal com a média μ e σ^2 em uma **variável normal padronizada** Z pela seguinte transformação:

$$Z = \frac{x - \mu}{\sigma}$$

Uma importante propriedade de qualquer variável padronizada é que o seu valor médio é zero e sua variância é a unidade. Assim, Z possui média zero e variância 1. Substituindo z na função FDP dada anteriormente, obtemos:

$$f(Z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}Z^2\right)$$

que é a FDP da variável normal padronizada. As probabilidades apresentadas no **Apêndice D**, Tabela D.1, são baseadas na variável normal padronizada.

Por convenção, denotamos uma variável distribuída de forma normal como:

$$X \sim N(\mu, \sigma^2)$$

em que $\sim$ significa “distribuído como”, N indica distribuição normal e as quantidades entre parênteses são os dois parâmetros da distribuição normal, ou seja, a média e a variância. Seguindo essa convenção,

$$X \sim N(0, 1)$$

significa que X é uma variável de distribuição normal com média zero e variância 1. Em outras palavras, ela é a variável normal padronizada Z .

EXEMPLO 18

Suponha que $X \sim N(8, 4)$. Qual a probabilidade de que X assumirá um valor entre $X_1 = 4$ e $X_2 = 12$? Para calcularmos a probabilidade requerida, estimamos os valores de Z como:

$$Z_1 = \frac{X_1 - \mu}{\sigma} = \frac{4 - 8}{2} = -2$$

$$Z_2 = \frac{X_2 - \mu}{\sigma} = \frac{12 - 8}{2} = +2$$

Agora, com base na Tabela D.1, observamos que $\Pr(0 \leq Z \leq 2) = 0,4772$. Então, por simetria, temos $\Pr(-2 \leq Z \leq 0) = 0,4772$. Por conseguinte, a probabilidade requerida é $0,4772 + 0,4772 = 0,9544$. (Veja a Figura A.4.)

EXEMPLO 19

Qual a probabilidade de, no exemplo anterior, X exceder 12? A probabilidade de que X exceda 12 é a mesma de que Z exceda 2. com base na Tabela D.1, é óbvio que essa probabilidade é $(0,5 - 0,4772)$ ou $0,0228$.

4. Sejam $X_1 \sim N(\mu_1, \sigma_1^2)$ e $X_2 \sim N(\mu_2, \sigma_2^2)$ e suponha que elas sejam independentes. Considere, agora, a combinação linear

$$Y = aX_1 + bX_2$$

em que a e b são constantes. Então, pode ser demonstrado que:

$$Y \sim N[(a\mu_1 + b\mu_2), (a^2\sigma_1^2 + b^2\sigma_2^2)]$$

Esse resultado, que afirma que uma *combinação linear de variáveis de distribuição normal é distribuída normalmente*, pode ser facilmente generalizado para uma combinação linear de mais de duas variáveis de distribuição normal.

5. **Teorema central do limite.** Considere que $X_1, X_2, \dots, X_n$ denotem n variáveis aleatórias independentes, todas elas possuem a mesma FDP com média $= \mu$ e variância $= \sigma^2$. Seja $\bar{X} = \sum X_i / n$ (a média amostral). À medida que n aumenta indefinidamente (i.e., $n \rightarrow \infty$)

$$\bar{X} \underset{n \rightarrow \infty}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

Isto é, $\bar{X}$ aproxima-se da distribuição normal com média μ e variância σ^2/n . Repare que esse resultado é verdadeiro não importando a forma da FDP. Como resultado, temos:

$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0, 1)$$

Ou seja, Z é uma variável normal padronizada.

6. O terceiro e quarto momento da distribuição normal em torno do valor médio são como se segue:

$$\text{Terceiro momento: } E(X - \mu)^3 = 0$$

$$\text{Quarto momento: } E(X - \mu)^4 = 3\sigma^4$$

Nota: todos os momentos de ordem ímpar em torno do valor médio de uma variável normalmente distribuída são zero.

7. Como resultado, e seguindo as medidas de assimetria e curtose discutidas anteriormente, para uma FDP normal, a simetria é $= 0$ e a curtose é $= 3$; uma distribuição normal é simétrica e mesocúrtica. Portanto, um teste simples de normalidade é descobrir se os valores calculados de assimetria e curtose afastam-se das normas de 0 e 3. Esta é, de fato, a lógica subjacente ao **teste de normalidade Jarque-Bera (JB)** discutido no livro:

$$JB = n \left[\frac{S^2}{6} + \frac{(K - 3)^2}{24} \right] \quad (5.12.1)$$

em que S representa a assimetria e K , a curtose. Sob a hipótese nula da normalidade, JB é distribuído como uma estatística **qui-quadrado** com 2 graus de liberdade.

8. A média e a variância de uma variável aleatória com distribuição normal são independentes no sentido de que uma não é função da outra.
9. Se X e Y são de distribuição conjunta normal, elas são independentes se, e apenas se, a covariância entre elas $[\text{cov}(X, Y)]$ é zero. (Veja o Exercício 4.1.)

A distribuição χ^2 (qui-quadrado)

Sejam $Z_1, Z_2, \dots, Z_k$ variáveis normais padronizadas independentes (variáveis normais com média zero e variância 1). Então a quantidade

$$Z = \sum_{i=1}^k Z_i^2$$

possui a distribuição χ^2 com k graus de de liberdade (gl), em que o termo gl significa o número de quantidades independentes na soma anterior. Uma variável com distribuição qui-quadrado é representada por χ_k^2 , em que o subscrito k indica o gl. Geometricamente, a distribuição qui-quadrada aparece na Figura A.5.

As propriedades da distribuição χ^2 são as seguintes:

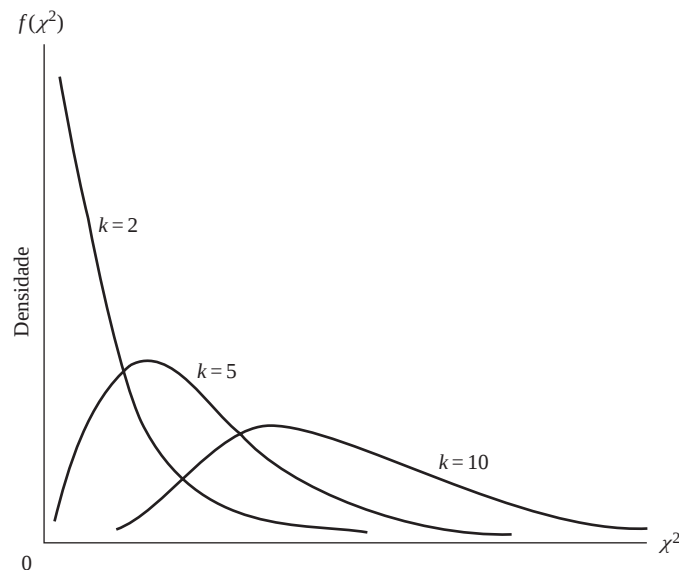
1. Como demonstra a Figura A.5, a distribuição χ^2 é uma distribuição assimétrica, o grau de assimetria dependendo do gl. Para um gl relativamente pequeno, a distribuição é altamente assimétrica para a direita; mas, à medida que o gl aumenta, a distribuição torna-se progressivamente simétrica. Na verdade, para o gl superior a 100, a variável

$$\sqrt{2\chi^2} - \sqrt{2k - 1}$$

pode ser tratada como uma variável normal padronizada, em que k é o gl.

FIGURA A.5

Função da densidade da variável χ^2 .



2. A média de distribuição qui-quadrado é k e sua variância é $2k$, em que k é o gl.
3. Se Z_1 e Z_2 são duas variáveis qui-quadrados independentes com gl k_1 e k_2 , então a soma $Z_1 + Z_2$ é também uma variável qui-quadrado com $gl = k_1 + k_2$.

EXEMPLO 20

Qual a probabilidade de obter um χ^2 com valor de 40 ou maior, dado o gl de 20? Como mostra a Tabela D.4, a probabilidade de obter um χ^2 com valor de 39,9968 ou maior (20 gl) é de 0,005. Portanto, a probabilidade de obter um χ^2 com valor de 40 ou maior é menor do que 0,005, uma probabilidade bem pequena.

Distribuição t de Student

Se Z_1 é uma variável normal padrão [$Z_1 \sim N(0, 1)$] e outra variável Z_2 segue a distribuição qui-quadrada com k graus de liberdade e é distribuída independentemente de Z_1 , a variável definida como

$$t = \frac{Z_1}{\sqrt{(Z_2/k)}} \\ = \frac{Z_1 \sqrt{k}}{\sqrt{Z_2}}$$

segue a distribuição t de Student com k graus de liberdade. Uma variável com distribuição t é frequentemente designada como t_k , em que o subscrito k denota os graus de liberdade. Geometricamente, a distribuição t é apresentada na Figura A.6.

As propriedades da distribuição t de Student são as seguintes:

1. Como a Figura A.6 demonstra, a distribuição t , assim como a distribuição normal, é simétrica, porém ela é mais achatada do que a distribuição normal. Contudo, à medida que aumentam os graus de liberdade, a distribuição t aproxima-se da distribuição normal.
2. A média da distribuição t é zero e sua variância é $k/(k - 2)$.

A distribuição t está tabulada na Tabela D.2.

EXEMPLO 21

Dado que os graus de liberdade são iguais a 13, qual a probabilidade de obter um valor t (a) de cerca de 3 ou maior, (b) de aproximadamente -3 ou menor, e (c) com valor $|t|$ ou cerca de 3 ou maior, em que $|t|$ significa o valor absoluto de t (não levando em conta o sinal $+$ ou $-$)?

Com base na Tabela D.2, as respostas são: (a) cerca de 0,005, (b) cerca de 0,005 devido à simetria da distribuição, e (c) cerca de $0,01 = 2(0,005)$.

FIGURA A.6

Distribuição de t de Student para graus de liberdade selecionados.

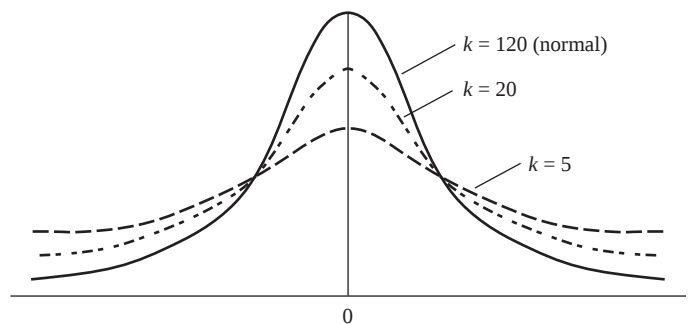
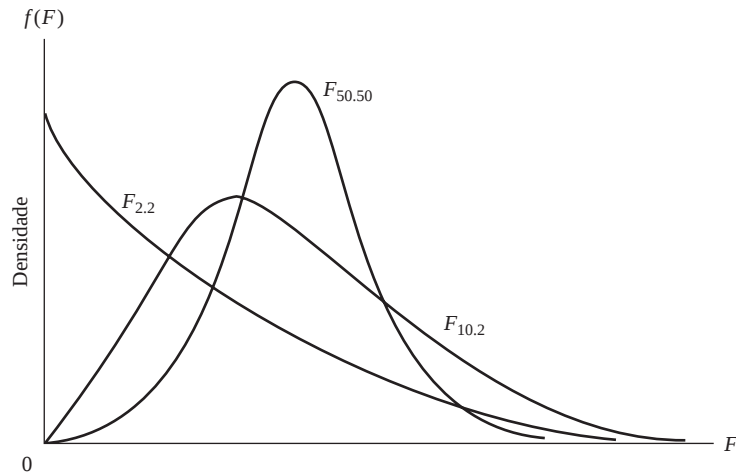


FIGURA A.7

Distribuição F para vários graus de liberdade.



A distribuição F

Se Z_1 e Z_2 são variáveis que possuem uma distribuição qui-quadrado independente com graus de liberdade k_1 e k_2 , respectivamente, a variável

$$F = \frac{Z_1/k_1}{Z_2/k_2}$$

segue a distribuição F (de Fisher) com graus de liberdade k_1 e k_2 . Uma variável com distribuição F é representada por F_{k_1, k_2} em que os subscritos indicam os graus de liberdade associados à duas variáveis Z , k_1 sendo denominado *grau de liberdade do numerador* e k_2 , *grau de liberdade do denominador*. Geometricamente, a distribuição F é demonstrada na Figura A.7

A distribuição F conta com as seguintes propriedades:

1. Como a distribuição qui-quadrado, a distribuição F tem viés para a direita. Porém, pode-se demonstrar que, à medida que k_1 e k_2 tornam-se maiores, a distribuição F aproxima-se da distribuição normal.
2. O valor médio de uma variável com distribuição F é $k_2/(k_2 - 2)$, que é definido por $k_2 > 2$, e sua variância é

$$\frac{2k_2^2(k_1 + k_2 - 2)}{k_1(k_2 - 2)^2(k_2 - 4)}$$

que é definida por $k_2 > 4$.

