capítulo 3

Distribuições de probabilidades

Objetivos do estudo

Ao final deste capítulo, você deverá ser capaz de:

- Reconhecer que o resultado da maioria dos experimentos probabilísticos (por exemplo, o resultado obtido no lançamento de um dado) pode ser descrito como uma variável aleatória.
- Perceber como o comportamento de uma variável aleatória pode ser freqüentemente sintetizado por uma distribuição de probabilidades (uma fórmula matemática).
- Identificar as distribuições de probabilidades mais comuns e estar ciente de suas aplicações.
- Resolver uma variedade de problemas de probabilidades usando a distribuição de probabilidades apropriada.

Neste capítulo, os conceitos de probabilidade apresentados no capítulo anterior são generalizados com o uso da idéia de **distribuição de probabilidades**. Uma distribuição de probabilidades enumera, de alguma forma, todos os resultados possíveis de um experimento probabilístico e a probabilidade associada a cada um dos resultados. Por exemplo, o experimento mais simples é o lançamento de uma moeda, para o qual os resultados possíveis são cara (C) ou coroa (Cr), cada um com probabilidade igual a 0,5. A distribuição de probabilidades pode ser expressa de várias maneiras: em palavras, por um gráfico ou uma fórmula. Para o lançamento de uma moeda, o gráfico é mostrado na figura 3.1, e a representação matemática é:

$$Pr(C) = \frac{1}{2}$$

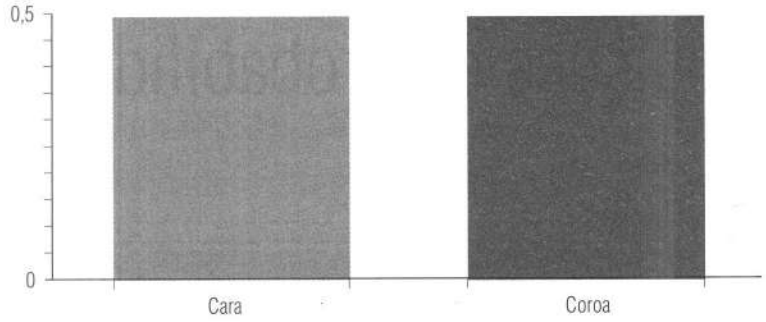
$$Pr(Cr) = \frac{1}{2}$$

As diferentes formas de apresentação são equivalentes, mas, dependendo da finalidade específica, uma pode ser mais apropriada do que a outra.

Algumas distribuições de probabilidades ocorrem com freqüência e, por causa disso, são bastante conhecidas. Por esse motivo, têm nomes pelos quais podemos chamá-las facilmente; por exemplo, **distribuição binomial** ou **distribuição normal**. Na realidade, essas são *famílias* de distribuição. Um único lançamento de uma moeda dá origem a um membro da distribuição binomial; dois lançamentos originariam outro membro dessa família. Essas duas distribuições diferem entre si apenas devido ao nú-

FIGURA 3.1

Distribuição de probabilidades do lançamento de uma moeda



mero de lançamentos. Se uma moeda enviesada fosse lançada, isso levaria a uma outra distribuição binomial, mas ela seria distinta das duas distribuições anteriores por causa da probabilidade diferenciada de sair o resultado cara.

Para compreender bem a idéia de distribuição de probabilidades, introduzimos inicialmente um novo conceito, o de **variável aleatória**. Como será visto mais adiante neste capítulo, a média amostral é uma variável aleatória importante, e para entender como se deve fazer inferências a partir dela é preciso reconhecê-la como uma variável aleatória.

Variáveis aleatórias

Alguns exemplos de variáveis aleatórias já foram mencionados no capítulo anterior, tais como o resultado do lançamento de uma moeda, ou o número de garotos numa família de cinco crianças. Uma variável aleatória é uma variável cujo resultado ou valor decorre do acaso e é, portanto, imprevisível – embora a gama de resultados possíveis e a probabilidade de cada um deles possam ser conhecidas. Por exemplo, é impossível saber com antecedência o resultado do lançamento de uma moeda, mas temos certeza de que será cara ou coroa, cada um com probabilidade igual a 0,5. O número de caras em 250 lançamentos é uma outra variável aleatória, que pode assumir qualquer valor entre 0 e 250, embora valores próximos de 125 sejam os mais prováveis. É muito improvável que você consiga 250 caras lançando uma moeda imparcial!

O horário de partida de um trem é uma variável aleatória. Ela pode estar marcada para as 11h15, mas provavelmente (ou é quase certo!) o trem não sairá no horário exato. Se fosse escolhida uma amostra de dez jogadores de basquete, calculando-se sua estatura média, esta seria uma variável aleatória. Nesse último caso, é o processo de extração de uma amostra que introduz a variabilidade, e esta, por sua vez, faz com que a média resultante seja uma variável aleatória. Se o experimento fosse repetido, seriam obtidos uma amostra distinta e um valor diferente da variável aleatória.

Os exemplos anteriores podem ser contrastados com algumas coisas que *não* são variáveis aleatórias. Se calculássemos a estatura média de *todos* os jogadores de basquete, o resultado não seria uma variável aleatória. Desta vez, não há procedimento de amostragem para introduzir variabilidade no resultado. Se o experimento fosse repetido, o mesmo resultado seria obtido, pois as mesmas pessoas seriam medidas pela segunda vez (supondo, é claro, que a população não tenha sido alterada). O fato de o valor de algo ser desconhecido não qualifica este algo como uma variável aleatória. Essa é uma distinção importante, pois é legítimo fazer afirmações probabilísticas a respeito de variáveis aleatórias ("a probabilidade de que a altura média numa amostra de jogadores de basquete seja superior a 1,95 m é 60%"), mas não sobre parâmetros ("a probabilidade de que o papa tenha mais de 1,80 m é 60%"). Eis aqui novamente uma diferença de opinião entre as escolas freqüentista e subjetiva. Este último grupo argumentaria que é possível fazer afirmações probabilísticas sobre a estatura do papa – o que configura uma maneira de expressar a falta de conhecimento a respeito do valor verdadeiro. Já os freqüentistas diriam que a altura do papa é um fato que não conhecemos, mas que isso não faz com que ela seja uma variável aleatória.

Distribuição binomial

Uma das distribuições mais simples que uma variável aleatória pode ter é a binomial. A distribuição binomial surge sempre que o experimento probabilístico subjacente tiver apenas dois resultados possíveis, como cara ou coroa no lançamento de uma moeda. Mesmo que uma moeda seja lançada muitas vezes (fazendo com que se possa terminar com uma, duas, três... etc. caras no total), o experimento *subjacente* tem somente dois resultados, de modo que se deve usar a distribuição binomial. Um contra-exemplo é o lançamento de um dado, que tem seis resultados possíveis (nesse caso, seria utilizada a distribuição multinomial, não discutida neste livro). Note, porém, que, se estivéssemos interessados somente no lançamento de um seis, *poderíamos* usar a distribuição binomial definindo os dois resultados possíveis como "seis" ou "não-seis". Em estatística, com uma transformação apropriada dos dados, é comum que se consiga usar distribuições diferentes para atacar o mesmo problema. Veremos mais a respeito disso neste capítulo.

A distribuição binomial pode, portanto, ser aplicada ao tipo de problema encontrado no capítulo anterior, relativo ao sexo de crianças. Ela oferece uma fórmula geral de cálculo da probabilidade de r garotos em n nascimentos ou, em termos mais gerais, r "sucessos" em n tentativas.[1] Nós a utilizaremos para calcular as probabilidades de 0, 1,..., 5 garotos em cinco nascimentos.

Para que a distribuição binomial seja aplicada, precisamos primeiro supor a independência de eventos sucessivos. Imagine que, para cada nascimento:

1. A associação de menino com "sucesso" é puramente formal, e não tem finalidade valorativa!

$Pr(garoto) = P = ½$

Segue-se que:

$Pr(menina) = 1 - Pr(garoto) = 1 - P = ½$

Embora tenhamos $P = ½$ nesse exemplo, a distribuição binomial pode ser aplicada com qualquer valor de P entre 0 e 1.

Inicialmente, consideraremos o caso de $r = 5$ e $n = 5$, isto é, cinco garotos em cinco nascimentos. A probabilidade é encontrada com o uso da regra de multiplicação:

$$Pr(r = 5) = P \times P \times P \times P \times P = P^5 = (½)^5 = 1/32$$

A probabilidade de quatro garotos (e implicitamente uma menina) é:

$$Pr(r = 4) = P \times P \times P \times P \times (1 - P) = 1/32$$

Mas isso considera apenas uma seqüência de quatro garotos e uma menina. O enunciado original do problema não mencionava uma seqüência específica de crianças. Há cinco seqüências possíveis (a única menina pode estar em qualquer uma das cinco posições na ordem). Podemos empregar a fórmula combinatória nCr para calcular o número de seqüências, produzindo 5C4 = 5. Portanto, a probabilidade de quatro garotos e uma menina em qualquer ordem é 5/32. Resumindo, a fórmula para quatro garotos e uma menina é:

$$Pr(r = 4) = 5C4 \times P^4 \times (1 - P)$$

Para três garotos (e duas meninas) obtemos:

$$Pr(r = 3) = 5C3 \times P^3 \times (1 - P)^2 = 10 \times 1/8 \times 1/4 = 10/32$$

De maneira análoga:

$$Pr(r = 2) = 5C2 \times P^2 \times (1 - P)^3 = 10/32$$

$$Pr(r = 1) = 5C1 \times P^1 \times (1 - P)^4 = 5/32$$

$$Pr(r = 0) = 5C0 \times P^0 \times (1 - P)^5 = 1/32$$

Para confirmar nossos cálculos, podemos observar que a soma das probabilidades é igual a 1, como deve ocorrer, já que enumeramos todas as possibilidades.

Surge um padrão bastante claro. A probabilidade de r garotos em n nascimentos é dada por:

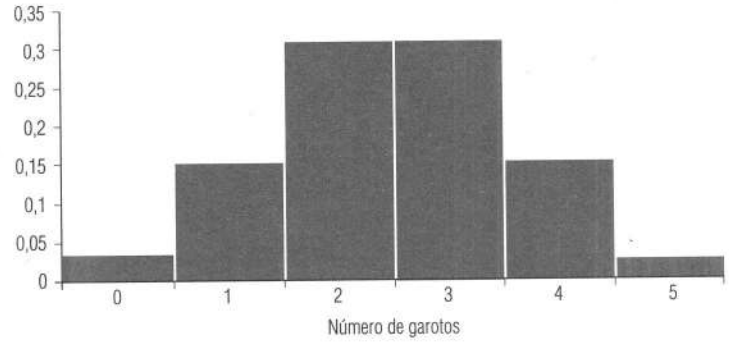
$$Pr(r) = nCr \times P^r \times (1 - P)^{n-r}$$

e esta é conhecida como fórmula ou distribuição binomial. A distribuição binomial é apropriada para a análise de problemas com as seguintes características:

- Há um número (n) de tentativas.
- Cada tentativa tem apenas dois resultados possíveis: "sucesso" (com probabilidade P) e "insucesso" (com probabilidade 1 – P), sendo que os resultados são independentes entre tentativas.
- A probabilidade P não se altera de uma tentativa para outra.