3. O quadrado de uma variável aleatória com distribuição t com k graus de liberdade possui uma distribuição F com 1 e k graus de liberdade. Simbolicamente,

$$t_k^2 = F_{1, k}$$

EXEMPLO 22

Dado $k_1 = 10$ e $k_2 = 8$, qual a probabilidade de obter um valor $F(a)$ de 3,4 ou maior e (b) de 5,8 ou maior? Como demonstra a Tabela D.3, essas probabilidades são (a) aproximadamente 0,05 e (b) aproximadamente 0,01.

4. Se o grau de liberdade do denominador, k_2 , é muito elevado, a seguinte relação ocorre entre as distribuições F e qui-quadrado:

$$k_1 F \sim \chi_{k_1}^2$$

Para um grau de liberdade do denominador bastante alto, o grau de liberdade do numerador multiplicado pelo valor F é aproximadamente o mesmo de um valor qui-quadrado com grau de liberdade do numerador.

EXEMPLO 23

Sejam $k_1 = 20$ e $k_2 = 120$. O valor F crítico de 5% para esses graus de liberdade é 1,48. Por conseguinte, o F de $k_1 = (20)(1,48) = 29,6$. Com base na distribuição qui-quadrado para 20 graus de liberdade, o valor qui-quadrado crítico de 5% é cerca de 31,41.

Por sinal, perceba que, como, para um grau de liberdade do denominador mais elevado, a distribuição t , a distribuição qui-quadrado e a distribuição F aproximam-se da distribuição normal, essas três distribuições são conhecidas como as *distribuições relacionadas à distribuição normal*.

Distribuição binomial de Bernoulli

Considera-se que uma variável aleatória X segue a distribuição de Bernoulli, denominada assim em homenagem ao matemático suíço, se a sua função de densidade (ou massa) de probabilidade (FDP) é:

$$P(X = 0) = 1 - p$$

$$P(X = 1) = p$$

em que p , $0 \leq p \leq 1$, é a probabilidade de que algum evento seja um “sucesso”, como a probabilidade de obter cara no lançamento de uma moeda. Para tal variável,

$$E(X) = [1 \times p(X = 1) + 0 \times p(X = 0)] = p$$

$$\text{var}(X) = pq$$

ou seja, $q = (1 - p)$, a probabilidade de um “fracasso”.

Distribuição binomial

A distribuição binomial é a generalização da distribuição de Bernoulli. Denotemos por n o número de tentativas independentes, cada uma delas resulta em um “sucesso” com probabilidade p e um “fracasso” com uma probabilidade $q = (1 - p)$. Se X representa o número do sucesso em n tentativas, então diz-se que X segue a distribuição binomial cuja FDP é:

$$f(X) = \binom{n}{x} p^x (1 - p)^{n-x}$$

em que x representa o número do sucesso em n tentativas e

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

em que $n!$, lido como n fatorial, significa $n(n-1)(n-2) \cdot \dots \cdot 1$.

A binomial é uma distribuição de dois parâmetros, n e p . Para essa distribuição:

$$E(X) = np$$

$$\text{var}(X) = np(1 - p) = npq$$

Por exemplo, se lançarmos uma moeda 100 vezes e quisermos descobrir a probabilidade de obter 60 caras, colocamos na fórmula acima $p = 0,5$, $n = 100$ e $x = 60$. Existem rotinas de cálculos para avaliação de tais probabilidades.

Podemos verificar como a distribuição binomial é uma generalização da distribuição de Bernoulli.

A distribuição de Poisson

Considera-se que uma variável aleatória X tem uma distribuição de Poisson se a sua FDP é:

$$f(X) = \frac{e^{-\lambda} \lambda^x}{x!} \quad \text{é } x = 0, 1, 2, \dots, \lambda > 0$$

A distribuição de Poisson depende de um parâmetro único, λ . Uma característica distintiva da distribuição de Poisson é que a sua variância é igual a seu valor esperado, que é λ . Isto é,

$$E(X) = \text{var}(X) = \lambda$$

O modelo de Poisson, como vimos no capítulo sobre os modelos de regressão não linear, é utilizado para modelar fenômenos raros ou infrequentes, como o número de chamadas telefônicas recebidas em um intervalo de 5 minutos, ou o número de multas por excesso de velocidade recebidas em um intervalo de uma hora, ou ainda os números de patentes recebidas por uma empresa em um ano.

A.7 Inferência estatística: estimação

Na Seção A.6, consideramos várias distribuições de probabilidade teóricas. Muito frequentemente, sabemos ou estamos propensos a admitir que uma variável aleatória X segue uma distribuição de probabilidade particular, mas não sabemos o(s) valor(es) do(s) parâmetro(s) da distribuição. Por exemplo, se X segue a distribuição normal, podemos querer saber o valor de seus dois parâmetros: a média e a variância. Para estimarmos as incógnitas, o procedimento habitual é supor que temos uma **amostra aleatória** de tamanho n com base na distribuição da probabilidade conhecida e utilizar os dados da amostra para estimar os parâmetros desconhecidos.⁵ Isso é chamado de **problema da estimação**. Nesta seção, examinaremos mais de perto esse problema. Ele pode ser dividido em duas categorias: estimação pontual e estimação intervalar.

Estimação pontual

Para melhor entendermos, seja X uma variável aleatória com FDP de $f(x; \theta)$, em que θ é o parâmetro da distribuição (para simplificar a discussão, supomos que há apenas um parâmetro desconhecido; nossa discussão pode ser facilmente generalizada). Suponha que conhecemos a forma funcional — conhecemos a FDP teórica, tal como a distribuição t —, mas não conhecemos o valor de θ . Portanto, sorteamos uma amostra aleatória de tamanho n a partir dessa FDP conhecida e desenvolvemos uma função dos valores da amostra, de modo que

$$\hat{\theta} = f(x_1, x_2, \dots, x_n)$$

forneça-nos uma estimativa do verdadeiro θ . $\hat{\theta}$ é conhecido como uma **estatística**, ou um **estimador**, e um valor numérico particular tomado pelo estimador é conhecido como **estimativa**. Perceba que $\hat{\theta}$ pode ser tratada como uma variável aleatória porque é uma função dos dados amostrais. $\hat{\theta}$ nos fornece uma regra, ou fórmula, que nos conta como estimamos o verdadeiro θ . Assim, se admitimos que

$$\hat{\theta} = \frac{1}{n}(x_1 + x_2 + \dots + x_n) = \bar{X}$$

em que $\bar{X}$ é a média da amostra, então $\bar{X}$ é um estimador do verdadeiro valor da média, por exemplo, μ . Se, em um caso específico, $\bar{X} = 50$, isso fornece uma *estimativa de* μ . O estimador $\hat{\theta}$ obtido previamente é conhecido como **estimador pontual**, por fornecer apenas uma estimativa única (pontual) de θ .

⁵ Sejam $X_1, X_2, \dots, X_n$ n variáveis aleatórias com FDP conjunta $f(x_1, x_2, \dots, x_n)$. Se podemos escrever

$$f(x_1, x_2, \dots, x_n) = f(x_1)f(x_2)\cdots f(x_n)$$

em que $f(x)$ é a FDP comum de cada X , então, diz-se que $x_1, x_2, \dots, x_n$ constituem uma amostra aleatória de tamanho n com base em uma população com FDP $f(x_n)$.

Estimação intervalar

Em vez de obtermos apenas uma estimativa única de θ , suponha que obtenhamos duas estimativas de θ ao construirmos dois estimadores $\hat{\theta}_1(x_1, x_2, \dots, x_n)$ e $\hat{\theta}_2(x_1, x_2, \dots, x_n)$, e, com alguma confiança (probabilidade), que o intervalo entre $\hat{\theta}_1$ e $\hat{\theta}_2$ inclui o verdadeiro θ . Na estimação intervalar, em contraste com a estimação pontual, fornecemos uma amplitude de valores possíveis dentro dos quais o verdadeiro θ pode estar.

O conceito principal por trás da estimação intervalar é a noção de **amostra**, ou **probabilidade de distribuição, de um estimador**. Por exemplo, pode-se demonstrar que, se uma variável X possui distribuição normal, a média da amostra $\bar{X}$ também possui distribuição normal com média $= \mu$ (a média verdadeira) e variância $= \sigma^2/n$, em que n é o tamanho da amostra. Em outras palavras, a distribuição amostral, ou probabilidade, do estimador $\bar{X}$ é $\bar{X} \sim N(\mu, \sigma^2/n)$. Como resultado, se construirmos o intervalo

$$\bar{X} \pm 2 \frac{\sigma}{\sqrt{n}}$$

e dissermos que a probabilidade é de aproximadamente 0,95, ou 95%, intervalos como esse incluirão o verdadeiro μ , estamos, de fato, construindo um estimador de intervalo para μ . Perceba que o intervalo fornecido anteriormente é aleatório, uma vez que é baseado em $\bar{X}$, que variará de amostra para amostra.

De forma mais geral, na estimação intervalar, construímos dois estimadores $\hat{\theta}_1$ e $\hat{\theta}_2$, ambos funções dos valores amostrais de X , de maneira que:

$$\Pr(\hat{\theta}_1 \leq \theta \leq \hat{\theta}_2) = 1 - \alpha \quad 0 < \alpha < 1$$

ou seja, podemos afirmar que é de $1 - \alpha$ a probabilidade de que o intervalo de $\hat{\theta}_1$ a $\hat{\theta}_2$ contenha o verdadeiro θ . Este é conhecido como intervalo de confiança de tamanho $1 - \alpha$ para θ , $1 - \alpha$ sendo conhecido como **coeficiente de confiança**. Se $\alpha = 0,05$, então $1 - \alpha = 0,95$, significando que, se construirmos um intervalo de confiança com um coeficiente de confiança de 0,95, então nas construções repetidas resultantes de amostras repetidas deveremos estar certos em 95 de 100 casos, se afirmarmos que o intervalo contém o verdadeiro θ . Quando o coeficiente de confiança é 0,95, frequentemente dizemos que temos um intervalo de confiança de 95%. Em geral, se o coeficiente de confiança é de $1 - \alpha$, dizemos que temos um intervalo de confiança de $100(1 - \alpha)\%$. Perceba que α é conhecido como o **nível de significância** ou a probabilidade de cometer um erro de Tipo I. Esse tópico é discutido na Seção A.8.

EXEMPLO 24

Suponha que a distribuição da altura dos homens de uma população possua distribuição normal com média $= \mu$ polegadas e $\sigma = 2,5$ polegadas. Uma amostra de 100 homens tirada de forma aleatória dessa população tem uma média de altura de 67 polegadas. Estabeleça um intervalo de confiança de 95% para a média de altura ($= \mu$) da população como um todo.

Como foi notado, $\bar{X} \sim N(\mu, \sigma^2/n)$, que, nesse caso, torna-se $\bar{X} \sim N(\mu, 2,5^2/100)$. Pela Tabela D.1, pode-se verificar que

$$\bar{X} - 1,96 \left(\frac{\sigma}{\sqrt{n}} \right) \leq \mu \leq \bar{X} + 1,96 \frac{\sigma}{\sqrt{n}}$$

cobre 95% da área sob a curva normal. Portanto, o intervalo anterior fornece um intervalo de confiança de 95% para μ . Inserindo os valores fornecidos de $\bar{X}$, σ e n , obtemos o intervalo de confiança de 95% como

$$66,51 \leq \mu \leq 67,49$$

(Continua)

EXEMPLO 24
(Continuação)

Em mensurações repetidas como essa, os intervalos assim estabelecidos incluirão o verdadeiro μ com 95% de confiança. Um comentário técnico pode ser feito aqui. Embora possamos dizer que a probabilidade de que o intervalo aleatório $[\bar{X} \pm 1,96(\sigma/\sqrt{n})]$ inclui μ seja de 95%, *não podemos* dizer que seja de 95% a probabilidade de que o intervalo particular (66,51, 67,49) inclua μ . Como esse intervalo é fixado, a probabilidade de que ele incluirá μ é 0 ou 1. O que podemos afirmar é que, se construirmos 100 desses intervalos, 95 dos 100 intervalos incluirão μ ; não podemos garantir que um intervalo em particular incluirá necessariamente μ .

Métodos de estimação

De maneira geral, há três métodos de estimação de parâmetros: (1) mínimos quadrados (MQ), (2) máxima verossimilhança (MV) e (3) método dos momentos (MM) e sua extensão, o método dos momentos generalizado (MMG). Temos dedicado tempo considerável para ilustrar o método dos mínimos quadrados. No Capítulo 4, introduzimos o método da máxima verossimilhança no contexto da regressão, mas esse método possui uma aplicação muito mais ampla.