FIGURA 3.2

Distribuição de probabilidades do número de meninos em cinco crianças



As probabilidades calculadas pela fórmula binomial podem ser ilustradas num diagrama, como mostra a figura 3.2. Isso é muito semelhante à distribuição de freqüências relativas apresentada no capítulo 1. Esta distribuição baseava-se em dados empíricos (de riqueza), enquanto a distribuição binomial de probabilidades é uma construção teórica apoiada nos princípios básicos da teoria de probabilidade.

Como afirmado anteriormente, a binomial é na verdade uma família de distribuições; cada membro dessa família se distingue de outro por dois **parâmetros**, n e P. A binomial, portanto, é uma distribuição de dois parâmetros e, uma vez conhecidos os seus valores, ela é determinada completamente (isto é, Pr(r) pode ser calculada para todos os valores de r). Para mostrar a diferença entre membros da família da distribuição binomial, a figura 3.3 apresenta três outras distribuições binomiais, com valores distintos de P e n. Pode-se observar que, para P = ½, a distribuição é simétrica, ao passo que para todos os outros valores ela é assimétrica à esquerda ou à direita. A parte (b) da figura ilustra a distribuição relacionada ao exemplo resolvido de lançamento de um dado, que é descrito a seguir.

Como a distribuição binomial depende somente dos dois valores n e P, pode-se usar uma notação simplificada em lugar da própria fórmula. Uma variável aleatória r, que tem distribuição binomial com os parâmetros n e P, é escrita, em termos gerais, da seguinte maneira:

(3.1) $r \sim B(n, P)$

Portanto, para o exemplo anterior de número de crianças, no qual r representa o número de meninos,

$r \sim B(5, ½)$

Essa é simplesmente uma maneira sucinta e conveniente de expressar a informação disponível; não envolve novos problemas de natureza conceitual. Escrever

$r \sim B(n, P)$

TABELA 3.1 Cálculo da média e da variância da distribuição binomial

r	Pr(r)	r × Pr(r)	$r^2 \times Pr(r)$
0	1/32	0	0
1	5/32	5/32	5/32
2	10/32	20/32	40/32
3	10/32	30/32	90/32
4	5/32	20/32	80/32
5	1/32	5/32	25/32
Total	32/32	80/32	240/32

é simplesmente uma simplificação de

$$Pr(r) = nCr \times P^r \times (1 - P)^{n-r}$$

Média e variância da distribuição binomial

No capítulo 1, calculamos a média e a variância de um conjunto de dados no caso da distribuição de riqueza entre famílias. A representação gráfica daquela distribuição (figura 1.9) não parece muito diferente de uma das ilustrações da distribuição binomial mostradas na figura 3.3. Isso sugere que podemos calcular a média e a variância de uma distribuição binomial assim como fizemos para a distribuição empírica de níveis de riqueza. O cálculo da média nos daria a resposta a uma pergunta tal como: "Se tivermos uma família com cinco crianças, quantas podemos esperar que sejam meninos?" Intuitivamente, a resposta parece clara, 2,5, muito embora tal família não possa existir! A fórmula binomial confirma essa intuição.

A média e a variância são calculadas mais facilmente montando-se uma tabela de freqüências relativas com base nas freqüências binomiais. Isso é mostrado na tabela 3.1 para os valores n = 5 e P = ½. Observe que r é equivalente a x em nossa notação usual, e Pr(r), a freqüência relativa, a $f(x)/\Sigma f(x)$. A média dessa distribuição é dada por:

$$(3.2)\quad E(r) = \frac{\sum r \times Pr(r)}{\sum Pr(r)} = \frac{80/32}{32/32} = 2,5$$

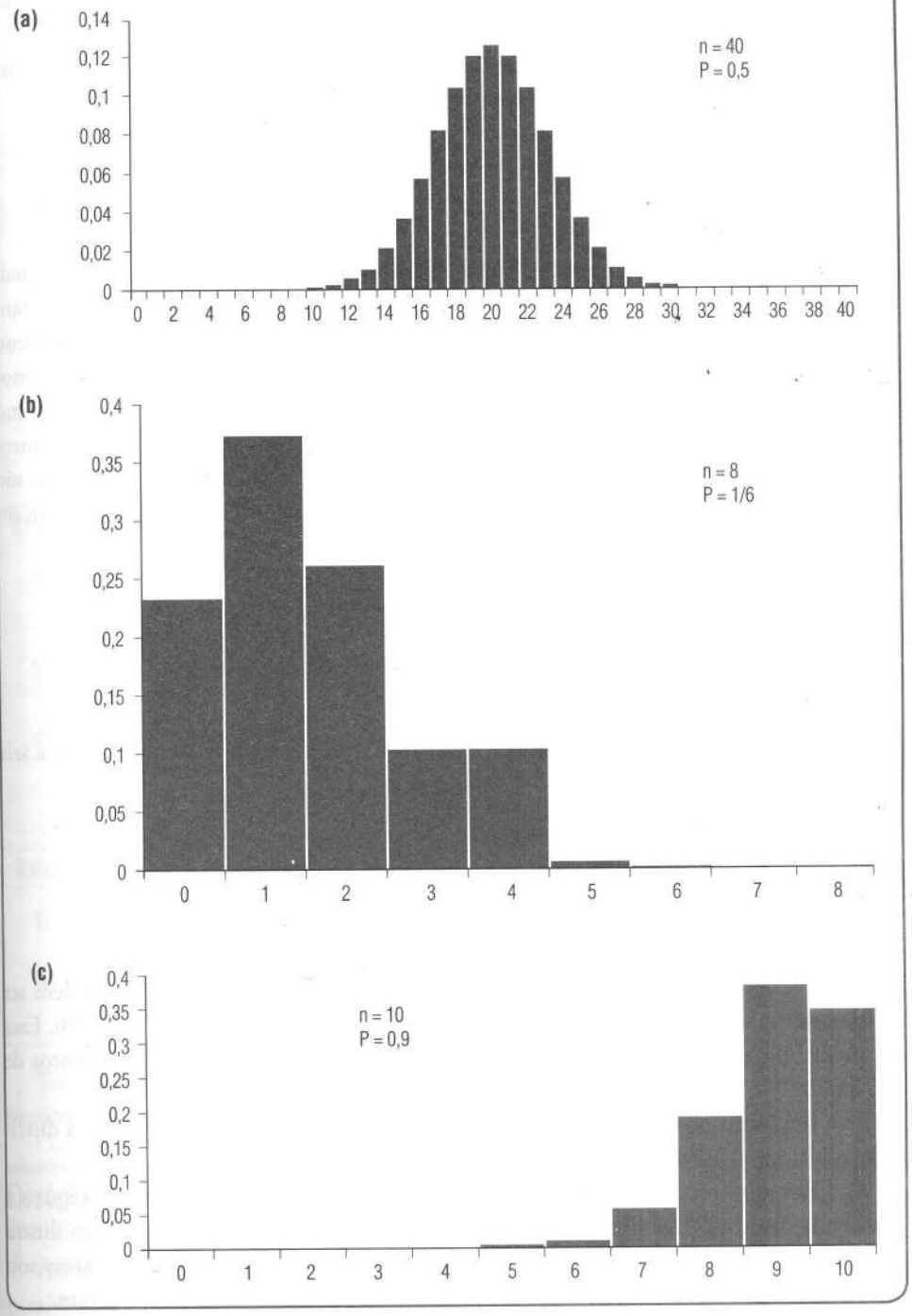
e a variância é dada por:

$$(3.3)\quad V(r) = \frac{\sum r^2 \times Pr(r)}{\sum Pr(r)} - \mu^2 = \frac{240/32}{32/32} - 2,5^2 = 1,25$$

O valor médio nos diz que, numa família de cinco crianças, esperaríamos, em média, dois meninos e meio. Obviamente, nenhuma família pode ser assim; essa é a média de todas as famílias. É mais difícil interpretar a variância intuitivamente, mas ela indica como o número de garotos se distribuiria em torno da média de 2,5.

FIGURA 3.3

Distribuições binomiais com parâmetros de valores diferentes



Existe uma maneira mais rápida de calcular a média e a variância da distribuição binomial. É possível mostrar que a média pode ser calculada por nP, ou seja, o número de tentativas vezes a probabilidade de sucesso. Por exemplo, numa família de cinco crianças e probabilidades iguais de que cada criança seja menino ou menina, esperaríamos que $nP = 5 \times ½ = 2,5$ fossem garotos.

A variância pode ser calculada por $nP(1 - P)$. Isso dá $5 \times ½ \times ½ = 1,25$, como obtido antes por meio de cálculo extensivo.

Exemplo resolvido 3.1

Lançamento de um dado

Se um dado for lançado quatro vezes, qual é a probabilidade de obter dois ou mais resultados iguais a seis? Esse é um problema que envolve experimentos repetidos (lançamentos de um dado), mas com dois tipos de resultado para cada lançamento: sucesso (um seis) ou insucesso (qualquer outro número, exceto seis). Note que combinamos diversas possibilidades (escores de 1, 2, 3, 4 ou 5), e representamos todas elas como insucessos. A probabilidade de sucesso (1/6) não varia de um experimento para outro, e por isso o uso da distribuição binomial é apropriado. Os valores dos parâmetros são $n = 4$ e $P = 1/6$. Sendo r a variável aleatória "de número seis em quatro lançamentos do dado", então

(3.4) $r \sim B(4, 1/6)$

Portanto:

$$Pr(r) = nCr \times P^r(1 - P)^{(n-r)}$$

em que $P = 1/6$ e $n = 4$. As probabilidades de dois, três e quatro resultados iguais a seis são dadas por:

$$Pr(r = 2) = 4C2(1/6)^2(5/6)^2 = 0,116$$

$$Pr(r = 3) = 4C3(1/6)^3(5/6)^1 = 0,015$$

$$Pr(r = 4) = 4C4(1/6)^4(5/6)^0 = 0,00077$$

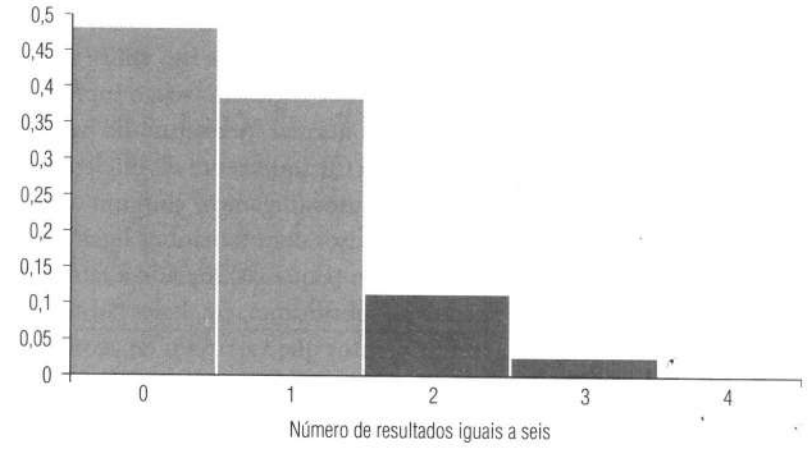
Como esses eventos são mutuamente exclusivos, as suas probabilidades podem ser simplesmente somadas para obter o resultado desejado, que é 0,132, ou 13,2%. Essa é a probabilidade de dois ou mais resultados iguais a seis em quatro lançamentos de um dado.