A ideia-chave por trás do método da verossimilhança é a **função de verossimilhança**. Para ilustrar, suponha que a variável aleatória X possui FDP $f(X, \theta)$ que depende de um parâmetro único θ . Conhecemos a FDP (por exemplo, de Bernoulli ou binomial), mas não conhecemos o valor do parâmetro. Suponha que obtenhamos uma amostra aleatória de nX valores. A FDP conjunta desses n valores é:

$$g(x_1, x_2, \dots, x_n; \theta)$$

Por ela ser uma amostra aleatória, podemos escrever a FDP conjunta anterior como um produto das FDPs individuais:

$$g(x_1, x_2, \dots, x_n; \theta) = f(x_1; \theta)f(x_2; \theta) \cdots f(x_n; \theta)$$

A FDP conjunta possui uma interpretação dual. Se θ é conhecido, interpretamos como uma FDP conjunta de se observar os dados de valores amostrais. Por outro lado, podemos tratá-la como uma função de θ para valores de $x_1, x_2, \dots, x_n$. Na segunda interpretação, chamamos a FDP conjunta de **função de verossimilhança** e escrevemos como:

$$L(\theta; x_1, x_2, \dots, x_n) = f(x_1; \theta)f(x_2; \theta) \cdots f(x_n; \theta)$$

Observe a inversão do papel de θ na função de densidade de probabilidade conjunta e na função de verossimilhança.

O estimador de máxima verossimilhança de θ é aquele valor de θ que maximiza a função de verossimilhança (da amostra), L . Por uma conveniência matemática, em geral tomamos o logaritmo da verossimilhança, chamado **função log de verossimilhança (log L)**. Seguindo as regras de cálculo da maximização, diferenciamos a função log de verossimilhança com respeito à incógnita e igualamos a derivada resultante a zero. O valor resultante do estimador é chamado **estimador de máxima verossimilhança**. Pode-se aplicar a condição de maximização de segunda ordem para assegurar que o valor que obtivemos é, de fato, o valor máximo.

No caso de haver mais de um parâmetro desconhecido, diferenciamos a função log de verossimilhança com respeito a cada incógnita, igualamos as expressões resultantes a zero e solucionamos simultaneamente para obter os valores dos parâmetros desconhecidos. Já demonstramos isso com relação ao modelo de regressão múltipla (veja o Capítulo 4, Apêndice 4A1.).

EXEMPLO 25

Suponha que a variável aleatória X siga a distribuição de Poisson com o valor médio de λ . Suponha que $x_1, x_2, \dots, x_n$ sejam variáveis aleatórias de Poisson independentes, cada uma com média λ . Suponha que queiramos descobrir o estimador de máxima verossimilhança de λ . A função de verossimilhança aqui é:

(Continua)

EXEMPLO 25
(Continuação)

$$L(x_1, x_2, \dots, x_n; \lambda) = \frac{e^{-\lambda} \lambda^{x_1}}{x_1!} \frac{e^{-\lambda} \lambda^{x_2}}{x_2!} \dots \frac{e^{-\lambda} \lambda^{x_n}}{x_n!}$$

$$= \frac{e^{-n\lambda} \lambda^{\sum x_i}}{x_1! x_2! \dots x_n!}$$

Essa é uma expressão razoavelmente difícil de manejar, mas, se tomarmos o seu log, ela se torna

$$\log(x_1, x_2, \dots, x_n; \lambda) = -n\lambda + \sum x_i \log \lambda - \log c$$

em que $\log c = \prod x_i!$. Diferenciando a expressão anterior com respeito a λ , obtemos $(-n + (\sum x_i)/\lambda)$. Igualando essa última expressão a zero, obtemos $\lambda_{ml} = (\sum x_i)/n = \bar{X}$, que é o estimador de máxima verossimilhança da incógnita λ .

Método dos momentos

Apresentamos uma noção do método dos momentos no Exercício 3.4 no chamado **princípio da analogia**, no qual os momentos da amostra tentam duplicar as propriedades de suas contrapartes na população. O método dos momentos generalizado (MMG), que é uma generalização do MM, agora está tornando-se mais popular, porém, não em um nível introdutório. Desse modo, por ora, não trataremos dele.

As propriedades estatísticas desejáveis agrupam-se em duas categorias: propriedades das amostras pequenas, ou amostras finitas, e propriedades das amostras grandes, ou assintóticas. Por trás desses conjuntos de propriedades está a noção de que um estimador possui uma distribuição em amostra, ou de probabilidade.

Propriedades de pequenas amostras**Sem viés**

Um estimador $\hat{\theta}$ é chamado de estimado não tendencioso (não viesado) e de θ se o valor esperado de $\hat{\theta}$ for igual ao verdadeiro θ ; isto é,

$$E(\hat{\theta}) = \theta$$

ou

$$E(\hat{\theta}) - \theta = 0$$

Se essa igualdade não se sustenta, o estimador é conhecido como viesado, e o viés é calculado como:

$$\text{viés}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

É claro, se $E(\hat{\theta}) = \theta$ — isto é, $\hat{\theta}$ é um estimador não viesado — o viés é zero.

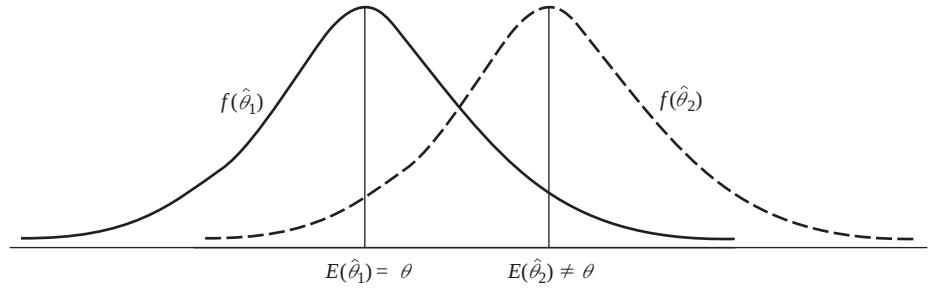
Geometricamente, a situação é representada na Figura A.8. Observe que a não tendenciosidade é uma propriedade das amostras repetidas, não de qualquer amostra: mantendo o tamanho da amostra fixo, extraímos várias amostras, obtendo, cada vez, uma estimativa do parâmetro desconhecido. Espera-se que o valor médio dessas estimativas seja igual ao valor verdadeiro se o estimador não possuir viés.

Variância mínima

Diz-se que $\hat{\theta}$ é um estimador de mínima variância de θ se a variância de $\hat{\theta}_1$ for menor, ou pelo menos igual, à variância de $\hat{\theta}_2$, que é qualquer outro estimador de θ . Geometricamente, temos

FIGURA A.8

Estimadores viesados e não viesados.



a Figura A.9, que mostra os três estimadores de θ , ou seja, $\hat{\theta}_1$, $\hat{\theta}_2$, e $\hat{\theta}_3$, e suas distribuições de probabilidade. Como demonstrado, a variância de $\hat{\theta}_3$ é menor que as de $\hat{\theta}_1$ e $\hat{\theta}_2$. Então, admitindo apenas os três estimadores possíveis, neste caso, $\hat{\theta}_3$ é um estimador de variância mínima. Porém, perceba que $\hat{\theta}_3$ é um estimador tendencioso (por quê?).

Melhor estimador não viesado ou estimador eficiente

Se $\hat{\theta}_1$ e $\hat{\theta}_2$ são dois estimadores não viesados de θ , e a variância de $\hat{\theta}_1$ é menor, ou no máximo, igual à variância de $\hat{\theta}_2$, $\hat{\theta}_1$ é um **estimador não viesado de variância mínima**, ou **melhor não viesado**, ou **eficiente**. Na Figura A.9, dos dois estimadores não viesados, $\hat{\theta}_1$ e $\hat{\theta}_2$, $\hat{\theta}_1$ é o melhor não viesado, ou eficiente.

Linearidade

Um estimador $\hat{\theta}$ é conhecido como um estimador linear de θ se ele é uma função linear das observações da amostra. A média da amostra definida como

$$\bar{X} = \frac{1}{n} \sum X_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

é um estimador linear, porque é uma função linear dos valores de X .

Melhor estimador linear não viesado ou estimador eficiente

Se $\hat{\theta}$ é linear, é não viesado, e possui uma variância mínima no grupo de todos os estimadores lineares não viesados de θ , ele é chamado de **melhor estimador linear não viesado**, ou, para resumir, **BLUE**.

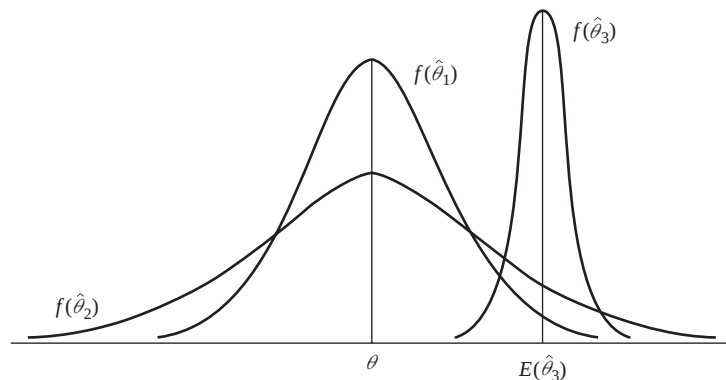
Estimador com erro quadrado médio mínimo (MSE)

O MSE de um estimador $\hat{\theta}$ é definido como

$$\text{MSE}(\hat{\theta}) = E(\hat{\theta} - \theta)^2$$

FIGURA A.9

Distribuição de três estimadores de θ .



Isso contrasta com a variância de $\hat{\theta}$, que se define como:

$$\text{var}(\hat{\theta}) = E[\hat{\theta} - E(\hat{\theta})]^2$$

A diferença entre as duas é que a $\text{var}(\hat{\theta})$ mensura a dispersão da distribuição de $\hat{\theta}$ em torno da sua média, ou valor esperado, enquanto o $\text{MSE}(\hat{\theta})$ mensura a dispersão em torno do valor verdadeiro do parâmetro. A relação entre as duas é como se segue:

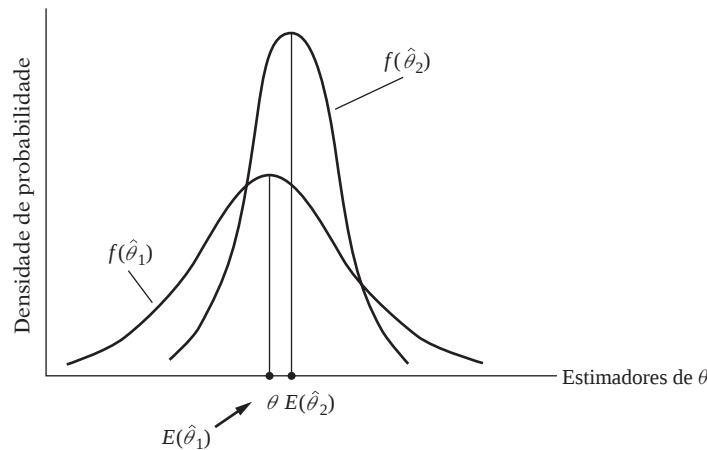
$$\begin{aligned} \text{MSE}(\hat{\theta}) &= E(\hat{\theta} - \theta)^2 \\ &= E[\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta]^2 \\ &= E[\hat{\theta} - E(\hat{\theta})]^2 + E[E(\hat{\theta}) - \theta]^2 + 2E[\hat{\theta} - E(\hat{\theta})][E(\hat{\theta}) - \theta] \\ &= E[\hat{\theta} - E(\hat{\theta})]^2 + E[E(\hat{\theta}) - \theta]^2 \quad \text{uma vez que o último termo é zero}^6 \\ &= \text{var}(\hat{\theta}) + \text{viés}(\hat{\theta})^2 \\ &= \text{variância de } \hat{\theta} \text{ mais o quadrado do viés} \end{aligned}$$

Naturalmente, se o viés é zero, $\text{MSE}(\hat{\theta}) = \text{var}(\hat{\theta})$. O critério MSE mínimo consiste em escolher um estimador cujo MSE seja o menor em um conjunto de estimadores concorrentes. Observe que, mesmo se tal estimador for encontrado, há um *trade-off* envolvido — para obter uma variância mínima, podemos ter de aceitar algum viés. Geometricamente, a situação é apresentada na Figura A.10. Nesta figura, $\hat{\theta}_2$ é levemente viesado, mas sua variância é menor do que a do estimador não viesado $\hat{\theta}_1$. Na prática, contudo, o critério MSE mínimo é utilizado quando o critério do melhor não viesado é incapaz de produzir estimadores com variâncias menores.

Propriedades de grandes amostras

Em geral, acontece de um estimador não satisfazer uma ou mais das propriedades estatísticas desejáveis em amostras pequenas. Contudo, à medida que o tamanho da amostra cresce indefinidamente, o estimador possui várias propriedades estatísticas desejáveis. Essas propriedades são conhecidas como **propriedades de amostras grandes**, ou **assintóticas**.

FIGURA A.10



Ausência assintótica de viés.