Esse resultado pode ser representado graficamente como parte da área sob a distribuição binomial apropriada, tal como na figura 3.4.

As áreas sombreadas indicam as probabilidades de dois ou mais resultados iguais a seis e, em conjunto, sua área corresponde a 13,2% de toda a distribuição. Isso ilustra um princípio importante: as probabilidades podem ser representadas por áreas sob uma distribuição de probabilidades apropriada. Veremos isso mais vezes adiante.

FIGURA 3.4

Probabilidade de dois ou mais resultados iguais a seis em quatro lançamentos de um dado



Exercício 3.1

a) A probabilidade de que um indivíduo escolhido ao acaso tenha olhos azuis é 0,6. Qual é a probabilidade de que quatro pessoas escolhidas ao acaso tenham todas olhos azuis?

b) Qual é a probabilidade de que duas pessoas na amostra de quatro tenham olhos azuis?

c) Nesse exemplo específico, escreva a fórmula binomial da probabilidade de r indivíduos com olhos azuis, para r = 0...4. Confirme que a soma das probabilidades é igual a 1.

Exercício 3.2

a) Calcule a média e a variância do número de pessoas com olhos azuis no exercício anterior.

b) Faça um gráfico dessa distribuição binomial e assinale o valor da média e o valor da média mais ou menos um desvio-padrão.

Tendo apresentado o conceito de distribuição de probabilidades usando a binomial, passamos agora à mais importante de todas as distribuições de probabilidades, a normal.

Distribuição normal

A distribuição binomial é apropriada quando um experimento tem dois resultados possíveis, mas nem todos os problemas pertencem a essa categoria. Por exemplo, o horário

(aleatório) de chegada de um trem é uma variável contínua e não pode ser analisado com a binomial. Há muitas distribuições de probabilidades em estatística, desenvolvidas para a análise de tipos distintos de problemas. Várias delas são abordadas neste livro, e a mais importante é a distribuição normal, à qual voltamos agora a nossa atenção. Ela foi descoberta pelo matemático alemão Gauss, no século XIX (e por isso também é conhecida como distribuição de Gauss), em seu trabalho sobre regressão (veja o capítulo 7).

Muitas variáveis aleatórias têm distribuição normal. A estatura de homens (e mulheres) tem distribuição normal, assim como o QI (quociente de inteligência). Outro exemplo é o de uma máquina que fabrica parafusos (digamos) com um comprimento-padrão de 5 cm, mas que de fato produz parafusos com tamanhos ligeiramente variáveis (essas diferenças provavelmente seriam bem pequenas), devido a fatores tais como desgaste da máquina, oscilações na pressão do lubrificante, etc. Esses fatores fariam com que se produzissem parafusos com comprimentos que variariam de acordo com a distribuição normal. Como esse tipo de processo é muito comum, a distribuição normal acaba ocorrendo freqüentemente em situações do dia-a-dia.

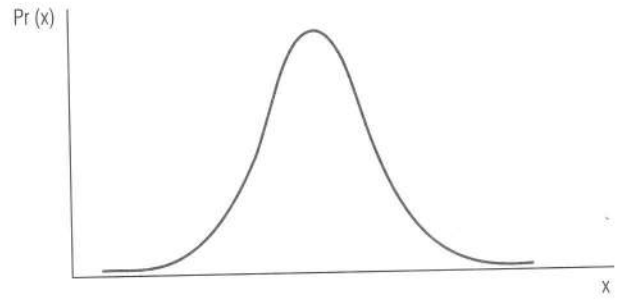
A distribuição normal tende a ocorrer quando uma variável aleatória resulta de muitas influências aleatórias independentes que são combinadas, e nenhuma delas domina as demais. A altura de um homem provém de muitas influências genéticas, bem como de fatores ambientais, como sua dieta, etc. Em conseqüência, a estatura tem distribuição normal. Se tomarmos a altura de homens e mulheres juntos, porém, o resultado não será uma distribuição normal. Isso se dá devido ao fato de que um fator domina o outro: gênero. Muitas variáveis conhecidas em economia não têm distribuição normal – renda é um exemplo, muito embora o logaritmo da renda seja aproximadamente normal. Mais para a frente, no momento adequado, passaremos algumas técnicas para lidar com tais circunstâncias.

Uma vez introduzida a idéia de distribuição normal, como seria a sua forma? Ela é apresentada a seguir em termos gráficos, e depois em termos matemáticos. Diferentemente da binomial, a distribuição normal é aplicável a variáveis aleatórias contínuas, como a estatura – uma distribuição normal típica é mostrada na figura 3.5. Como a distribuição normal é contínua, ela pode ser avaliada em todos os valores de x, e não somente em valores inteiros. A figura ilustra as principais características da distribuição:

- É unimodal, tendo um pico central. Se o gráfico tratasse da altura de homens, mostraria que a maioria se concentra em torno da estatura média, com alguns indivíduos muito altos e outros muito baixos.
- É simétrica, com as metades esquerda e direita parecendo espelhos uma da outra.
- Tem forma de sino.
- Estende-se continuamente por todos os valores de x, desde menos infinito a mais infinito, embora o valor de Pr(x) se torne muito pequeno quando nos aproximamos desses valores (como as páginas deste livro não são infinitamente largas, essa característica não é fielmente reproduzida aqui!). Isso também demonstra que muitas das distribuições empíricas (como a de estatura de homens) apenas se aproximam do ideal teórico, embora a aproximação seja suficientemente boa para fins práticos.

FIGURA 3.5

Distribuição normal



Em termos matemáticos, a fórmula da distribuição normal é (x é a variável aleatória):

$$(3.5)\quad Pr(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

A representação matemática não é tão formidável quanto parece: μ e σ são os parâmetros da distribuição, como n e P para a distribuição binomial (muito embora tenham significados diferentes); π é igual a 3,1416; e é igual a 2,7183. Caso a fórmula seja estimada com o uso de diferentes valores de x, os valores de Pr(x) obtidos mapearão uma distribuição normal. Felizmente, como veremos, não vamos precisar utilizar a fórmula na maioria dos problemas práticos.

Tal como a binomial, a normal é uma família de distribuições, variando de uma para a outra somente em termos dos valores dos parâmetros μ e σ. Diversas distribuições normais foram desenhadas na figura 3.6 para vários valores dos parâmetros.

FIGURA 3.6(a)

Distribuição normal, $\mu = 20$, $\sigma = 5$

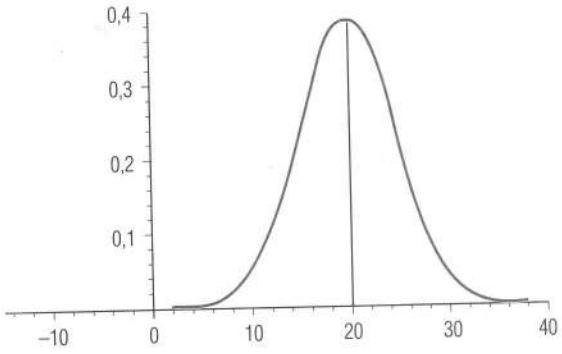


FIGURA 3.6(b)

Distribuição normal, $\mu = 15$, $\sigma = 2$

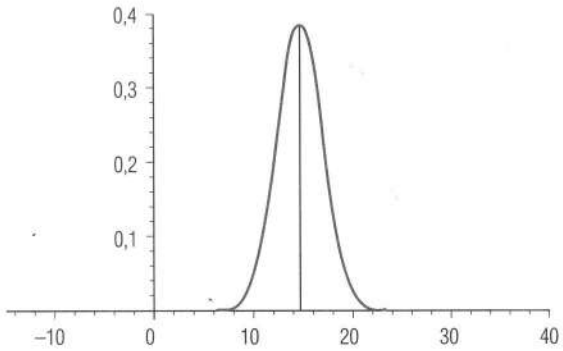
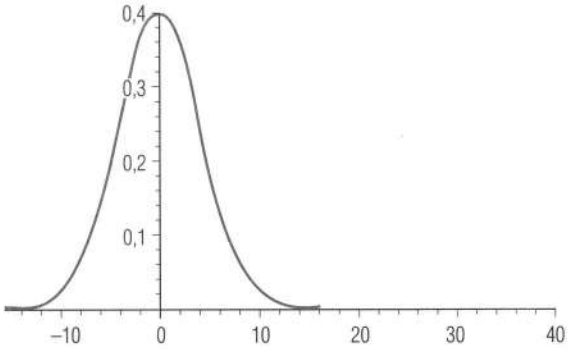


FIGURA 3.6(c)

Distribuição normal, $\mu = 0$, $\sigma = 4$



Qualquer que seja o valor de μ escolhido, ele será o centro da distribuição. Como a distribuição é simétrica, μ é a sua média. O efeito da variação de σ consiste em estreitar (quando σ é baixo) ou alargar (quando σ é alto) a distribuição – σ é o desvio-padrão da distribuição. A normal é outra família de distribuições, de dois parâmetros, como a binomial, e, uma vez que a média μ e o desvio-padrão σ (ou, o que seria equivalente, a variância, σ^2) sejam conhecidos, a distribuição inteira pode ser desenhada. A notação simplificada de uma distribuição normal é:

(3.6) $x \sim N(\mu, \sigma^2)$

o que significa que "a variável x tem distribuição normal com média μ e variância σ^2". Essa expressão é semelhante à da distribuição binomial, embora os significados dos parâmetros sejam diferentes.

O uso da distribuição normal pode ser ilustrado por meio de um exemplo simples. A altura de homens adultos tem distribuição normal com média $\mu = 174$ cm e desvio-padrão $\sigma = 9{,}6$ cm. Seja x a altura de homens adultos, então

$$(3.7)\quad x \sim N(174, 92{,}16)$$

como mostra a figura 3.7. Note que 3.7 contém a variância, e não o desvio-padrão.

Qual é a probabilidade de que um homem escolhido ao acaso tenha altura superior a 180 cm? Se todos os homens tivessem a mesma probabilidade de ser escolhidos, isso seria equivalente a perguntar que proporção deles tem altura superior a 180 cm. Isso é dado pela área sob a distribuição normal, à direita de $x = 180$, ou seja, a área sombreada na figura 3.7. Quanto mais distantes estamos da média de 174 cm, menor a área na cauda da distribuição. Uma maneira de descobrir essa área é usar a equação 3.5, o que exige, porém, o emprego de matemática sofisticada.

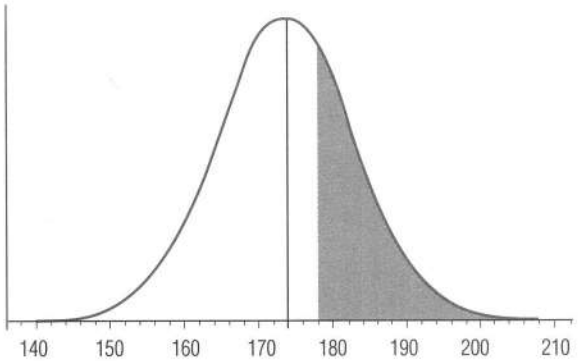
Como esse é um problema freqüente, as respostas já foram colocadas nas tabelas da **distribuição normal padronizada**. Entretanto, como há um número infinitamente grande de distribuições normais (uma para cada combinação de μ e σ^2), seria impossível tabular todas elas. A distribuição normal padronizada, que tem média igual a 0 e variância igual a 1, é usada, portanto, para representar todas as distribuições normais. Antes de se consultar a tabela, porém, os dados precisam ser transformados para que sejam compatíveis com a distribuição normal padronizada.

A transformação exigida é o cálculo do escore z, que foi apresentado no capítulo 1. Ele mede a distância entre o valor que nos interessa (180) e a média, medida em termos de desvios-padrão. Portanto, calculamos:

$$(3.8)\quad z = \frac{x - \mu}{\sigma}$$

FIGURA 3.7

Ilustração da distribuição de altura de homens



e z é uma variável aleatória normalmente distribuída, com média igual a 0 e variância igual a 1, ou seja, $z \sim N(0, 1)$. Essa transformação desloca a distribuição original μ unidades para a esquerda, e depois ajusta a dispersão dividindo por σ, o que resulta em média 0 e variância 1. z tem distribuição normal porque x também tem distribuição normal. A transformação em 3.8 retém a forma da distribuição normal, apesar das alterações de média e variância. Se x tivesse alguma outra distribuição, z também não teria distribuição normal.

É fácil verificar a média e a variância de z se usarmos as regras para os operadores E e V encontradas no capítulo 1:

$$E(z) = E\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sigma}(E(x)-\mu) = 0, \text{ pois } E(x) = \mu$$

$$V(z) = V\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sigma^2}V(x) = \frac{\sigma^2}{\sigma^2} = 1$$

Avaliando o escore z com os nossos dados, obtemos:

$$(3.9)\quad z = \frac{180-174}{9,6} = 0,63$$

Isso mostra que o valor 180 está 0,63 de um desvio-padrão acima da média, 174, da distribuição. Essa é uma medida de quão longe o valor 180 está de 174, e nos permite consultar a resposta em uma tabela. A tarefa consiste agora em encontrar a área sob a distribuição normal padronizada à direita de 0,63 de um desvio-padrão acima da média. A resposta pode ser lida diretamente na tabela da distribuição normal padronizada, incluída na tabela A.2 no apêndice deste livro. Um trecho dela é apresentado na tabela 3.2.

A coluna da esquerda fornece o escore z com uma casa decimal. A linha apropriada da tabela a ser consultada é a linha de z = 0,6. Para a segunda casa decimal (0,03), consultamos a coluna apropriada. Em sua interseção encontramos o valor 0,2643, que é a área desejada e, portanto, a probabilidade que buscamos. Em outras palavras, 26,43%

TABELA 3.2 Áreas da distribuição normal padronizada (extraídas da tabela A.2)

z	0,00	0,01	0,02	0,03	...	0,09
0,0	0,5000	0,4960	0,4920	0,4880	...	0,4641
0,1	0,4602	0,4562	0,4522	0,4483	...	0,4247
⋮	⋮	⋮	⋮	⋮	...	⋮
0,5	0,3085	0,3050	0,3015	0,2981	...	0,2776
0,6	0,2743	0,2709	0,2676	0,2643	...	0,2451
0,7	0,2420	0,2389	0,2358	0,2327	...	0,2148

da distribuição se situa à direita de 0,63 de um desvio-padrão acima da média. Portanto, 26,43% dos homens têm altura superior a 180 cm.

O uso da tabela de distribuição normal padronizada é possível porque, embora haja um número infinito de distribuições normais, fundamentalmente elas são todas iguais, de modo que a área à direita de 0,63 de um desvio-padrão acima da média é a mesma para todas. Se conseguirmos medir a distância em termos de desvios-padrão, podemos usar a tabela da normal padronizada. O processo de padronização transforma todas as distribuições normais em uma distribuição normal padronizada com média igual a 0 e variância igual a 1. Esse processo está representado na figura 3.8.

A área na cauda da direita é a mesma nas duas distribuições. Trata-se da distribuição normal padronizada, que aparece na figura 3.8(b), tabulada na tabela A.2. Para demonstrar que a padronização transforma todas as distribuições normais na normal padronizada, repetimos o problema anterior, agora usando todas as medidas em polegadas. A resposta, obviamente, deve ser a mesma. Como 1 polegada = 2,54 cm, os dados são:

$x = 70,87$ $\sigma = 3,78$ $\mu = 68,50$

Que proporção dos homens tem altura superior a 70,87 polegadas? A distribuição normal apropriada agora é:

(3.10) $x \sim N(68,50, 3,78^2)$

O escore z é:

$$(3.11)\quad z = \frac{70,87 - 68,50}{3,78} = 0,63$$

ou seja, o mesmo escore z obtido anteriormente, o que nos dá a mesma probabilidade.

FIGURA 3.8(a)

Distribuição normal

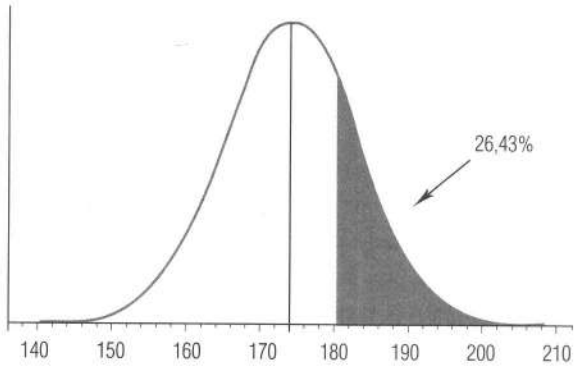
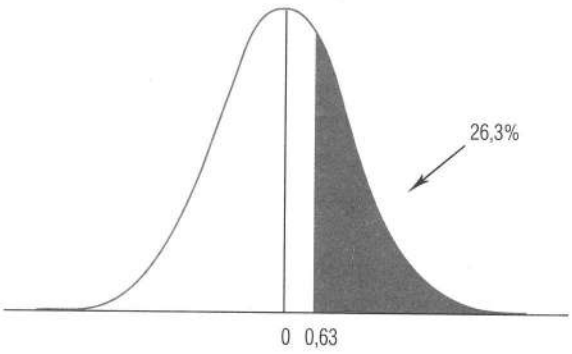


FIGURA 3.8(b)

Distribuição normal padronizada correspondente à figura 3.8(a)



Exemplo resolvido 3.2

As embalagens de cereal matinal têm peso nominal de 750 g, mas há certa variação em torno disso, pois as máquinas que colocam cereal nas caixas são imperfeitas. Vamos supor que os pesos obedeçam a uma distribuição normal, e que o desvio-padrão em torno da média de 750 seja igual a 5 g. Que proporção de caixas de cereal pesa mais de 760 g?

Resumindo nossas informações, temos $x \sim N(750, 25)$, em que x representa o peso. Desejamos encontrar $Pr(x > 760)$. Para descobrir a resposta, precisamos medir a distância entre 760 e 750 em termos de desvios-padrão. Esse cálculo é:

$$z = \frac{760 - 750}{5} = 2,0$$

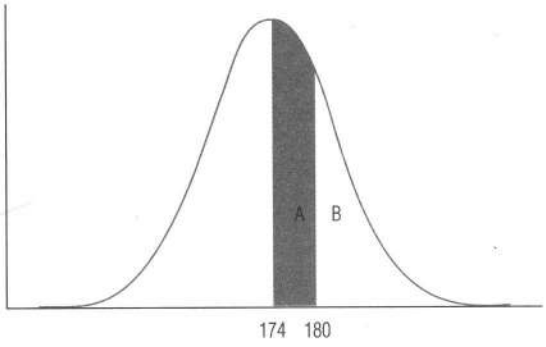
Procurando z = 2,0 na tabela A.2, vemos uma área de 0,0228 na cauda da distribuição. Portanto, 2,28% das caixas pesam mais de 760 g.

Como as tabelas da distribuição normal padronizada são muito utilizadas, vale a pena examinar mais alguns exemplos para reforçar o método. Até agora, calculamos que $Pr(z > 0,63) = 0,2643$. Como a área total sob o gráfico da distribuição é igual a 1 (isto é, a soma das probabilidades deve ser 1), a área à esquerda de z = 0,63 deve ser igual a 0,7357, ou seja, 73,57% dos homens têm altura inferior a 180 cm. É muito fácil trabalhar com áreas sob o gráfico da distribuição para chegar a qualquer área que se queira. Por exemplo, qual é a proporção de homens com altura entre 174 e 180 cm? Recorrer à figura 3.9 é de grande utilidade para isso.

Pede-se o valor da área A. A área B já foi determinada, e é igual a 0,2643. Como a distribuição é simétrica, a área A + B deve ser igual a 0,5, pois 174 está no centro (média) da distribuição. A área A, portanto, é igual a 0,5 – 0,2643 = 0,2357. O resultado desejado é 23,57%.

FIGURA 3.9

Proporção de homens com altura entre 174 cm e 180 cm



Como exercício final, considere a questão da proporção de homens com altura entre 166 e 178 cm. Como mostrado na figura 3.10, deseja-se a área C + D. A única maneira de obtê-la é calculando as duas áreas separadamente e depois somando-as. No caso da área D, o escore z correspondente a 178 é:

$$(3.12)\quad z_D = \frac{178-174}{9,6} = 0,42$$

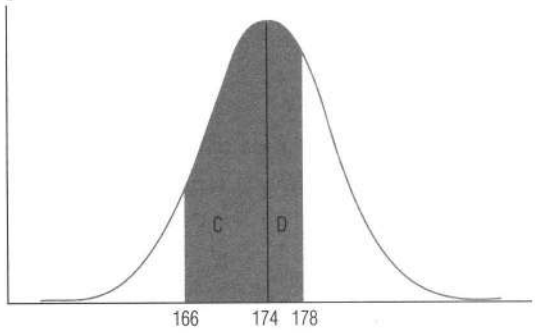
A tabela A.2 indica que a área na cauda da direita, além de z = 0,42, é igual a 0,3372, de modo que a área D vale 0,5 – 0,3372 = 0,1628. No caso da área C, o escore z é

$$(3.13)\quad z_C = \frac{166-174}{9,6} = -0,83$$

O sinal negativo indica que a área que está sendo considerada situa-se na cauda esquerda da distribuição, abaixo da média. Como a distribuição é simétrica, ela é idênti-

FIGURA 3.10

Proporção de homens com altura entre 166 cm e 178 cm



Uso de *software* para encontrar áreas sob a curva da distribuição normal padronizada

Se você utiliza uma planilha eletrônica, pode obter a distribuição de z diretamente, dispensando o uso de tabelas. No Excel, por exemplo, a função "=DIST.NORMP(0,63)" fornece a resposta 0,7357, ou seja, a área à esquerda do escore z. A área na cauda direita, portanto, é obtida subtraindo esse valor de 1, ou seja, 1 – 0,7357 = 0,2643. Portanto, ao escrever a fórmula "=1 – DIST.NORMP(0,63)" numa célula você obterá diretamente a área na cauda direita.

ca à que estaria na cauda direita, de modo que ao consultar a tabela o sinal negativo pode ser ignorado. Procurando-se z = 0,83 na tabela A.2, obtém-se uma área de 0,2033 na cauda, e assim a área C é igual a 0,5 – 0,2033 = 0,2967. Somando-se as áreas C e D obtém-se 0,1628 + 0,2967 = 0,4595. Portanto, quase metade dos homens tem altura entre 166 e 178 cm.