Um estimador $\hat{\theta}$ é considerado um estimador assintoticamente não viesado de θ se

$$\lim_{n \rightarrow \infty} E(\hat{\theta}_n) = \theta$$

⁶O último termo pode ser escrito como $2\{[E(\hat{\theta})]^2 - [E(\hat{\theta})]^2 - \theta E(\hat{\theta}) + \theta E(\hat{\theta})\} = 0$. Observe também que $E[E(\hat{\theta}) - \theta]^2 = [E(\hat{\theta}) - \theta]^2$, posto que o valor esperado de uma constante é simplesmente a própria constante.

em que $\hat{\theta}_n$ significa que o estimador é baseado no tamanho da amostra de n , $\lim$ significa limite e $n \rightarrow \infty$ indica que n cresce indefinidamente. Em outras palavras, $\hat{\theta}$ é um estimador assintoticamente não viesado de θ se o seu valor esperado, ou média, aproxima-se do valor verdadeiro à medida que o tamanho da amostra torna-se cada vez maior. Como exemplo, considere a seguinte mensuração da variância amostral de uma variável aleatória X :

$$S^2 = \frac{\sum (X_i - \bar{X})^2}{n}$$

Pode-se demonstrar que

$$E(S^2) = \sigma^2 \left(1 - \frac{1}{n}\right)$$

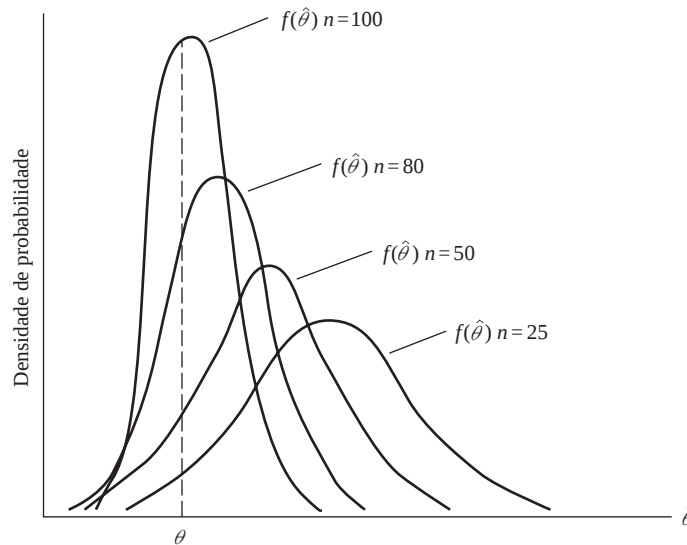
em que σ^2 é a verdadeira variância. É óbvio que, em uma amostra pequena, S^2 é viesado, mas à medida que n cresce indefinidamente, $E(S^2)$ aproxima-se do verdadeiro σ^2 ; portanto, é assintoticamente não viesado.

Consistência

Diz-se que $\hat{\theta}$ é um estimador consistente se ele se aproxima do valor verdadeiro θ à medida que o tamanho da amostra torna-se cada vez maior. A Figura A.11 ilustra a propriedade. Na figura, temos a distribuição de $\hat{\theta}$ baseada no tamanho das amostras de 25, 50, 80 e 100. Como mostra a figura, $\hat{\theta}$ baseado em $n = 25$ é viesado, posto que sua distribuição amostral não é centrada no verdadeiro θ . Porém, à medida que n cresce, a distribuição de $\hat{\theta}$ não apenas tende a ser mais proximamente fechada em θ ($\hat{\theta}$ torna-se menos viesado), mas sua variância também torna-se menor. Se, no limite (quando n cresce indefinidamente), a distribuição de $\hat{\theta}$ convergir para um único ponto θ , isto é, se a distribuição de $\hat{\theta}$ tiver dispersão, ou variância, zero, dizemos que $\hat{\theta}$ é um **estimador consistente** de θ .

FIGURA A.11

A distribuição de θ à medida que a amostra cresce.



Diz-se, mais formalmente, que $\hat{\theta}$ é um estimador consistente de θ se a probabilidade de que o valor absoluto da diferença entre $\hat{\theta}$ e θ seja menor do que δ (uma quantidade positiva arbitrariamente pequena) aproxima-se de 1 quando n tende ao infinito. Simbolicamente,

$$\lim_{n \rightarrow \infty} P\{|\hat{\theta} - \theta| < \delta\} = 1 \quad \delta > 0$$

em que P significa probabilidade. Isso é frequentemente expresso como

$$\text{plim}_{n \rightarrow \infty} \hat{\theta} = \theta$$

em que *plim* indica o limite em probabilidade.

Perceba que as propriedades de não tendenciosidade de consistência são conceitualmente muito diferentes. A propriedade de não tendenciosidade pode compreender qualquer tamanho de amostra, enquanto a consistência é estritamente uma propriedade das amostras grandes.

Um *condição suficiente* para a consistência é que tanto o viés quanto a variância tendam a zero à medida que o tamanho da amostra cresce indefinidamente.⁷ Por outro lado, uma condição suficiente para a consistência é que o quadrado médio mínimo $\text{MSE}(\hat{\theta})$ tende a zero à medida que n cresce indefinidamente. (Para $\text{MSE}[\hat{\theta}]$, veja a discussão anteriormente apresentada.)

EXEMPLO 26

Seja $X_1, X_2, \dots, X_n$ uma amostra aleatória com base em uma distribuição com média μ e variância σ^2 . Demonstre que a média $\bar{X}$ da amostra é um estimador consistente de μ .

Por meio de estatísticas elementares, sabe-se que $E(\bar{X}) = \mu$ e $\text{var}(\bar{X}) = \sigma^2/n$. Uma vez que $E(\bar{X}) = \mu$ independentemente do tamanho da amostra, ele é não viesado. Além disso, à medida que n cresce indefinidamente, a $\text{var}(\bar{X})$ tende a zero. Por isso, $\bar{X}$ é um estimador consistente de μ .

As seguintes regras sobre a probabilidade são dignas de nota.

1. *Invariância (propriedade de Slutsky)*. Se $\hat{\theta}$ for um estimador consistente de θ , e se $h(\hat{\theta})$ for qualquer função de $\hat{\theta}$, então

$$\text{plim}_{n \rightarrow \infty} h(\hat{\theta}) = h(\theta)$$

O que isso significa é que, se $\hat{\theta}$ for um estimador consistente de θ , $1/\hat{\theta}$ será também um estimador consistente de $1/\theta$ e que $\log(\hat{\theta})$ será também um estimador consistente de $\log(\theta)$. Perceba que essa propriedade não é válida para o operador de expectativa E ; isto é, se $\hat{\theta}$ for um estimador não viesado de θ (isto é, $E[\hat{\theta}] = \theta$), *não será verdade* que $1/\hat{\theta}$ é um estimador não viesado de $1/\theta$; isto é, $E(1/\hat{\theta}) \neq 1/E(\hat{\theta}) = 1/\theta$.

2. Se b é uma constante,

$$\text{plim}_{n \rightarrow \infty} b = b$$

Ou seja, o limite em probabilidade de uma constante é a mesma constante.

3. Se $\hat{\theta}_1$ e $\hat{\theta}_2$ forem estimadores consistentes,

$$\text{plim}(\hat{\theta}_1 + \hat{\theta}_2) = \text{plim} \hat{\theta}_1 + \text{plim} \hat{\theta}_2$$

$$\text{plim}(\hat{\theta}_1 \hat{\theta}_2) = \text{plim} \hat{\theta}_1 \text{plim} \hat{\theta}_2$$

$$\text{plim}\left(\frac{\hat{\theta}_1}{\hat{\theta}_2}\right) = \frac{\text{plim} \hat{\theta}_1}{\text{plim} \hat{\theta}_2}$$

As duas últimas propriedades, em geral, não são válidas para o operador de expectativa E . Assim, $E(\hat{\theta}_1/\hat{\theta}_2) \neq E(\hat{\theta}_1)/E(\hat{\theta}_2)$. De maneira semelhante, $E(\hat{\theta}_1 \hat{\theta}_2) \neq E(\hat{\theta}_1)E(\hat{\theta}_2)$. Se, entretanto, $\hat{\theta}_1$ e $\hat{\theta}_2$ forem distribuídos independentemente, $E(\hat{\theta}_1 \hat{\theta}_2) = E(\hat{\theta}_1)E(\hat{\theta}_2)$, como observado anteriormente.

⁷Mais tecnicamente, $\lim_{n \rightarrow \infty} E(\hat{\theta}_n) = \theta$ e $\lim_{n \rightarrow \infty} \text{var}(\hat{\theta}_n) = 0$.

Eficiência assintótica

Seja $\hat{\theta}$ um estimador de θ . A variância da distribuição assintótica de $\hat{\theta}$ é chamada de **variância assintótica** de $\hat{\theta}$. Se $\hat{\theta}$ for consistente e a sua variância assintótica for menor do que a variância assintótica de todos os estimadores consistentes de θ , $\hat{\theta}$ é chamado de **assintoticamente eficiente**.

Normalidade assintótica

Um estimador $\hat{\theta}$ é considerado ter distribuição assintoticamente normal se sua distribuição amostral tende a aproximar-se da distribuição normal à medida que o tamanho da amostra n cresce indefinidamente. Por exemplo, a teoria estatística demonstra que, se $X_1, X_2, \dots, X_n$ são variáveis independentes com distribuição normal e possuem a mesma média μ e a mesma variância σ^2 , a média da amostra $\bar{X}$ também possui distribuição normal com média μ e variância σ^2/n em amostras pequenas e também em amostras grandes. Contudo, se os X_i forem independentes com média μ e variância σ^2 , mas não necessariamente pertencerem à distribuição normal, a média da amostra $\bar{X}$ possuirá distribuição assintoticamente normal com média μ e variância σ^2/n ; ou seja, à medida que o tamanho da amostra n cresce indefinidamente, a média da amostra tende a ser normalmente distribuída com média μ e variância σ^2/n . Na verdade, esse é o teorema central do limite previamente discutido.

A.8 Inferência estatística: testando as hipóteses

A estimação e o teste da hipótese constituem os ramos gêmeos da inferência estatística clássica. Ao examinarmos o problema da estimação, examinaremos brevemente o problema do teste estatístico de hipóteses.

O problema do teste de hipótese pode ser estabelecido da seguinte forma: admita que tenhamos uma variável aleatória X com uma FDP conhecida $f(x; \theta)$, em que θ é o parâmetro da distribuição. Ao obtermos uma amostra aleatória de tamanho n , obtemos o estimador pontual $\hat{\theta}$. Uma vez que o verdadeiro θ é raramente conhecido, levantamos a questão: o estimador $\hat{\theta}$ é “compatível” com algum valor hipotético de θ , por exemplo, $\theta = \theta^*$, em que θ^* é um valor numérico específico de θ ? Em outras palavras, poderia a nossa amostra ser proveniente da FDP $f(x; \theta) = \theta^*$? Na linguagem de teste da hipótese, $\theta = \theta^*$ é chamado **hipótese nula** (ou sustentada) e é geralmente denotada por H_0 . A hipótese nula é testada contra uma **hipótese alternativa**, denotada por H_1 , que, por exemplo, pode estabelecer que $\theta \neq \theta^*$. (Nota: em alguns livros, H_0 e H_1 são designados por H_1 e H_2 , respectivamente.)

A hipótese nula e a hipótese alternativa podem ser **simples** ou **compostas**. Uma hipótese é denominada *simples* se especifica o(s) valor(es) do(s) parâmetro(s) de distribuição; do contrário, é chamada de hipótese *composta*. Assim, se $X \sim N(\mu, \sigma^2)$ e afirmamos que

$$H_0: \mu = 15 \quad \text{e} \quad \sigma = 2$$

é uma hipótese simples, enquanto

$$H_0: \mu = 15 \quad \text{e} \quad \sigma > 2$$

é uma hipótese composta porque aqui o valor de σ não é especificado.

Para testarmos a hipótese nula (por exemplo, para testar sua validade), utilizamos a informação da amostra para obter o que é conhecido como **estatística de teste**. Muito frequentemente, essa estatística de teste torna-se o estimador pontual do parâmetro desconhecido. Então, tentamos descobrir a *distribuição da amostra ou da probabilidade da estatística de teste e utilizamos a abordagem do intervalo de confiança ou o teste de significância para testar a hipótese nula*. O mecanismo é ilustrado a seguir.

Para melhor entendermos, vamos voltar ao Exemplo 24, que diz respeito à altura (X) dos homens em uma população. Dizemos que

$$X_i \sim N(\theta, \sigma^2) = N(\mu, 2,5^2)$$

$$\bar{X} = 67 \quad n = 100$$

Vamos admitir que

$$H_0: \mu = \mu^* = 69$$

$$H_1: \mu \neq 69$$

A questão é: poderia a amostra com $\bar{X} = 67$, a estatística de teste, ter sido extraída da população com o valor médio de 69? Intuitivamente, não podemos rejeitar a hipótese nula se $\bar{X}$ é “suficientemente próximo” de μ^* ; ou então podemos rejeitá-la em favor da hipótese alternativa. Como decidimos que $\bar{X}$ é “suficientemente próximo” de μ^* ? Podemos adotar duas abordagens, (1) intervalo de confiança e (2) teste de significância, ambas levando a conclusões idênticas em qualquer aplicação específica.