Uma interpretação alternativa dos resultados obtidos acima é que, se um homem for escolhido ao acaso na população adulta, a probabilidade de que tenha altura superior a 180 cm é 26,43%. Isso está de acordo com a visão freqüentista. Como 26,43% da população tem altura superior a 180 cm, então essa é a probabilidade de que um homem escolhido aleatoriamente tenha mais de 180 cm de altura.

Exercício 3.3

a) A variável x tem distribuição normal, com $x \sim N(40, 36)$. Determine a probabilidade de que $x > 50$.
b) Encontre $Pr(x < 45)$.
c) Encontre $Pr(36 < x < 44)$.

Exercício 3.4

A média mais ou menos 0,67 desvios-padrão corta 25% em cada cauda da distribuição normal. Portanto, os 50% intermediários da distribuição ficam a no máximo mais ou menos 0,67 desvio-padrão da média. Use esse fato para calcular o intervalo entre quartis da distribuição $x \sim N(200, 256)$.

Exercício 3.5

Tal como sugerido no texto, o logaritmo da renda tem distribuição aproximadamente normal. Supondo que o log (na base 10) da renda tenha a distribuição $x \sim N(4{,}18, 2{,}56)$, calcule o intervalo entre quartis para x e depois use antilogs para encontrar o intervalo entre quartis da renda.

Média amostral como variável com distribuição normal

Um dos conceitos mais importantes em inferência estatística é a distribuição de probabilidades da média de uma amostra ao acaso, pois é comum usá-la para dizer algo a respeito da população a ela associada. Suponhamos que, da população de homens adultos, uma amostra de tamanho n = 36 é tirada ao acaso; a altura de todos é medida, e calcula-se a estatura média na amostra. O que podemos inferir, a partir disso, sobre a verdadeira altura média da população? Para fazer isso, precisamos conhecer as propriedades estatísticas da média amostral. Esta é uma variável aleatória, pois há o elemento probabilístico associado ao processo de amostragem ao acaso (amostras diferentes produziriam valores distintos para a média amostral). Como a média amostral é uma variável aleatória, deve ter uma distribuição de probabilidades.

Portanto, precisamos conhecer, primeiro, qual é a distribuição apropriada e, em segundo lugar, quais são seus parâmetros. Partindo da definição da média amostral, temos:

$$(3.14)\quad \bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n)$$

em que cada observação, x_i, é por si mesma uma variável aleatória com distribuição normal, com $x_i \sim N(\mu, \sigma^2)$, pois cada variável provém da distribuição básica, que tem tais características. (Afirmamos anteriormente que a altura dos homens tem distribuição normal.) Faremos uso agora do seguinte teorema para demonstrar que x tem distribuição normal:

> **Teorema** Qualquer combinação linear de variáveis aleatórias independentes, com distribuição normal, também tem distribuição normal.

Uma combinação linear de duas variáveis x_1 e x_2 é da forma $w_1x_1 + w_2x_2$, em que w_1 e w_2 são constantes. Isso pode ser generalizado para qualquer número de valores de x. Desde que as observações sejam independentes, portanto, a média amostral tem distribuição normal. Nesse caso, o peso w de cada observação é igual a 1/n.

Precisamos agora determinar os parâmetros (média e variância) da distribuição. Para isso, usamos mais uma vez os operadores E e V:

$$(3.15)\quad E(\bar{x}) = \frac{1}{n}(E(x_1) + E(x_2) + \ldots + E(x_n)) = \frac{1}{n}(\mu + \mu + \ldots + \mu) = \frac{1}{n}n\mu = \mu$$

$$(3.16)\quad V(\bar{x}) = V\left(\frac{1}{n}[x_1 + x_2 + \ldots + x_n]\right) = \frac{1}{n^2}(V(x_1) + V(x_2) + \ldots + V(x_n))$$

$$= \frac{1}{n^2}(\sigma^2 + \sigma^2 + \ldots + \sigma^2) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

Reunindo tudo isso, temos:[2]

$$(3.17)\quad \bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

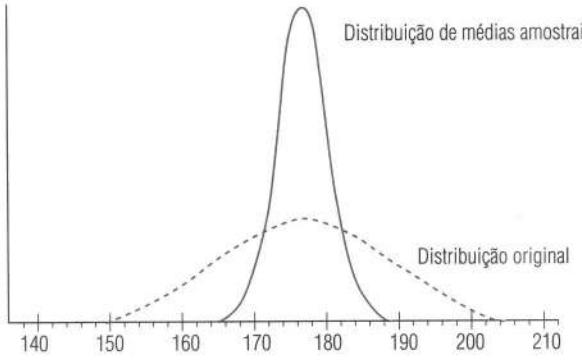
Isso pode ser resumido no seguinte teorema:

Teorema A média amostral, $\bar{x}$, extraída de uma população normalmente distribuída com média μ e variância σ^2, tem uma distribuição amostral normal, com média μ e variância σ^2/n, em que n é o tamanho da amostra.

O significado desse teorema é o seguinte. Em primeiro lugar, supõe-se que a população da qual as amostras são extraídas também tem distribuição normal (mas essa hipótese deixará de ser feita daqui a pouco), com média μ e variância σ^2. Dessa população, muitas amostras são extraídas, cada uma delas de tamanho n, e calcula-se a média de cada amostra. As amostras são independentes, o que quer dizer que as observações escolhidas em uma amostra não influenciam a seleção de observações nas outras amostras. Isso produz muitas médias amostrais, como $\bar{x}_1$, $\bar{x}_2$, etc. Se essas médias amostrais forem tratadas como um novo conjunto de observações, então é possível obter a distribuição de probabilidades dessas observações. O teorema diz que essa distribuição é normal, com as médias amostrais centradas em torno da média μ, a média da população, e com variância σ^2/n. O argumento é representado graficamente na figura 3.11.

FIGURA 3.11

Distribuição original e distribuição de médias amostrais



Nota: A distribuição de $\bar{x}$ é extraída de uma amostra de tamanho n = 9. Uma amostra maior estreitaria a distribuição de $\bar{x}$; uma amostra menor a alargaria.

2. Não se preocupe se não conseguiu acompanhar o desenvolvimento dessa fórmula, mas aceite o resultado porque é correto.

Intuitivamente, esse teorema pode ser compreendido da seguinte maneira: se a altura de homens adultos for uma variável aleatória com distribuição normal, com média $\mu = 174$ cm e variância $\sigma^2 = 92{,}16$, então seria esperado que uma amostra de (digamos) nove homens adultos gerasse uma altura média amostral de 174 cm, talvez pouco mais, ou pouco menos. Em outras palavras, a média amostral está centrada em torno de 174 cm, ou seja, a média da distribuição de médias amostrais é 174 cm.

Quanto maior fosse o tamanho das amostras individuais (ou seja, quanto maior o valor de n), mais próxima a média amostral tenderia a estar de 174 cm. Por exemplo, se o tamanho da amostra fosse igual a apenas dois, uma amostra de duas pessoas muito altas seria bastante possível, resultando numa média amostral elevada, bem superior a 174 cm; 182 cm, por exemplo. Mas se o tamanho da amostra fosse igual a 20, seria pouco provável que 20 homens adultos muito altos fossem selecionados, e a média da amostra tenderia a estar muito mais próxima de 174. É por isso que o tamanho n da amostra aparece na fórmula da variância da distribuição da média amostral, σ^2/n.

Note, mais uma vez, que transformamos uma ou mais variáveis aleatórias, os x_i, com uma particular distribuição de probabilidades, em outra variável aleatória, $\bar{x}$, com uma distribuição ligeiramente diferente. Essa é uma prática comum em estatística: transformar uma variável geralmente faz com que ela fique num formato mais útil, por exemplo, um em que a distribuição de probabilidades seja bem conhecida.

O teorema acima pode ser utilizado para resolver uma variedade de problemas estatísticos. Por exemplo, qual é a probabilidade de que uma amostra obtida ao acaso de nove homens adultos tenha uma altura média superior a 180 cm? Sabe-se que a estatura de todos os homens adultos tem distribuição normal com média $\mu = 174$ cm e variância $\sigma^2 = 92{,}16$. O teorema pode ser utilizado para determinar a distribuição de probabilidades da média amostral. Para a população, temos:

$x \sim N(\mu, \sigma^2)$, isto é, $x \sim N(174, 92{,}16)$

Portanto, para a média amostral:

$\bar{x} \sim N(\mu, \sigma^2/n)$, isto é, $\bar{x} \sim N(174, 92{,}16/9)$

Isso é representado graficamente na figura 3.12.

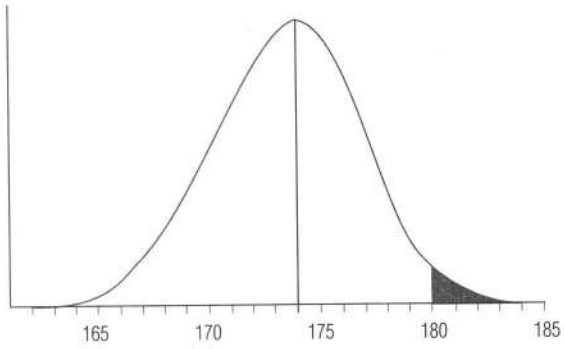
Para responder a pergunta proposta, é preciso determinar a área sombreada à direita de 180 na figura 3.12. Esse já deve ser, a esta altura, um problema conhecido. Primeiro, calculamos o escore z:

$$(3.18)\quad z = \frac{\bar{x} - \mu}{\sqrt{\sigma^2/n}} = \frac{180 - 174}{\sqrt{92{,}16/9}} = 1{,}88$$

Note que a fórmula do escore z é ligeiramente diferente, porque estamos lidando com a média amostral $\bar{x}$, e não com o próprio x. No numerador, usamos $\bar{x}$ em lugar

FIGURA 3.12

Proporção de médias amostrais superiores a $\bar{x} = 180$



de x, e no denominador usamos σ^2/n, e não σ^2. Isso porque $\bar{x}$ tem variância σ^2/n, e não σ^2, que é a variância da população. $\sqrt{\sigma^2/n}$ é conhecido como **erro-padrão**, para diferenciá-lo de σ, que é o desvio-padrão da população. Entretanto, o princípio por trás do escore z é o mesmo: ele mede a distância entre a média amostral de 180 e a média da população de 174, em termos do número de desvios-padrão.

Ao consultar o valor de z = 1,88 na tabela A.2, encontramos uma área de 0,0311 na cauda da direita da distribuição normal. Portanto, 3,11% das médias amostrais serão maiores ou iguais a 180 cm quando o tamanho da amostra for igual a nove. A probabilidade desejada, então, é igual a 3,11%.

Como essa probabilidade é bastante pequena, poderíamos considerar os motivos pelos quais isso ocorre. Há duas possibilidades:

1. por azar, a amostra coletada não é muito representativa da população como um todo, ou;
2. a amostra é representativa da população, mas a média da população não é realmente igual a 174 cm.

Somente uma dessas possibilidades pode estar correta. A maneira de decidir entre elas será analisada mais adiante, no capítulo 5, que trata de testes de hipóteses.