A abordagem do intervalo de confiança

Posto que $X_i \sim N(\theta, \sigma^2)$, sabemos que a estatística de teste $\bar{X}$ é distribuída como

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

Uma vez que conhecemos a distribuição de probabilidade de $\bar{X}$, por que não estabelecer, por exemplo, um intervalo de confiança de $100(1 - \alpha)$ para μ baseado em $\bar{X}$ e verificar se esse intervalo de confiança inclui $\mu = \mu^*$? Se incluir, não poderemos rejeitar a hipótese nula; se não incluir, poderemos rejeitar a hipótese nula. Assim, se $\alpha = 0,05$, teremos um intervalo de confiança de 95%, e, se este intervalo de confiança incluir, μ^* , não poderemos rejeitar a hipótese nula — 95 dentre 100 intervalos assim estabelecidos deverão provavelmente incluir μ^* .

O procedimento é como se segue: uma vez que $\bar{X} \sim N(\mu, \sigma^2/n)$, segue-se que

$$Z_i = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

ou seja, uma variável normal padrão. Por meio da tabela de distribuição normal, sabemos que

$$\Pr(-1,96 \leq Z_i \leq 1,96) = 0,95$$

Isto é,

$$\Pr\left(-1,96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1,96\right) = 0,95$$

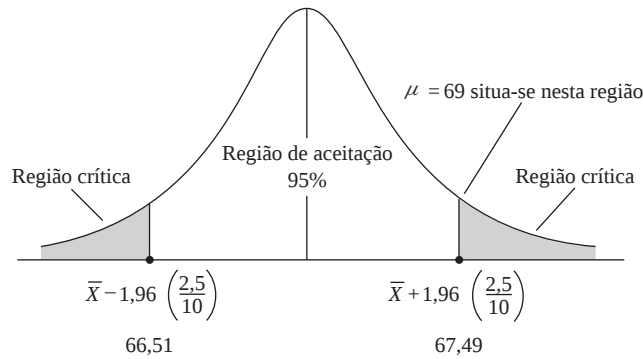
que, rearranjada, resulta em

$$\Pr\left[\bar{X} - 1,96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1,96 \frac{\sigma}{\sqrt{n}}\right] = 0,95$$

Isso é um intervalo de confiança para μ . Uma vez que esse intervalo foi estabelecido, o teste da hipótese nula é simples. Tudo o que temos de fazer é verificar se $\mu = \mu^*$ está nesse intervalo. Se estiver, não poderemos rejeitar a hipótese nula; se não estiver, poderemos rejeitá-la.

FIGURA A.12

Intervalo de confiança de 95% para μ .



Voltando ao Exemplo 24, já estabelecemos um intervalo de confiança de 95% para μ , que é

$$66,51 \leq \mu \leq 67,49$$

Obviamente, esse intervalo não inclui $\mu = 69$. Por conseguinte, podemos rejeitar a hipótese nula de que o verdadeiro μ é 69 com um coeficiente de confiança de 95%. Geometricamente, a situação é como apresentada na Figura A.12.

Na linguagem do teste de hipóteses, o intervalo de confiança que estabelecemos é chamado de **região de aceitação** e a(s) área(s) fora da(s) região(ões) é(são) chamada(s) **região(ões) crítica(s)** ou **região(ões) de rejeição** da hipótese nula. Os limites inferior e superior da região de aceitação (que a separam das regiões de rejeição) são chamados **valores críticos**. Nessa linguagem do teste de hipóteses, se o valor hipotético recair na região de aceitação, não se poderá rejeitar a hipótese nula; caso contrário, pode-se rejeitá-la.

É importante observar que, ao decidir rejeitar ou não a H_0 , pode vir a ocorrer dois tipos de erros: (1) podemos rejeitar H_0 quando ela for, de fato, verdadeira; este é o chamado **erro tipo I** (no exemplo anterior, $\bar{X} = 67$ poderia ser proveniente da população com um valor médio de 69); ou (2) podemos não rejeitar H_0 quando ela for, de fato, falsa; este é chamado de **erro tipo II**. Portanto, um teste de hipótese não estabelece o valor do verdadeiro μ . Ele apenas fornece meios de decidir se podemos agir como se $\mu = \mu^*$.

Erros do tipo I e do tipo II

Esquematicamente, temos

Decisão	Situação	
	H_0 é verdadeira	H_0 é falsa
Rejeitar	Erro do tipo I	Não há erro
Não rejeitar	Não há erro	Erro do tipo II

Idealmente, gostaríamos de minimizar tanto os erros do tipo I quanto os do tipo II. Infelizmente, para qualquer tamanho de amostra, não é possível minimizar ambos os erros simultaneamente. A abordagem clássica a esse problema, incorporada ao trabalho de Neyman e Pearson, é supor que um erro do tipo I seja provavelmente mais sério, na prática, do que um erro do tipo II. Deveríamos manter a probabilidade de cometer um erro do tipo I em um nível bem baixo, como 0,01 ou 0,05, e então tentar minimizar a probabilidade de cometer um erro do tipo II quanto for possível.

Na literatura, a probabilidade de um erro do tipo I é designada como α e é chamada de **nível de significância**, e a probabilidade de um erro do tipo II é designada como β . A probabilidade de não cometer um erro do tipo II é chamada de **potência do teste**. Em outras palavras, a potência de um teste é a sua capacidade de rejeitar uma falsa hipótese nula. A abordagem clássica ao teste de hipótese é fixar α em níveis como 0,01 (ou 1%) ou 0,05 (5%) e tentar maximizar a potência do teste; ou seja, minimizar β .

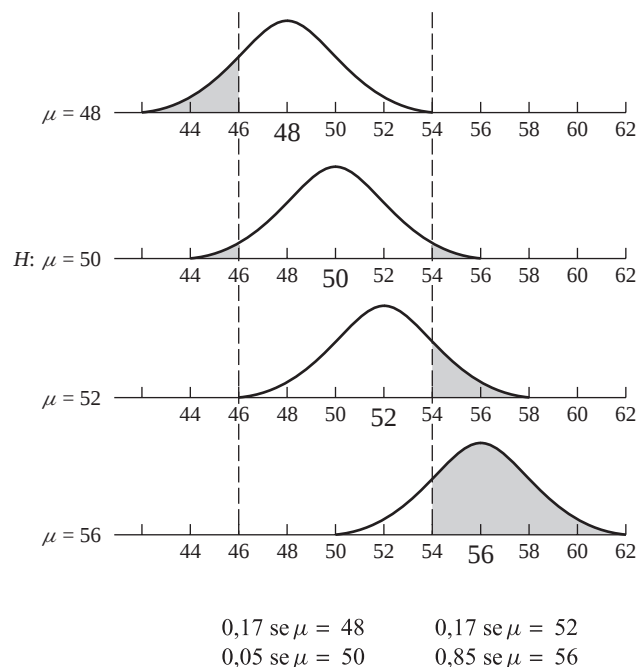
É importante que o leitor compreenda o conceito da potência de um teste, que é mais bem explicado com um exemplo.⁸ Seja $X \sim N(\mu, 100)$, ou seja, X tem distribuição normal com média μ e variância 100. Suponha que $\alpha = 0,05$. Suponha que tenhamos uma amostra de 25 observações, que forneça um valor médio da amostra de $\bar{X}$. Além disso, suponha que consideremos a hipótese $H_0: \mu = 50$. Posto que X é normalmente distribuído, sabemos que a média da amostra é também normalmente distribuída como: $\bar{X} \sim N(\mu, 100/25)$. Daí, estabelecida a hipótese nula de que $\mu = 50$, o intervalo de confiança de 95% para $(\mu \pm 1,96(\sqrt{100/25}) = \mu \pm 3,92$, ou seja, (46,08 a 53,92). Portanto, a região crítica consiste em todos os valores de $\bar{X}$ menores que 46,08 ou maiores que 53,92. Rejeitaremos a hipótese nula de que a média verdadeira é 50 se o valor da média da amostra estiver abaixo de 46,08 ou maior que 53,92.

Porém, qual a probabilidade de que $\bar{X}$ esteja situado na(s) região(ões) crítica(s) anterior(es) se o verdadeiro μ possui um valor diferente de 50? Suponha que haja três hipóteses alternativas: $\mu = 48$, $\mu = 52$ e $\mu = 56$. Se alguma dessas alternativas for verdadeira, ela será a média real da distribuição de $\bar{X}$. O desvio padrão não é modificado para as três alternativas, uma vez que σ^2 ainda se pressupõe como 100.

As áreas sombreadas na Figura A.13 demonstram as possibilidades de que $\bar{X}$ recairá sobre a região crítica se cada uma das hipóteses alternativas for verdadeira. Como se pode verificar, essas possibilidades são 0,17 (para $\mu = 48$), 0,05 (para $\mu = 50$), 0,17 (para $\mu = 52$) e 0,85 (para $\mu = 56$). Como se pode verificar nessa figura, sempre que o verdadeiro valor de μ difere substancialmente da hipótese em consideração (que aqui é $\mu = 50$), a probabilidade de rejeitar a hipótese é alta; porém, quando o verdadeiro valor não é muito diferente do valor dado para a hipótese nula, a probabilidade de rejeição é menor. Intuitivamente, isso deveria fazer sentido se as hipóteses nula e alternativa fossem muito proximamente agrupadas.

FIGURA A.13

Distribuição de X quando $N = 25$, $\sigma = 10$, e $\mu = 48, 50, 52$, ou 56 . Na $H_0: \mu = 50$, a região crítica com $\alpha = 0,05$ é $\bar{X} < 46,1$ e $\bar{X} > 53,9$. A área sombreada indica a probabilidade de que $\bar{X}$ recaia sobre a região crítica. Essa probabilidade é:



Isso pode ser visto mais adiante quando consideramos a Figura A.14, chamada **gráfico da função potência**; a curva demonstra que há a chamada **curva de potência**.

⁸ A próxima discussão e os gráficos são baseados em Walker, Helen M.; Lev, Joseph. *Statistical inference*. Nova York: Holt, Rinehart e Winston, 1953. p. 161–162.

O leitor perceberá que o coeficiente de confiança ($1 - \alpha$) discutido anteriormente é apenas 1 menos a probabilidade de que se cometa um erro do tipo I. Assim, um coeficiente de confiança de 95% significa que estamos preparados para aceitar no máximo uma probabilidade de 5% de cometer um erro do tipo I — não queremos rejeitar a hipótese verdadeira mais do que 5 vezes em 100.

O valor p ou nível exato de significância

Em vez de fazer uma pré-seleção de α em níveis arbitrários, como 1, 5 ou 10%, pode-se obter o **valor p (probabilidade)** ou **nível exato de significância** de uma estatística de teste. O valor p é definido como *o menor nível de significância a que uma hipótese nula pode ser rejeitada*.

Suponhamos que, em uma aplicação envolvendo 20 graus de liberdade, obtenhamos um valor t de 3,552. Agora, o valor p , ou probabilidade exata, de obter um valor t de 3,552 ou superior a isso pode ser verificado na Tabela D.2 como 0,001 (unicaudal) ou 0,002 (bicaudal). Podemos afirmar que o valor t observado de 3,552 é estatisticamente significativo no nível 0,001 ou 0,002, dependendo de utilizarmos um teste unicaudal ou bicaudal.

Agora, vários pacotes estatísticos rotineiramente apresentam o valor p das estatísticas de teste estimadas. Portanto, aconselha-se ao leitor a observar o valor p sempre que possível.

Tamanho da amostra e testes de hipótese

Em dados de pesquisa envolvendo centenas de observações, a hipótese nula parece ser rejeitada com mais frequência do que em amostras pequenas. Vale a pena citar aqui Angus Deaton:

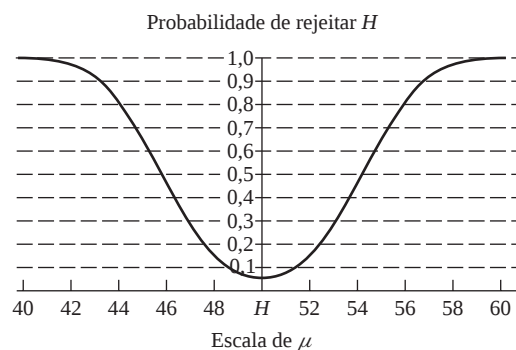
À medida que o tamanho da amostra cresce, e, desde que utilizemos um procedimento de estimação consistente, nossas estimativas estarão próximas da verdade e menos dispersas ao redor dessa verdade para que as discrepâncias que não são detectáveis com o tamanho da amostra pequena levem-nos à rejeição em amostras grandes. Amostras de tamanhos grandes assemelham-se ao grande poder resolutivo de um telescópio; características que não são visíveis a uma certa distância tornam-se mais e mais definitivamente delineadas à medida que acontece a magnificação.⁹

Seguindo Leamer e Schwarz, Deaton sugere ajustar os valores críticos padrão dos testes F e χ^2 como se segue: *rejeitar a hipótese nula quando o valor F calculado exceder o logaritmo do tamanho da amostra, ou seja, $\ln n$, e quando a estatística χ^2 calculada para a restrição q exceder $q \ln n$, em que l é o logaritmo natural e n é o tamanho da amostra*. Esses valores críticos são conhecidos como valores críticos **Leamer-Schwarz**.

Utilizando o exemplo de Deaton, se $n = 100$, a hipótese nula seria rejeitada apenas se o valor F calculado fosse maior do que 4,6, porém, se $n = 10.000$, a hipótese nula seria rejeitada quando o valor F calculado excedesse 9,2.