É interessante examinar a diferença entre a resposta obtida para uma amostra de tamanho nove (3,11%) e a obtida antes para um único indivíduo (26,43%). Este último caso pode ser considerado como uma amostra de uma única observação extraída da população. Os exemplos ilustram o fato de que, quanto maior o tamanho da amostra, mais próxima a média amostral está da média da população. Portanto, amostras maiores tendem a fornecer melhores estimativas da média da população.

Amostragem de uma população não-normal

O teorema e os exemplos anteriores se apoiaram no fato de que a população obedecia a uma distribuição normal. Mas o que acontece quando isso não é verdade?

Afinal de contas, não se sabe com certeza se a altura de todos os homens adultos tem, exatamente, uma distribuição normal, e há muitas populações que não possuem essa distribuição (por exemplo, a distribuição de riqueza, como vimos no capítulo 1). O que se pode fazer nessas circunstâncias? A resposta consiste em usar um outro teorema para a distribuição de médias amostrais, apresentado aqui sem demonstração. Ele é conhecido como **Teorema do Limite Central**

> **Teorema do Limite Central** A média amostral $\bar{x}$, extraída de uma população com média μ e variância σ^2, tem uma distribuição amostral que se aproxima de uma distribuição normal, com média μ e variância σ^2/n, à medida que o tamanho da amostra tende a infinito.

Esse teorema é de grande utilidade, pois elimina a suposição de que a população se distribui normalmente. Note que a distribuição de médias amostrais é normal somente quando o tamanho da amostra é infinitamente grande; para qualquer tamanho de amostra finito, a distribuição é apenas aproximadamente normal. Entretanto, a aproximação será suficientemente boa para fins práticos quando o tamanho da amostra for de pelo menos 25 observações. Se a própria população tiver distribuição aproximadamente normal, uma amostra menor seria suficiente. Se a distribuição populacional for particularmente assimétrica, seria desejável contar com mais de 25 observações. Esse número representa uma regra prática adequada à maioria das circunstâncias. Trata-se de mais um exemplo de como a estatística é uma ciência inexata: ela não fornece respostas absolutamente claras, mas nos ajuda a chegar a conclusões razoáveis.

Como exemplo do uso do Teorema do Limite Central, voltemos aos dados de riqueza do capítulo 1. Lembre-se que o nível médio de riqueza era igual a 131,443 (em milhares), e que a variância era 52.880. Suponhamos que n = 50 pessoas fossem escolhidas ao acaso nessa população. Qual é a probabilidade de que a média da amostra seja maior do que 150 (isto é, £150.000)?

Nesse caso, sabemos que a distribuição de origem é fortemente assimétrica, de modo que é importante saber que contamos com 50 observações. Isso deve ser suficiente para justificar a aplicação do Teorema do Limite Central. Portanto, a distribuição de $\bar{x}$ é

$$(3.19)\ \bar{x} \sim N(\mu, \sigma^2/n)$$

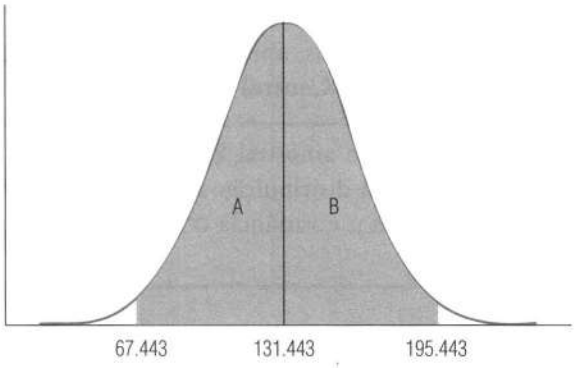
e, inserindo os valores dos parâmetros, obtemos[3]

$$(3.20)\ \bar{x} \sim N(131{,}443,\ 52.880/50)$$

3. Note que se tivéssemos usado 131.443 para a média, teríamos 52.880.000.000 para a variância. O uso de milhares de libras como unidade faz com que seja mais fácil trabalhar com os números. O escore z é o mesmo nos dois casos.

FIGURA 3.13

Probabilidade de $\bar{x}$ ficar a menos de £64.000 de cada lado de £131.443



Para encontrar a área além de uma média amostral de 150, calcula-se primeiramente o escore z:

$$(3.21)\quad z = \frac{150 - 131,443}{\sqrt{52.880/50}} = 0,57$$

Consultando as tabelas da distribuição normal padronizada, vemos que a área na cauda é 28,43%. Essa é a probabilidade desejada. Portanto, há uma probabilidade de 28,43% de encontrar uma média de pelo menos £150.000 com uma amostra de 50 observações. Isso demonstra que existe uma probabilidade razoavelmente elevada de se obter uma média amostral relativamente distante de £131.443. Isso decorre do elevado grau de dispersão na distribuição de riqueza.

Estendendo um pouco esse exemplo, podemos perguntar qual é a probabilidade de que a média da amostra fique, digamos, a £64.000 de cada lado do valor médio de £131.443 (isto é, entre 67.443 e 195.443). A figura 3.13 ilustra a situação, com a área desejada sombreada. Por simetria, as áreas A e B devem ser iguais, o que significa que precisamos calcular apenas uma delas. Para B, calculamos o escore z:

$$(3.22)\quad z = \frac{195,443 - 131,443}{\sqrt{52.880/50}} = 1,968$$

Com base na tabela da normal padronizada, deixa-se aproximadamente 2,5% na cauda superior, de modo que a área B é igual a 0,475. Portanto, as áreas A e B, juntas, representam 95% da distribuição. Existe assim uma probabilidade de 95% de que a média amostral fique no intervalo [67.443, 195.443], que chamamos de **intervalo de probabilidade** de 95% para a média amostral. Escrevemos o intervalo da seguinte maneira:

(3.23) $\Pr(67.443 \leq \bar{x} \leq 195.443) = 0,95$

ou, em termos das fórmulas que temos utilizado:[4]

(3.24) $\Pr(\mu - 1,96\sqrt{\sigma^2/n} \leq \bar{x} \leq \mu + 1,96\sqrt{\sigma^2/n}) = 0,95$

O intervalo de probabilidade de 95% e o conceito correspondente de intervalo de confiança de 95% (apresentado no capítulo 4) desempenham papéis importantes em inferência estatística. Montamos deliberadamente o exemplo anterior para chegar a uma resposta de 95% por esse motivo.

Exercício 3.6

a) Se x tem distribuição $x \sim N(50, 64)$, e são obtidas amostras com $n = 25$, qual é a distribuição de $\bar{x}$?
b) Caso o tamanho da amostra dobre para 50, como isso alterará o erro-padrão de $\bar{x}$?
c) Sendo o tamanho de amostra igual a 25, (1) qual é a probabilidade de $\bar{x} > 51$? (2) Quanto é $\Pr(\bar{x} < 48)$? (3) Quanto é $\Pr(49 < \bar{x} < 50,5)$?

Relação entre as distribuições binomial e normal

Muitas distribuições estatísticas são relacionadas umas às outras de alguma forma. Isso significa que muitos problemas podem ser resolvidos por uma variedade de métodos distintos (usando-se distribuições diferentes), muito embora geralmente uma seja mais conveniente do que as outras. Esse aspecto pode ser ilustrado examinando-se a relação entre as distribuições binomial e normal.

Lembre o experimento do lançamento repetido de uma moeda, registrando-se o número de caras. Dissemos anteriormente que isso poderia ser analisado com a distribuição binomial. Note, porém, que o número de caras, uma variável aleatória, é influenciado por muitos eventos independentes (os lançamentos individuais) combinados. Além do mais, cada um dos lançamentos tem o mesmo peso, ou seja, nenhum deles domina os outros. Essas são exatamente as condições sob as quais ocorre uma distribuição normal, de modo que parece haver uma ligação entre as duas distribuições.

Essa idéia é correta. E, como já vimos, se uma variável aleatória r obedece a uma distribuição binomial, então:

(3.1) $r \sim B(n, P)$

a média da distribuição é nP e a sua variância, $nP(1 - P)$. Ocorre que, à medida que n se eleva, a distribuição binomial se aproxima da distribuição normal com média nP e variância

4. 1,96 é o valor exato que deixa 2,5% em cada cauda.

nP(1 – P). Essa aproximação é suficientemente precisa desde que nP > 5 e n(1 – P) > 5, o que significa que a aproximação pode não ser muito boa (mesmo para valores elevados de n) se P for um valor muito próximo de zero ou de um. No caso do experimento de lançamento de uma moeda, em que P = 0,5, dez lançamentos seriam o bastante. Note que essa aproximação é suficientemente boa com apenas dez observações, muito embora a distribuição subjacente de probabilidades em nada se assemelhe a uma distribuição normal.

Para demonstrar, resolvemos o problema a seguir usando a distribuição binomial e a distribuição normal. Quarenta estudantes fazem uma prova de estatística que é avaliada somente com duas notas: aprovado/reprovado. Se a probabilidade, P, de um estudante passar é 60%, qual é a probabilidade de que pelo menos 30 estudantes passem nessa prova?

Os dados amostrais são:

$$P = 0,6$$
$$1 - P = 0,4$$
$$n = 40$$

Método da distribuição binomial

Para resolver o problema usando a distribuição binomial, é necessário encontrar a probabilidade de que exatamente 30 estudantes sejam aprovados, mais a probabilidade de que 31 passem, de que 32 passem, etc., até a probabilidade de que 40 passem na prova (o fato de os eventos serem mutuamente exclusivos permite o uso desse procedimento). A probabilidade de que 30 estudantes sejam aprovados é:

$$\begin{aligned} Pr(r = 30) &= nCr \times P^r(1 - P)^{n-r} \\ &= 40C^{30} \times 0,6^{30} \times 0,4^{10} \\ &= 0,020 \end{aligned}$$

(Observação: esse cálculo pressupõe que as probabilidades sejam independentes, ou seja; que não haja "cola"!) Em si, este já é um cálculo bastante tedioso, mas Pr(31), Pr(32), etc. ainda precisam ser determinadas. Calculando e somando essas probabilidades, temos o resultado 3,52%, ou seja, a probabilidade de que pelo menos 30 estudantes sejam aprovados. (Fazer os cálculos seria um exercício útil para o leitor, pelo menos para ter uma noção de quanto tempo isso leva.)

Método da distribuição normal

Como afirmado acima, a distribuição binomial pode ser aproximada por uma distribuição normal com média nP e variância nP(1 – P). Nesse caso, nP é igual a 24 (40 × 0,6) e n(1 – P) é igual a 16, ambos maiores do que cinco, de modo que a aproximação pode ser usada com segurança. Portanto:

$$r \sim N(nP, nP(1 - P))$$

Inserindo-se os valores dos parâmetros, obtém-se:

$$r \sim N(24, 9,6)$$

Os métodos usuais são empregados a seguir para encontrar a área apropriada sob a distribuição. Entretanto, antes de se fazê-lo, é preciso realizar um ajuste (isso só ocorre quando se aproxima a distribuição binomial pela distribuição normal). A distribuição normal é contínua, ao passo que a distribuição binomial é discreta. Portanto, o valor 30 na distribuição binomial é representado pela área sob a distribuição normal entre 29,5 e 30,5; 31, pela área entre 30,5 e 31,5, etc. Portanto, é a área sob a distribuição normal à direita de 29,5, e não de 30, que deve ser calculada. Isso é chamado de **correção de continuidade**. O cálculo do escore z é:

$$(3.25)\quad z = \frac{29,5 - 24}{\sqrt{9,6}} = 1,78$$

Isso nos dá uma área de 3,75%, não muito distante da resposta correta, calculada pela distribuição binomial. O tempo economizado e a facilidade de cálculo parecem compensar a pequena perda de precisão.