FIGURA A.14

Função da potência do teste de hipótese $\mu = 50$ quando $N = 25$, $\sigma = 10$, e $\alpha = 0,05$.



⁹ Deaton, Angus. *The analysis of household surveys: a microeconomic approach to development policy*. Baltimore: The Johns Hopkins University Press, 2000. p. 130.

A abordagem do teste de significância

Lembre-se de que

$$Z_i = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Em qualquer aplicação dada, $\bar{X}$ e n são conhecidos (ou podem ser estimados), contudo, os verdadeiros μ e σ não são conhecidos. Porém, se σ for especificado e considerarmos (fazendo uso da H_0) que $\mu = \mu^*$, um valor numérico específico, então Z_i poderá ser diretamente calculado e poderemos facilmente observar a tabela de distribuição normal para encontrar a probabilidade de obter o valor Z calculado. Se essa probabilidade for pequena, por exemplo, menor do que 5% ou 1%, poderemos rejeitar a hipótese nula — se a hipótese fosse verdadeira, as chances de obter o valor particular de Z deveriam ser muito altas. Essa é a ideia geral por trás da abordagem do teste de significância para o teste de hipótese. A ideia central em questão é a estatística de teste (aqui a estatística Z) e sua distribuição de probabilidade sob o valor presumido de $\mu = \mu^*$. Apropriadamente, neste caso, o teste é conhecido como **teste Z** , uma vez que utilizamos o valor Z (normal padronizado).

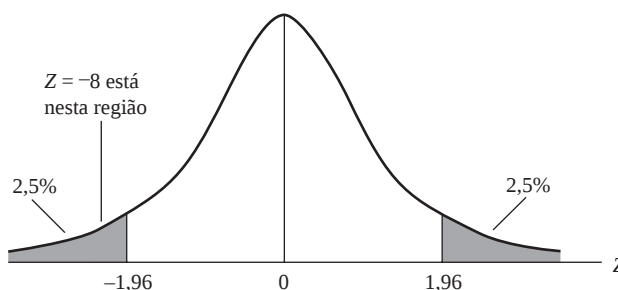
Voltando ao nosso exemplo, se $\mu = \mu^* = 69$, a estatística Z torna-se

$$\begin{aligned} Z &= \frac{\bar{X} - \mu^*}{\sigma/\sqrt{n}} \\ &= \frac{67 - 69}{2,5/\sqrt{100}} \\ &= -2/0,25 = -8 \end{aligned}$$

Se observarmos a Tabela D.1, de distribuição normal, podemos verificar que a probabilidade de obter tal valor de Z é extremamente pequena. (Nota: a probabilidade de um valor Z exceder 3 ou -3 é de aproximadamente 0,001. A probabilidade de Z exceder 8 é ainda menor.) Podemos rejeitar a hipótese nula de que $\mu = 69$; dado esse valor, a nossa chance de obter um $\bar{X}$ de 67 é extremamente pequena. Portanto, duvidamos que a nossa amostra venha da população com um valor médio de 69. Por meio do diagrama, a situação é apresentada na Figura A.15.

FIGURA A.15

A distribuição da estatística Z .



Na linguagem do teste de significância, quando dizemos que uma estatística de teste é significativa, em geral queremos dizer que podemos rejeitar a hipótese nula. Considera-se que a estatística de teste é significativa se a probabilidade de obtê-la for igual ou menor do que α , a probabilidade de cometer um erro do tipo I. Assim, se $\alpha = 0,05$, sabemos que a probabilidade de obter um valor Z de $-1,96$ ou $1,96$ é de 5% (ou de 2,5% em cada cauda da distribuição normal padrão). Em nosso exemplo ilustrativo, Z era -8 . Daí a probabilidade de obter tal valor de Z ser muito menor do que 2,5%, bem abaixo de nossa probabilidade pré-especificada de cometer um erro do tipo I. É por isso que o valor calculado de $Z = -8$ é estatisticamente significativo; rejeitamos a hipótese nula de que o verdadeiro μ^* seja 69. É claro, chegamos à mesma conclusão utilizando a abordagem do intervalo de confiança para o teste de hipótese.

Agora, vamos resumir os passos envolvidos no teste da hipótese estatística:

Passo 1. Formule a hipótese nula H_0 e a hipótese alternativa H_1 (por exemplo: $H_0: \mu = 69$ e $H_1: \mu = 69$).

Passo 2. Selecione a estatística de teste (por exemplo: $\bar{X}$).

Passo 3. Determine a distribuição de probabilidade da estatística de teste (por exemplo: $\bar{X} \sim N(\mu, \sigma^2/n)$).

Passo 4. Escolha o nível de significância α (a probabilidade de cometer um erro do tipo I).

Passo 5. Utilizando a distribuição de probabilidade da estatística de teste, estabeleça um valor de confiança $100(1 - \alpha)\%$. Se o valor do parâmetro submetido à hipótese nula (por exemplo: $\mu = \mu^* = 69$) estiver na região de confiança, a região de aceitação, não rejeite a hipótese nula. Porém, se ele estiver fora desse intervalo (ou seja, dentro da região de rejeição), pode-se rejeitar a hipótese nula. Tenha em mente que, ao não rejeitar ou rejeitar uma hipótese nula, corre-se o risco de estar errado em uma porcentagem de α .

Referências

Para mais detalhes do material tratado neste apêndice, o leitor pode consultar as seguintes referências:

HOEL, Paul G. *Introduction to mathematical statistics*. 4. ed. Nova York John Wiley & Sons, 1974.

Este livro fornece uma introdução bem simples a vários aspectos da estatística matemática.

FREUND, John E.; e WALPOLE, Ronald E. *Mathematical statistics*. 3. ed. Englewood Cliffs, NJ.: Prentice Hall, 1980. Outro livro introdutório em estatística matemática.

MOOD, Alexander, M.; GRAYBILL, Franklin A.; BOE, Duane C. *Introduction to the theory of statistics*. 3. ed., Nova York: McGraw-Hill, 1974. Esta é uma introdução abrangente da teoria estatística, porém, é, de certa forma, mais difícil do que os dois livros anteriores.

NEWBOLD, Paul. *Statistics for business and economics*. Englewood Cliffs, NJ.: Prentice Hall, 1984. Uma introdução não matemática abrangente à estatística com vários problemas solucionados.

Este apêndice fornece o essencial sobre álgebra matricial para a compreensão do Apêndice C e de parte do conteúdo do Capítulo 18. A discussão não é rigorosa e não são dadas quaisquer demonstrações. Para demonstrações e mais detalhes, o leitor pode consultar as referências.

Matriz

$$\mathbf{A} = [a_{ij}] = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1N} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{M1} & a_{M2} & a_{M3} & \cdots & a_{MN} \end{bmatrix}$$
$$\mathbf{A}_{2 \times 3} = \begin{bmatrix} 2 & 3 & 5 \\ 6 & 1 & 3 \end{bmatrix} \quad \mathbf{B}_{3 \times 3} = \begin{bmatrix} 1 & 5 & 7 \\ -1 & 0 & 4 \\ 8 & 9 & 11 \end{bmatrix}$$

O escalar é um único número (real). Em outros termos, um escalar é uma matriz 1×1 .

Uma matriz constituída de M linhas e apenas uma coluna é chamada **vetor coluna**. Empregando letras minúsculas em negrito para denotar vetores, um exemplo de vetor coluna pode ser

$$\mathbf{x}_{4 \times 1} = \begin{bmatrix} 3 \\ 4 \\ 5 \\ 9 \end{bmatrix}$$

Vetor linha

Uma matriz que consiste em uma única linha e N colunas é denominada **vetor linha**.

$$\mathbf{x}_{1 \times 4} = [1 \quad 2 \quad 5 \quad -4] \quad \mathbf{y}_{1 \times 5} = [0 \quad 5 \quad -9 \quad 6 \quad 10]$$

Transposição

A transposição de uma matriz $\mathbf{A}$ $M \times N$, indicada por $\mathbf{A}'$ (que se lê como “ $\mathbf{A}$ linha” ou “ $\mathbf{A}$ transposta”) é uma matriz $N \times M$ obtida por meio da troca das linhas pelas colunas de $\mathbf{A}$; ou seja, a i -ésima linha de $\mathbf{A}$ torna-se a i -ésima coluna de $\mathbf{A}'$. Por exemplo,

$$\mathbf{A}_{3 \times 2} = \begin{bmatrix} 4 & 5 \\ 3 & 1 \\ 5 & 0 \end{bmatrix} \quad \mathbf{A}'_{2 \times 3} = \begin{bmatrix} 4 & 3 & 5 \\ 5 & 1 & 0 \end{bmatrix}$$

Na medida em que um vetor é um tipo especial de matriz, a transposição de um vetor linha é a transposição de um vetor coluna e a transposição de um vetor coluna é um vetor linha. Portanto,

$$\mathbf{x} = \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} \quad \text{e} \quad \mathbf{x}' = [4 \quad 5 \quad 6]$$

Seguiremos a convenção de indicar os vetores linha com “linha” (‘).

Submatriz

Dada a matriz $\mathbf{A}$ ($M \times N$), se todas as colunas e linhas de $\mathbf{A}$ forem eliminadas, com exceção das r linhas e s colunas, a matriz resultante da ordem $r \times s$ será denominada **submatriz** de $\mathbf{A}$. Sendo assim, se

$$\mathbf{A}_{3 \times 3} = \begin{bmatrix} 3 & 5 & 7 \\ 8 & 2 & 1 \\ 3 & 2 & 1 \end{bmatrix}$$

e se eliminarmos a terceira linha e a terceira coluna de $\mathbf{A}$, obteremos

$$\mathbf{B}_{2 \times 2} = \begin{bmatrix} 3 & 5 \\ 8 & 2 \end{bmatrix}$$

que corresponde a uma submatriz de $\mathbf{A}$ cuja ordem é 2×2 .

B.2 Tipos de matrizes

Matriz quadrada

Uma matriz que possui o mesmo número de linhas e colunas é denominada **matriz quadrada**.

$$\mathbf{A} = \begin{bmatrix} 3 & 4 \\ 5 & 6 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 3 & 5 & 8 \\ 7 & 3 & 1 \\ 4 & 5 & 0 \end{bmatrix}$$

Matriz diagonal

Uma matriz quadrada que possua pelo menos um elemento diferente de zero na diagonal principal (do canto superior esquerdo ao canto inferior direito) e possua zeros nas demais posições será chamada de **matriz diagonal**.

$$\mathbf{A}_{2 \times 2} = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} \quad \mathbf{B}_{3 \times 3} = \begin{bmatrix} -2 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Matriz escalar

Uma matriz diagonal cujos elementos diagonais são todos iguais é designada **matriz escalar**. Um exemplo é a matriz de variância-covariância de um termo de erro populacional do modelo clássico de regressão linear dado na Equação (C.2.3), ou seja,

$$\text{var-cov}(\mathbf{u}) = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2 \end{bmatrix}$$

Matriz identidade ou unidade

Uma matriz diagonal cujos elementos diagonais são todos 1 é chamada **matriz identidade** ou **unidade** e é denotada por **I**. Esse é um tipo especial de matriz escalar.

$$\mathbf{I}_{3 \times 3} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \mathbf{I}_{4 \times 4} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Matriz simétrica

Uma matriz quadrada cujos elementos acima da diagonal principal são imagens espelhadas dos elementos que estão abaixo da diagonal principal é chamada de **matriz simétrica**. Em outros termos, uma matriz simétrica corresponde àquela cuja transposição é igual a si mesma; ou seja, $\mathbf{A} = \mathbf{A}'$. Ou então, o elemento a_{ij} de **A** é igual ao elemento a_{ji} de **A'**. Um exemplo é a matriz de variância-covariância dado na Equação (C.2.2). Outro é a matriz de correlação apresentada em (C.5.1).

Matriz nula

Uma matriz cujos elementos são todos zero é chamada **matriz nula** e é denotada por **0**.

Vetor nulo

Uma linha ou coluna cujos elementos são todos zero é designada **vetor nulo** e também é denotada por **0**.

Matrizes iguais

Duas matrizes **A** e **B** denominam-se iguais se são da mesma ordem e seus elementos correspondentes são iguais; isto é, $a_{ij} = b_{ij}$ para todos os i e j . Por exemplo, as matrizes

$$\mathbf{A}_{3 \times 3} = \begin{bmatrix} 3 & 4 & 5 \\ 0 & -1 & 2 \\ 5 & 1 & 3 \end{bmatrix} \quad \text{e} \quad \mathbf{B}_{3 \times 3} = \begin{bmatrix} 3 & 4 & 5 \\ 0 & -1 & 2 \\ 5 & 1 & 3 \end{bmatrix}$$

são iguais, ou seja $\mathbf{A} = \mathbf{B}$.

B.3 Operações com matrizes

Soma de matrizes

Sendo $\mathbf{A} = [a_{ij}]$ e $\mathbf{B} = [b_{ij}]$. Se **A** e **B** forem da mesma ordem, definiremos a soma das matrizes como

$$\mathbf{A} + \mathbf{B} = \mathbf{C}$$

em que **C** é da mesma ordem de **A** e **B** e são obtidas por meio de $c_{ij} = a_{ij} + b_{ij}$ para todos os i e j ; **C** é obtida pela adição dos elementos correspondentes de **A** e **B**. Se essa adição pode ser efetuada, podemos afirmar que **A** e **B** são *conformes* para adição. Por exemplo, se

$$\mathbf{A} = \begin{bmatrix} 2 & 3 & 4 & 5 \\ 6 & 7 & 8 & 9 \end{bmatrix} \quad \text{e} \quad \mathbf{B} = \begin{bmatrix} 1 & 0 & -1 & 3 \\ -2 & 0 & 1 & 5 \end{bmatrix}$$

e $\mathbf{C} = \mathbf{A} + \mathbf{B}$, então

$$\mathbf{C} = \begin{bmatrix} 3 & 3 & 3 & 8 \\ 4 & 7 & 9 & 14 \end{bmatrix}$$

Subtração de matrizes

A subtração de matrizes segue o mesmo princípio da adição, exceto pelo fato de que $\mathbf{C} = \mathbf{A} - \mathbf{B}$; ou seja, subtraímos os elementos de **B** dos elementos correspondentes de **A** para obtermos **C**, desde que **A** e **B** sejam da mesma ordem.

Multiplicação escalar

Para multiplicar uma matriz **A** por um escalar λ (um número real), multiplicamos cada elemento da matriz por λ :

$$\lambda \mathbf{A} = [\lambda a_{ij}]$$

Por exemplo, se $\lambda = 2$ e

$$\mathbf{A} = \begin{bmatrix} -3 & 5 \\ 8 & 7 \end{bmatrix}$$

então,

$$\lambda \mathbf{A} = \begin{bmatrix} -6 & 10 \\ 16 & 14 \end{bmatrix}$$

Multiplicação de matrizes

Consideremos $\mathbf{A}$ como $M \times N$ e $\mathbf{B}$ como $N \times P$. O produto $\mathbf{AB}$ (nesta ordem) é definido como uma nova matriz $\mathbf{C}$ de ordem $M \times P$ de modo que:

$$c_{ij} = \sum_{k=1}^N a_{ik} b_{kj} \quad \begin{matrix} i = 1, 2, \dots, M \\ j = 1, 2, \dots, P \end{matrix}$$

Isto é, o elemento na i -ésima linha e na j -ésima coluna de $\mathbf{C}$ é obtido por meio da multiplicação dos elementos da i -ésima linha de $\mathbf{A}$ pelos elementos correspondentes da j -ésima coluna de $\mathbf{B}$ e por meio da adição de todos os termos; tal procedimento é conhecido como regra da multiplicação *linha por coluna*. Para obtermos c_{11} , que corresponde ao elemento na primeira linha e a primeira coluna de $\mathbf{C}$, multiplicamos os elementos da primeira linha de $\mathbf{A}$ pelos elementos correspondentes na primeira coluna de $\mathbf{B}$ e somamos todos os termos. De modo semelhante, para obtermos c_{12} , multiplicamos os elementos que estão na primeira linha de $\mathbf{A}$ pelos elementos correspondentes que estão na segunda coluna de $\mathbf{B}$ e somamos todos os termos e assim em diante.

Observe que, para que a multiplicação exista, as matrizes $\mathbf{A}$ e $\mathbf{B}$ devem conformar-se em relação à multiplicação; o número de colunas em $\mathbf{A}$ deve ser igual ao número de linhas em $\mathbf{B}$. Se, por exemplo,

$$\begin{aligned} \mathbf{A}_{2 \times 3} &= \begin{bmatrix} 3 & 4 & 7 \\ 5 & 6 & 1 \end{bmatrix} \quad \text{e} \quad \mathbf{B}_{3 \times 2} = \begin{bmatrix} 2 & 1 \\ 3 & 5 \\ 6 & 2 \end{bmatrix} \\ \mathbf{AB} = \mathbf{C}_{2 \times 2} &= \begin{bmatrix} (3 \times 2) + (4 \times 3) + (7 \times 6) & (3 \times 1) + (4 \times 5) + (7 \times 2) \\ (5 \times 2) + (6 \times 3) + (1 \times 6) & (5 \times 1) + (6 \times 5) + (1 \times 2) \end{bmatrix} \\ &= \begin{bmatrix} 60 & 37 \\ 34 & 37 \end{bmatrix} \end{aligned}$$

Mas se

$$\mathbf{A}_{2 \times 3} = \begin{bmatrix} 3 & 4 & 7 \\ 5 & 6 & 1 \end{bmatrix} \quad \text{e} \quad \mathbf{B}_{2 \times 2} = \begin{bmatrix} 2 & 3 \\ 5 & 6 \end{bmatrix}$$

o produto de $\mathbf{AB}$ não é definido, na medida em que $\mathbf{A}$ e $\mathbf{B}$ não são conformes à multiplicação.

Propriedades da multiplicação de matrizes

1. A multiplicação de matrizes não é necessariamente *comutativa*; em geral $\mathbf{AB} \neq \mathbf{BA}$. Portanto, a ordem em que as matrizes são multiplicadas é muito importante. $\mathbf{AB}$ significa que $\mathbf{A}$ é *pós-multiplicada* por $\mathbf{B}$ ou $\mathbf{B}$ é *pré-multiplicada* por $\mathbf{A}$.
2. Ainda que $\mathbf{AB}$ e $\mathbf{BA}$ existam, as matrizes resultantes podem não ser de mesma ordem. Assim, se $\mathbf{A}$ é $M \times N$ e $\mathbf{B}$ é $N \times M$, $\mathbf{AB}$ é $M \times M$ enquanto $\mathbf{BA}$ é $N \times N$ e, por conseguinte, de ordens diferentes.
3. Ainda que $\mathbf{A}$ e $\mathbf{B}$ sejam matrizes quadradas, de modo que $\mathbf{AB}$ e $\mathbf{BA}$ sejam ambas definidas, as matrizes resultantes não serão necessariamente iguais. Por exemplo, se

$$\mathbf{A} = \begin{bmatrix} 4 & 7 \\ 3 & 2 \end{bmatrix} \quad \text{e} \quad \mathbf{B} = \begin{bmatrix} 1 & 5 \\ 6 & 8 \end{bmatrix}$$

então,

$$\mathbf{AB} = \begin{bmatrix} 46 & 76 \\ 15 & 31 \end{bmatrix} \quad \text{e} \quad \mathbf{BA} = \begin{bmatrix} 19 & 17 \\ 48 & 58 \end{bmatrix}$$

e $\mathbf{AB} \neq \mathbf{BA}$. Um exemplo de $\mathbf{AB} = \mathbf{BA}$ ocorre quando tanto $\mathbf{A}$ quanto $\mathbf{B}$ são matrizes identidade.

4. Um vetor linha pós-multiplicado por um vetor coluna é um escalar. Desse modo, considere os resíduos dos mínimos quadrados ordinários $\hat{u}_1, \hat{u}_2, \dots, \hat{u}_n$. Se $\mathbf{u}'$ for um vetor coluna e $\mathbf{u}$ for um vetor linha, teremos

$$\begin{aligned}\hat{\mathbf{u}}'\hat{\mathbf{u}} &= [\hat{u}_1 \quad \hat{u}_2 \quad \hat{u}_3 \quad \cdots \quad \hat{u}_n] \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \hat{u}_3 \\ \vdots \\ \hat{u}_n \end{bmatrix} \\ &= \hat{u}_1^2 + \hat{u}_2^2 + \hat{u}_3^2 + \cdots + \hat{u}_n^2 \\ &= \sum_i \hat{u}_i^2 \quad \text{um escalar [veja a Equação (C.3.5)]}\end{aligned}$$

5. Um vetor coluna pós-multiplicado por um vetor linha é uma matriz. Como exemplo, considere os termos de erro de população no modelo clássico de regressão linear, ou seja, $u_1, u_2, \dots, u_n$. Se $\mathbf{u}$ for um vetor coluna e $\mathbf{u}'$ um vetor linha, obteremos

$$\begin{aligned} \mathbf{u}\mathbf{u}' &= \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_n \end{bmatrix} \begin{bmatrix} u_1 & u_2 & u_3 & \cdots & u_n \end{bmatrix} \\ &= \begin{bmatrix} u_1^2 & u_1u_2 & u_1u_3 & \cdots & u_1u_n \\ u_2u_1 & u_2^2 & u_2u_3 & \cdots & u_2u_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ u_nu_1 & u_nu_2 & u_nu_3 & \cdots & u_n^2 \end{bmatrix} \end{aligned}$$

que é uma matriz de ordem $n \times n$. Observe que a matriz anterior é simétrica.

6. Uma matriz pós-multiplicada por um vetor coluna é um vetor coluna.
7. Um vetor linha pós-multiplicado por uma matriz é um vetor linha.
8. A multiplicação de matrizes é *associativa*; $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$, em que $\mathbf{A}$ é $M \times N$, $\mathbf{B}$ é $N \times P$ e $\mathbf{C}$ é $P \times K$.
9. A multiplicação de matrizes é distributiva em relação à adição; $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$ e $(\mathbf{B} + \mathbf{C})\mathbf{A} = \mathbf{BA} + \mathbf{CA}$.

Transposição de matrizes

Já definimos o processo de transposição de matrizes como o intercâmbio de linhas e colunas de uma matriz (ou um vetor). Vamos expor agora algumas das propriedades da transposição.

1. A transposição de uma matriz transposta é a própria matriz original. Assim, $(\mathbf{A}')' = \mathbf{A}$.
2. Se $\mathbf{A}$ e $\mathbf{B}$ são conformes para a adição, então $\mathbf{C} = \mathbf{A} + \mathbf{B}$ e $\mathbf{C}' = (\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}'$. A transposição da soma de duas matrizes é a soma de suas transposições.
3. Se $\mathbf{AB}$ é definida, $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$. A transposição do produto de duas matrizes é o produto de suas transposições na ordem inversa. Isso pode ser generalizado: $(\mathbf{ABCD})' = \mathbf{D}'\mathbf{C}'\mathbf{B}'\mathbf{A}'$.
4. A transposição de uma matriz identidade $\mathbf{I}$ corresponde à própria matriz identidade; $\mathbf{I}' = \mathbf{I}$.
5. A transposição de um escalar é o próprio escalar. Assim, se λ é um escalar, $\lambda' = \lambda$.
6. A transposição de $(\lambda\mathbf{A})'$ é $\lambda\mathbf{A}'$ em que λ é um escalar. [Observe: $(\lambda\mathbf{A})' = \mathbf{A}'\lambda' = \mathbf{A}'\lambda = \lambda\mathbf{A}'$.]

7. Se $\mathbf{A}$ é uma matriz quadrada de modo que $\mathbf{A} = \mathbf{A}'$, então $\mathbf{A}$ é uma matriz simétrica. (Veja a definição de matriz simétrica na Seção B.2.)

Inversão de matrizes

A inversa de uma matriz quadrada $\mathbf{A}$, denotada por $\mathbf{A}^{-1}$ (lida como “ $\mathbf{A}$ inversa”), se existir, é uma única matriz quadrada, de modo que

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

em que $\mathbf{I}$ é uma matriz identidade cuja ordem é a mesma de $\mathbf{A}$. Por exemplo,

$$\mathbf{A} = \begin{bmatrix} 2 & 4 \\ 6 & 8 \end{bmatrix} \quad \mathbf{A}^{-1} = \begin{bmatrix} -1 & \frac{1}{2} \\ \frac{6}{8} & -\frac{1}{4} \end{bmatrix} \quad \mathbf{A}\mathbf{A}^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \mathbf{I}$$

Veremos como $\mathbf{A}^{-1}$ é calculado depois de estudarmos o tópico dos determinantes. Enquanto isso, observe as seguintes propriedades da matriz inversa:

1. $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$; ou seja, a inversa do produto de duas matrizes é o produto de suas inversas na ordem inversa.
2. $(\mathbf{A}^{-1})' = (\mathbf{A}')^{-1}$; ou seja, a transposição de $\mathbf{A}$ inversa é a inversa de $\mathbf{A}$ transposta.

B.4 Determinantes

Para cada matriz quadrada, $\mathbf{A}$ corresponde um número (escalar) conhecido como o determinantes da matriz, que é denotado por $\det \mathbf{A}$ ou pelo símbolo $|\mathbf{A}|$, em que $|\quad|$ significa “o determinante de”. Observe que a matriz por si não possui qualquer valor numérico, mas o determinante de uma matriz é um número.

$$\mathbf{A} = \begin{bmatrix} 1 & 3 & -7 \\ 2 & 5 & 0 \\ 3 & 8 & 6 \end{bmatrix} \quad |\mathbf{A}| = \begin{vmatrix} 1 & 3 & -7 \\ 2 & 5 & 0 \\ 3 & 8 & 6 \end{vmatrix}$$

O $|\mathbf{A}|$ neste exemplo é chamado de determinante de ordem 3 por ser associado a uma matriz de ordem 3×3 .