Outros exemplos podem ser montados para testar esse método com valores diferentes de P e n. Valores pequenos de n, ou valores de nP ou n(1 – P) inferiores a 5, fornecerão resultados inadequados, isto é, a aproximação normal à binomial não será muito boa.

Exercício 3.7

a) Uma moeda é lançada 20 vezes. Qual é a probabilidade de ocorrência de mais de 14 caras? Faça o cálculo usando tanto a distribuição binomial quanto a normal, e compare os resultados.
b) Uma moeda enviesada, para a qual Pr(cara) = 0,7, é lançada seis vezes. Qual é a probabilidade de ocorrência de mais de quatro caras? Compare os métodos binomial e normal nesse caso. Quão precisa é a aproximação normal?
c) Repita o item (b), mas para mais de cinco caras.

Distribuição de Poisson

A seção anterior mostrou como a distribuição binomial pode ser aproximada por uma distribuição normal sob determinadas circunstâncias. A aproximação não funciona particularmente bem para valores muito pequenos de P, quando nP é inferior a 5. Nessas condições, a binomial pode ser aproximada pela distribuição de Poisson, que é dada pela fórmula:

$$(3.26)\quad Pr(x) = \frac{\mu^x e^{-\mu}}{x!}$$

em que μ é a média da distribuição (comso μ na distribuição normal e nP na binomial). Tal como esta última, mas diferentemente da normal, a distribuição de Poisson é uma

distribuição discreta de probabilidades, de modo que a equação 3.26 é definida apenas para valores inteiros de x. Além disso, é aplicável a uma série de tentativas independentes, como no caso da binomial.

O uso da distribuição binomial mostra-se apropriado quando a probabilidade de "sucesso" é muito pequena e o número de tentativas, grande. Demonstra-se seu emprego no exemplo a seguir. Um fabricante oferece uma garantia de dois anos para a tela de TV que produz. Com base na experiência passada, ele sabe que 0,5% de seus tubos serão defeituosos e o problema ocorrerá durante o prazo da garantia. Qual é a probabilidade de que, num lote de 500 tubos, (a) nenhum tenha defeito, (b) mais de três apresentem defeitos?

A média da distribuição de Poisson, nesse caso, é $\mu = 2,5$ (0,5% de 500). Portanto:

$$(3.27)\quad Pr(x = 0) = \frac{2,5^0 e^{-2,5}}{0!} = 0,082$$

ou seja, a probabilidade de não ocorrer defeito algum é 8,2%. A resposta desse problema, com o uso do método binomial, é:

$$Pr(r = 0) = 0,995^{500} = 0,0816$$

Portanto, o método Poisson fornece uma resposta razoavelmente precisa. A aproximação Poisson à binomial é satisfatória se nP for inferior a 7.

A probabilidade de que mais de três tubos falhem é calculada da seguinte maneira:

$$Pr(x > 3) = 1 - Pr(x = 0) - Pr(x = 1) - Pr(x = 2) - Pr(x = 3)$$

$$Pr(x = 1) = \frac{2,5^1 e^{-2,5}}{1!} = 0,205$$

$$Pr(x = 2) = \frac{2,5^2 e^{-2,5}}{2!} = 0,256$$

$$Pr(x = 3) = \frac{2,5^3 e^{-2,5}}{3!} = 0,214$$

$$\text{Portanto, } Pr(x > 3) = 1 - 0,082 - 0,205 - 0,256 - 0,214 = 0,242$$

Assim sendo, há uma probabilidade de aproximadamente 24% de ocorrência de mais de três falhas. O cálculo binomial é muito mais trabalhoso, mas fornece 24,2% como resposta.

A distribuição de Poisson também é utilizada em problemas nos quais os eventos ocorrem com o passar do tempo, como a marcação de gols num jogo de futebol (veja o problema 3.25), ou em problemas relativos a filas (por exemplo, chegadas em um caixa automático). Nessas questões, não há um "número" natural de tentativas, mas está claro que, se considerarmos um período curto, a probabilidade de ocorrência de um evento é pequena. Podemos, portanto, levar em conta o número de tentativas como sendo idên-

tico ao número de períodos. Tal fato é ilustrado pelo seguinte exemplo: um time de futebol marca, em média, dois gols por jogo (você pode usar como exemplo seu time favorito e o desempenho dele!). Qual é a probabilidade de que o time marque zero ou um gol durante um jogo?

A média da distribuição é igual a 2; temos, assim, usando a distribuição de Poisson:

$$\Pr(x = 0) = \frac{2^0 e^{-2}}{0!} = 0{,}135$$

$$\Pr(x = 1) = \frac{2^1 e^{-2}}{1!} = 0{,}271$$

Você deve continuar calculando as probabilidades até 2 ou mais gols, para verificar se a soma das probabilidades é igual a 1.

Um exemplo de problema relativo a filas pode ser assim ilustrado: se uma loja recebe, em média, 20 clientes por hora, qual é a probabilidade de que nenhum cliente apareça, num período de cinco minutos, justamente quando o proprietário estiver tomando um cafezinho?

O número médio de clientes por período de cinco minutos é 20 × 5/60 = 1,67. A probabilidade de um período livre de cinco minutos, portanto, é igual a:

$$\Pr(x = 0) = \frac{1{,}67^0 e^{-1{,}67}}{0!} = 0{,}189$$

ou seja, cerca de 19%. Note que esse problema não poderia ser resolvido pelo método binomial, pois não conhecemos os valores de n e P separadamente, mas apenas o seu produto.

Exercício 3.8

a) A probabilidade de ganhar um prêmio numa loteria é igual a 1 em 50. Se você comprar 50 bilhetes, qual é a probabilidade de que (1) zero bilhete seja premiado, (2) um bilhete seja premiado, (3) dois bilhetes sejam premiados? (4) Qual é a probabilidade de que pelo menos um seja premiado?

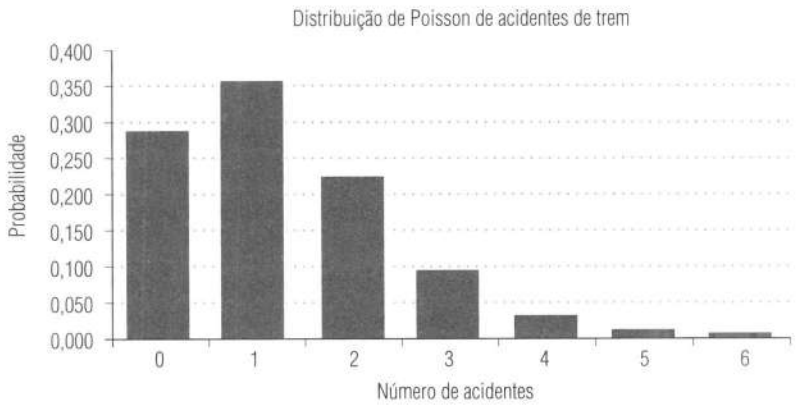
b) Em média, uma pessoa compra um bilhete de loteria num supermercado a cada cinco minutos. Qual é a probabilidade de que se passem dez minutos sem que haja um comprador sequer?

Acidentes de trem

Andrew Evans, do University College, em Londres, usou a distribuição de Poisson para examinar o número de acidentes fatais de trem na Grã-Bretanha entre 1967 e 1997. Como esses acidentes são raros, felizmente, a probabilidade de um acidente em qualquer período é muito pequena e, portanto, o uso da distribuição de Poisson se justifica.

Ele constatou que o número médio de acidentes tem caído com o tempo e, em 1997, havia chegado a 1,25 por ano. Esse número é usado, portanto, como média μ da distribuição de Poisson, e podemos calcular as probabilidades de 0, 1, 2, etc. acidentes por ano. Usando μ = 1,25 e inserindo esse valor na equação 3.26, obtemos a seguinte tabela:

Número de acidentes	0	1	2	3	4	5	6
Probabilidade	0,287	0,358	0,224	0,093	0,029	0,007	0,002



Assim sendo, o resultado mais provável é um acidente fatal por ano; qualquer número acima de quatro é extremamente improvável. Na realidade, Evans descobriu que a distribuição de Poisson não se ajustava com perfeição aos dados: a variação efetiva foi inferior à prevista pelo modelo.

Fonte: A. W. Evans, "Fatal train accidents on Britain's mainline railways". *J. Royal Statistical Society*, Series A, v. 163 (1), 2000.

Resumo

- O comportamento de muitas variáveis aleatórias (por exemplo, o resultado do lançamento de uma moeda) pode ser descrito por uma distribuição de probabilidades (nesse caso, a distribuição binomial).
- A distribuição binomial é apropriada para problemas nos quais há somente dois resultados possíveis para um evento aleatório (por exemplo, cara/coroa, sucesso/insucesso), e quando a probabilidade de sucesso é a mesma a cada vez que o experimento é realizado.
- A distribuição normal é apropriada para problemas em que a variável aleatória tem a conhecida distribuição em forma de sino. É comum isso ocorrer quando a variável

é influenciada por muitos fatores independentes, e nenhum deles domina os demais. Um exemplo é o da altura de homens, que tem distribuição normal.

- A distribuição de Poisson é usada em circunstâncias nas quais há uma probabilidade muito baixa de "sucesso" e um número elevado de tentativas.
- Cada uma dessas distribuições é, na realidade, uma família de distribuição, diferindo em termos de parâmetros. Tanto a distribuição binomial quanto a normal têm dois parâmetros: n e P, no primeiro caso; μ e σ^2, no segundo. A distribuição de Poisson tem um parâmetro, sua média μ.
- A média de uma amostra ao acaso obedece a uma distribuição normal porque é influenciada por muitos fatores independentes (as observações da amostra), nenhum dos quais predominante no cálculo da média. Essa afirmação é sempre verdadeira quando a população da qual a amostra é tirada obedece a uma distribuição normal.
- Se a população não tem distribuição normal, então o Teorema do Limite Central diz que a média amostral tem distribuição normal em amostras grandes. Nesse caso, "grande" quer dizer aproximadamente 25 ou mais.

Termos e conceitos fundamentais

- distribuição de probabilidades
- parâmetros de uma distribuição
- erro-padrão
- variável aleatória
- distribuição normal
- Teorema do Limite Central
- distribuição binomial
- distribuição normal padronizada
- distribuição de Poisson

Problemas

Os problemas mais difíceis têm o **número** em cor.

Problema 3.1

Dois dados são lançados e registra-se a soma dos dois resultados. Faça um gráfico da distribuição de probabilidades da soma resultante e calcule sua média e sua variância. Qual é a probabilidade de que a soma seja pelo menos igual a 9?

Problema 3.2

Dois dados são lançados e registra-se a diferença dos dois resultados. Faça um gráfico da distribuição de probabilidades resultante e calcule sua média e sua variância. Qual é a probabilidade de que o módulo da diferença seja pelo menos igual a 4?

Problema 3.3

Represente graficamente a distribuição de probabilidades do tempo provável de partida de um trem. Localize o horário programado oficial no seu gráfico.

Problema 3.4

Um trem parte a cada meia hora. Você chega à estação num momento inteiramente aleatório. Represente graficamente a distribuição de probabilidades de seu tempo de espera. Qual é o valor esperado de seu tempo de espera?

Problema 3.5

Represente graficamente a distribuição de probabilidades do número de acidentes por dia num trecho de estrada.

Problema 3.6

Represente graficamente a distribuição de probabilidades do número de acidentes por ano no mesmo trecho de estrada. Como e por que isso difere da sua resposta anterior?

Problema 3.7

Seis dados são lançados e registra-se o número de resultados iguais a 6. Calcule as probabilidades de ocorrência de 0, 1,..., 6 resultados iguais a 6 e faça um gráfico da distribuição de probabilidades.

Problema 3.8

Se a probabilidade de um menino num único nascimento é ½, independentemente do sexo das crianças anteriores, então o número de garotos numa família com dez crianças obedece a uma distribuição binomial com média 5 e variância 2,5. Em cada um dos casos a seguir, descreva como a distribuição do número de meninos difere da binomial descrita acima.

a) A probabilidade de um menino é 6/10.

b) A probabilidade de um menino é ½, mas os nascimentos não são independentes. O nascimento de um garoto faz com que seja superior a ½ a probabilidade de que a criança seguinte seja um menino.

ç) Como no item (b), exceto pelo fato de que o nascimento de um garoto faz com que seja inferior a ½ a probabilidade de que a criança seguinte seja um menino.

d) A probabilidade de um menino é 6/10 no primeiro nascimento. O nascimento de um garoto faz com que seja superior a ½ a probabilidade de que a criança seguinte seja um menino.

Problema 3.9

Uma empresa recebe componentes de um fornecedor em grandes lotes, para uso em seu processo de produção. A produção não é economicamente viável caso um lote contendo pelo menos 10% de componentes com defeito seja utilizado. A empresa verifica a qualidade de cada lote a partir de uma amostra de 15 unidades, e rejeita o lote inteiro caso encontre mais de um componente com defeito.

a) Caso seja entregue um lote com 10% de componentes com defeito, qual é a probabilidade de que seja aceito?

b) Como a empresa pode reduzir a probabilidade de aceitar incorretamente lotes insatisfatórios?

c) Se o fornecedor produzir um lote com 3% de componentes defeituosos, qual é a probabilidade de que a empresa devolva o lote?

d) Que papel exerce a hipótese de lote "grande" nos cálculos?

Problema 3.10

O recorde de número de filhos de uma mãe no Reino Unido é 39, sendo 32 deles meninas. Supondo que a probabilidade de uma menina num único nascimento seja igual a 0,5, e que essa probabilidade é independente dos nascimentos anteriores:

a) Calcule a probabilidade de 32 meninas em 39 nascimentos (você precisará usar uma calculadora científica ou um computador para ajudá-lo neste problema).

b) Esse resultado lança alguma dúvida sobre as hipóteses?

Problema 3.11

Usando a equação 3.5, que descreve a distribuição normal, e sendo $\mu = 0$ e $\sigma^2 = 1$, faça o gráfico da distribuição para os valores $x = -2; -1{,}5; -1; -0{,}5; 0; 0{,}5; 1; 1{,}5; 2$.

Problema 3.12

Repita o problema anterior para os valores $\mu = 2$ e $\sigma^2 = 3$. Utilize valores de x entre –2 e +6 com incrementos iguais a 1.

Problema 3.13

Para a variável normal padronizada z, determine:

a) $Pr(z > 1{,}64)$

b) $Pr(z > 0{,}5)$

c) $Pr(z > -1{,}5)$

d) $Pr(-2 < z < 1{,}5)$

e) $Pr(z = -0{,}75)$

Para os itens (a) e (d), faça um sombreamento das áreas relevantes do gráfico desenhado no problema 3.11.

Problema 3.14

Determine os valores de z que cortam:

a) os 10% superiores;

b) os 15% inferiores;

c) os 50% intermediários da distribuição normal padronizada.

Problema 3.15

Se $x \sim N(10, 9)$, calcule:

a) $Pr(x > 12)$

b) $Pr(x < 7)$

c) $Pr(8 < x < 15)$

d) $Pr(x = 10)$

Problema 3.16

O QI (quociente de inteligência) tem distribuição normal com média 100 e desvio-padrão igual a 16.

a) Que proporção da população tem QI acima de 120?

b) Que proporção da população tem QI entre 90 e 110?

c) No passado, cerca de 10% da população freqüentava a universidade. Atualmente, a proporção é de aproximadamente 30%. Qual era o QI do estudante "marginal" no passado? Qual é esse QI agora?

Problema 3.17

Dez adultos, selecionados ao acaso na população, têm o seu QI medido.

a) Qual é a distribuição de probabilidades do QI médio amostral?

b) Qual é a probabilidade de o QI médio da amostra ser superior a 110?

c) Se muitas amostras como essa fossem feitas, em que proporção delas você esperaria QI médio superior a 110?

d) Qual é a probabilidade de o QI médio ficar entre 90 e 110? Como se compara essa resposta à do item (b) do problema 3.16? Explique a diferença.

e) Qual é a probabilidade de uma amostra aleatória de dez estudantes universitários apresentar um QI médio superior a 110?

f) O primeiro adulto escolhido ao acaso tem QI igual a 150. Qual você esperaria ser o QI médio da amostra?

Problema 3.18

Sabe-se que a renda média de um país é igual a £10.000, com desvio-padrão de £2.500. Escolhe-se uma amostra de 40 indivíduos e sua renda média é calculada.

a) Qual é a distribuição de probabilidades dessa média amostral?

b) Qual é a probabilidade de que a média amostral seja superior a £10.500?

c) Qual é a probabilidade de que a média amostral seja inferior a £8.000?

d) Se o tamanho da amostra fosse igual a 10, por que você não poderia utilizar os mesmos métodos para responder aos demais itens?

Problema 3.19

Uma moeda é lançada dez vezes. Escreva a distribuição do número de caras:

a) Exatamente, usando a distribuição binomial.

b) Aproximadamente, usando a distribuição normal.

c) Encontre a probabilidade de quatro ou mais caras, usando os dois métodos. Quão preciso é o método normal, com e sem a correção de continuidade?

Problema 3.20

Uma máquina que produz circuitos eletrônicos tem um índice médio de defeitos de 15% (a fabricação do produto é difícil!). O custo de produção de um lote de 500 circuitos é £8.400, e os circuitos sem defeito são vendidos a £20 por unidade. Qual é a probabilidade de que a empresa tenha prejuízo em algum lote?

Problema 3.21

Um funcionário experiente na emissão de faturas comete, em média, um erro uma vez a cada cem faturas.

a) Qual é a probabilidade de encontrar um lote de cem faturas sem erro algum?

b) Qual é a probabilidade de encontrar um lote de cem faturas com mais de dois erros?

Calcule as respostas usando a distribuição binomial e a de Poisson. Se você tentar resolver o problema usando o método normal, quão precisa será a sua resposta?

Problema 3.22

Uma empresa que emprega cem funcionários tem uma taxa média de absenteísmo de 4%. Num dia qualquer, qual é a probabilidade de que (a) nenhum funcionário, (b) um funcionário, (c) mais de seis funcionários estejam ausentes?

Problema 3.23 (Projeto no computador)

Esse problema demonstra o funcionamento do Teorema do Limite Central. Em sua planilha, use a função =ALEATÓRIO() para gerar uma amostra aleatória de 25 observações (sugiro lançar a função nas células A4:A28, por exemplo). Copie essas células em cem colunas, para gerar cem amostras. Na linha 29, calcule a média de cada amostra. Agora examine a distribuição dessas médias amostrais. (Dica: você notará que a função ALEATÓRIO() recalcula automaticamente cada vez que você efetua uma operação na planilha. Isso dificulta a conclusão da análise. A solução é copiar e depois usar "Editar, Colar especial, Valores" para criar uma cópia dos valores das médias amostrais. Esses valores permanecerão estáveis.)

a) Que distribuição você esperaria para essas médias?

b) Qual é a distribuição de origem das amostras?

c) Quais são os parâmetros das distribuições de origem e das médias amostrais?

d) Os resultados estão de acordo com o que você esperava?

e) Construa uma tabela de freqüências das médias amostrais e faça um gráfico desses resultados. É parecido com o que você esperava?

f) Experimente utilizar tamanhos de amostras distintos e com distribuições originais diferentes para examinar o efeito dessas alternativas.

Problema 3.24 (Projeto)

Um jornaleiro excepcionalmente hábil com números (com o apoio de uma planilha eletrônica, como a que você precisará usar) está tentando calcular quantas cópias de um jornal ele deve encomendar. O custo do jornal para ele é igual a 15 centavos, e ele o vende por 45 centavos. As vendas têm distribuição normal, com média 250 e variância 625 por dia. As cópias não vendidas não podem ser devolvidas em troca de créditos ou reembolsadas; o jornaleiro é obrigado a jogá-las fora, perdendo 15 centavos por cópia.

a) Qual você acha que deve ser o objetivo do jornaleiro?

b) Quantas cópias ele deve encomendar?

c) O que ocorre com a variância do lucro na medida em que o jornaleiro encomenda mais cópias?

d) Calcule a probabilidade de vender mais de X cópias. (Crie uma coluna extra na planilha para esse fim.) Qual é o valor dessa probabilidade no número ótimo de cópias encomendadas?

e) Qual deveria ser o quociente entre preço e custo para justificar a encomenda de X cópias?

f) O distribuidor propõe um acordo de venda ou devolução, mas eleva o custo da cópia para 16 centavos. Deve o jornaleiro aceitar essa nova oferta?

g) Há alguma outra consideração que pode influenciar a decisão do jornaleiro?

Sugestões:

Prepare sua planilha da seguinte maneira:

col. A: (células A10:A160) 175, 176,... até 325 em incrementos unitários (representando níveis de vendas).

col. B: (células B10:B160) a probabilidade de que as vendas fiquem entre 175 e 176, entre 176 e 177, etc. até 325-326. (No Excel existe a função "=DIST. NORM()" para fazer isso – consulte a "Ajuda" do programa.)

col. C: (células C10:C160) custo total (= 0,15 × número de jornais encomendados. Coloque este último na célula F3 para que você possa usá-la como referência e alterar seu valor.)

col. D: (células D10:D160) receita total ("=MINIMO(vendas, jornais encomendados) × 0,45").

col. E: lucro (receita – custo).

col. F: lucro × probabilidade (isto é, col. E × col. B).

célula F161: soma de F10:F160 (esse é o lucro esperado).

Varie agora o número de jornais encomendados (célula F3) para determinar o valor máximo em F161. Você também pode calcular a variância do lucro de maneira bastante simples usando uma coluna adicional.

Problema 3.25 (Projeto)

Com os resultados do futebol da primeira divisão no fim de semana, veja se o número de gols marcados por jogo pode ser modelado adequadamente por um processo Poisson. Calcule inicialmente o número médio de gols marcados por jogo para toda a divisão, e depois determine a distribuição de gols por jogo usando a distribuição de Poisson. O número real de gols marcados obedece a essa distribuição? Você talvez queira usar os escores de várias semanas para conseguir resultados mais confiáveis.