Avaliação de um determinante

O processo de encontrar o valor de um determinante é conhecido como avaliação, expansão ou redução do determinante. Isso é feito ao manipular as entradas da matriz de uma forma bem definida.

Avaliação de um determinante 2×2

Se

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

seu determinante é avaliado como se segue:

$$|\mathbf{A}| = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}$$

que é obtido pela multiplicação cruzada dos elementos na diagonal principal e subtraindo desse produto a multiplicação cruzada dos elementos na outra diagonal da matriz $\mathbf{A}$, como indicado pelas setas.

Avaliação de um determinante 3×3

Se

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

então

$$|\mathbf{A}| = a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} + a_{12}a_{23}a_{31} - a_{12}a_{21}a_{33} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31}$$

Um exame cuidadoso da avaliação do determinante 3×3 demonstra que:

1. Cada termo na expansão do determinante contém um e apenas um elemento de cada linha e de cada coluna.
2. O número de elementos em cada termo é o mesmo do número de linhas (ou colunas) na matriz. Portanto, um determinante 2×2 possui dois elementos em cada termo de sua expansão, um determinante 3×3 possui três elementos em cada termo de sua expansão e assim por diante.
3. Os termos na expansão alternam-se em sinal de $+$ para $-$.
4. Um determinante 2×2 possui dois termos em sua expansão e um determinante 3×3 possui seis termos. A regra geral é: o determinante de ordem $N \times N$ possui $N! = N(N-1)(N-2) \cdots 3 \cdot 2 \cdot 1$ termos em sua expansão, em que $N!$ lê-se “fatorial de N ”. Seguindo essa regra, um determinante de ordem 5×5 possuirá $5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$ termos em sua expansão.¹

Propriedades dos determinantes

1. Uma matriz cujo valor do determinante é zero é chamada de **matriz singular**, enquanto uma matriz com um determinante não zero é chamada de matriz **não singular**. O inverso de uma matriz, como anteriormente definido, não existe para uma matriz singular.
2. Se todos os elementos de toda linha de $\mathbf{A}$ forem zero, seu determinante será zero. Assim,

$$|\mathbf{A}| = \begin{vmatrix} 0 & 0 & 0 \\ 3 & 4 & 5 \\ 6 & 7 & 8 \end{vmatrix} = 0$$

3. $|\mathbf{A}'| = |\mathbf{A}|$; isto é, os determinantes de $\mathbf{A}$ e da transposta $\mathbf{A}$ são os mesmos.
4. Intercambiando quaisquer das duas linhas ou das duas colunas de uma matriz $\mathbf{A}$, modifica-se o sinal de $|\mathbf{A}|$.

EXEMPLO 1 SE

$$\mathbf{A} = \begin{bmatrix} 6 & 9 \\ -1 & 4 \end{bmatrix} \quad \text{e} \quad \mathbf{B} = \begin{bmatrix} -1 & 4 \\ 6 & 9 \end{bmatrix}$$

em que $\mathbf{B}$ é obtido intercambiando das linhas de $\mathbf{A}$, então

$$\begin{aligned} |\mathbf{A}| &= 24 - (-9) & \text{e} & & |\mathbf{B}| &= -9 - (24) \\ &= 33 & & & &= -33 \end{aligned}$$

5. Se cada elemento de uma linha ou de uma coluna de $\mathbf{A}$ for multiplicado por um escalar λ , então $|\mathbf{A}|$ é multiplicado por λ .

¹ Para avaliar o determinante de uma matriz $N \times N$, $\mathbf{A}$, veja as referências.

EXEMPLO 2 SE

$$\mathbf{A} = \begin{bmatrix} 6 & 9 \\ -1 & 4 \end{bmatrix} \quad \text{e} \quad \mathbf{B} = \begin{bmatrix} -1 & 4 \\ 6 & 9 \end{bmatrix}$$

e multiplicarmos a primeira linha de $\mathbf{A}$ por 5 para obter

$$\mathbf{B} = \begin{bmatrix} 25 & -40 \\ 2 & 4 \end{bmatrix}$$

podemos verificar que $|\mathbf{A}| = 36$ e $|\mathbf{B}| = 180$, que é 5 $|\mathbf{A}|$.

6. Se duas linhas ou colunas de uma matriz forem idênticas, seu determinante será zero.
7. Se uma linha ou uma coluna de uma matriz for múltipla de outra linha ou coluna daquela matriz, seu determinante será zero. Assim, se

$$\mathbf{A} = \begin{bmatrix} 4 & 8 \\ 2 & 4 \end{bmatrix}$$

em que a primeira linha de $\mathbf{A}$ é duas vezes a segunda linha, $|\mathbf{A}| = 0$. De maneira geral, se qualquer linha (coluna) de uma matriz for uma combinação linear de outras linhas (colunas), seu determinante será zero.

8. $|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}|$; o determinante do produto de duas matrizes é o produto dos seus determinantes (individuais).

Posto de uma matriz

O posto de uma matriz é a ordem da maior submatriz quadrada cujo determinante não é zero.

EXEMPLO 3

$$\mathbf{A} = \begin{bmatrix} 3 & 6 & 6 \\ 0 & 4 & 5 \\ 3 & 2 & 1 \end{bmatrix}$$

Pode-se verificar que $|\mathbf{A}| = 0$. Em outras palavras, $\mathbf{A}$ é uma matriz singular. Embora sua ordem seja 3×3 , seu posto é menor do que 3. Na verdade, ele é 2, porque podemos encontrar uma submatriz 2×2 cujo determinante não é zero. Por exemplo, se excluirmos a primeira linha e a primeira coluna de $\mathbf{A}$, obtemos

$$\mathbf{B} = \begin{bmatrix} 4 & 5 \\ 2 & 1 \end{bmatrix}$$

cujo determinante é -6 , que é não zero. Portanto, o posto de $\mathbf{A}$ é 2. Como anteriormente observado, o inverso de uma matriz singular não existe. Para uma matriz $\mathbf{A}$ de origem $N \times N$, seu posto tem de ser N para que a sua inversa exista; se seu posto for menor do que N , $\mathbf{A}$ será singular.

Menor

Se a i -ésima linha e a j -ésima coluna de uma matriz $\mathbf{A}$ de origem $N \times N$ são excluídas, o determinante da submatriz resultante é chamado de o **menor** do elemento a_{ij} (o elemento na interseção da i -ésima linha e a j -ésima coluna) e é denotado por $|\mathbf{M}_{ij}|$.

EXEMPLO 4

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

O menor de a_{11} é

$$|\mathbf{M}_{11}| = \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} = a_{22}a_{33} - a_{23}a_{32}$$

De forma semelhante, o menor de a_{21} é

$$|\mathbf{M}_{21}| = \begin{vmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{vmatrix} = a_{12}a_{33} - a_{13}a_{32}$$

Os menores de outros elementos de $\mathbf{A}$ podem ser encontrados de maneira parecida.

Cofator

O cofator do elemento a_{ij} de uma matriz $\mathbf{A}$ de origem $N \times N$, denotado por c_{ij} , é definido como:

$$c_{ij} = (-1)^{i+j} |\mathbf{M}_{ij}|$$

Em outras palavras, um cofator é um menor *sinalizado*: com sinal positivo se $i + j$ for par e negativo se $i + j$ for ímpar. Assim, o cofator do elemento a_{11} da matriz $\mathbf{A}$ 3 x 3 anteriormente dada é $a_{22}a_{33} - a_{23}a_{32}$, enquanto o cofator do elemento a_{21} é $-(a_{12}a_{33} - a_{13}a_{32})$, uma vez que a soma dos subscritos 2 e 1 é 3, que é um número ímpar.

Matriz de cofator

Substituindo os elementos a_{ij} de uma matriz $\mathbf{A}$ pelos seus cofatores, obtemos uma matriz conhecida como **matriz de cofator** de $\mathbf{A}$, denotada por $(\text{cof } \mathbf{A})$.

Matriz adjunta

A matriz adjunta, escrita como $(\text{adj } \mathbf{A})$, é a transposta da matriz de cofator; $(\text{adj } \mathbf{A}) = (\text{cof } \mathbf{A})'$.

B.5 Encontrando a inversa de uma matriz quadrada

Se $\mathbf{A}$ é quadrada e não singular ($|\mathbf{A}| \neq 0$), a sua inversa $\mathbf{A}^{-1}$ pode ser encontrada da seguinte forma:

$$\mathbf{A}^{-1} = \frac{1}{|\mathbf{A}|} (\text{adj } \mathbf{A})$$

Os passos envolvidos no cálculo são os seguintes:

1. Descubra o determinante de $\mathbf{A}$. Se não for zero, execute o passo 2.
2. Substitua cada elemento a_{ij} de $\mathbf{A}$ por seu cofator para obter a matriz de cofator.
3. Transponha a matriz de cofator para obter a matriz adjunta.
4. Divida cada elemento da matriz adjunta por $|\mathbf{A}|$.

EXEMPLO 5

Descubra a inversa da matriz

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 5 & 7 & 4 \\ 2 & 1 & 3 \end{bmatrix}$$

Passo 1. Primeiro, descobrimos o determinante da matriz. Aplicando as regras de expansão de um determinante 3×3 dado previamente, obtemos $|\mathbf{A}| = -24$.

Passo 2. Agora obtemos a matriz de cofator, por exemplo, $\mathbf{C}$:

$$\begin{aligned} \mathbf{C} &= \begin{bmatrix} \begin{vmatrix} 7 & 4 \\ 1 & 3 \end{vmatrix} & -\begin{vmatrix} 5 & 4 \\ 2 & 3 \end{vmatrix} & \begin{vmatrix} 5 & 7 \\ 2 & 1 \end{vmatrix} \\ -\begin{vmatrix} 2 & 3 \\ 1 & 3 \end{vmatrix} & \begin{vmatrix} 1 & 3 \\ 2 & 3 \end{vmatrix} & -\begin{vmatrix} 1 & 2 \\ 2 & 1 \end{vmatrix} \\ \begin{vmatrix} 2 & 3 \\ 7 & 4 \end{vmatrix} & -\begin{vmatrix} 1 & 3 \\ 5 & 4 \end{vmatrix} & \begin{vmatrix} 1 & 2 \\ 5 & 7 \end{vmatrix} \end{bmatrix} \\ &= \begin{bmatrix} 17 & -7 & -9 \\ -3 & -3 & 3 \\ -13 & 11 & -3 \end{bmatrix} \end{aligned}$$

Passo 3. Transpondo a matriz de cofator anterior, obtemos a seguinte matriz adjunta:

$$(\text{adj } \mathbf{A}) = \begin{bmatrix} 17 & -3 & -13 \\ -7 & -3 & 11 \\ -9 & 3 & -3 \end{bmatrix}$$

Passo 4. Agora dividimos os elementos de $(\text{adj } \mathbf{A})$ pelo valor do determinante obtido, -24 , para obter

$$\begin{aligned} \mathbf{A}^{-1} &= -\frac{1}{24} \begin{bmatrix} 17 & -3 & -13 \\ -7 & -3 & 11 \\ -9 & 3 & -3 \end{bmatrix} \\ &= \begin{bmatrix} -\frac{17}{24} & \frac{3}{24} & \frac{13}{24} \\ \frac{7}{24} & \frac{3}{24} & -\frac{11}{24} \\ \frac{9}{24} & -\frac{3}{24} & \frac{3}{24} \end{bmatrix} \end{aligned}$$

Podemos facilmente verificar que

$$\mathbf{A}\mathbf{A}^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

que é uma matriz identidade. O leitor deve verificar que, para o exemplo ilustrativo dado no Apêndice C (veja a Seção C.10), a inversa da matriz $\mathbf{X}'\mathbf{X}$ é semelhante à demonstrada na Equação (C.10.5).

B.6 Diferenciação matricial

Para seguirmos o material no Apêndice CA, Seção CA.2, precisamos considerar algumas regras da diferenciação matricial.

REGRA 1 Se $\mathbf{a}' = [a_1 \ a_2 \ \dots \ a_n]$ é um vetor linha de números e

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

é um vetor coluna das variáveis $x_1, x_2, \dots, x_n$, então

$$\frac{\partial(\mathbf{a}'\mathbf{x})}{\partial\mathbf{x}} = \mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$$

REGRA 2 Considere a matriz $\mathbf{x}'\mathbf{A}\mathbf{x}$ tal que

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Então

$$\frac{\partial(\mathbf{x}'\mathbf{A}\mathbf{x})}{\partial\mathbf{x}} = 2\mathbf{A}\mathbf{x}$$

que é um vetor coluna de n elementos, ou

$$\frac{\partial(\mathbf{x}'\mathbf{A}\mathbf{x})}{\partial\mathbf{x}} = 2\mathbf{x}'\mathbf{A}$$

que é um vetor linha de n elementos.

Referências

- CHIANG, Alpha C. *Fundamental methods of mathematical economics*. 3. ed. Nova York: McGraw-Hill, 1984, caps. 4 e 5. A obra apresenta uma discussão avançada sobre álgebra linear.
- HADLEY, G. *Linear algebra*. Reading, Mass.: Addison-Wesley, 1961. A obra apresenta uma discussão avançada.