

Pedro Alberto Barbetta
Marcelo Menezes Reis
Antonio Cezar Bornia

ESTATÍSTICA PARA CURSOS DE ENGENHARIA E INFORMÁTICA

3ª EDIÇÃO

editora
atlas



Estatística

Para alguns livros é disponibilizado Material
Complementar e/ou de Apoio no site da editora.
Verifique se há material disponível para este livro em

atlas.com.br

**Pedro Alberto Barbetta
Marcelo Menezes Reis
Antonio Cezar Bornia**

Estatística

**Para Cursos de Engenharia
e Informática**

3ª Edição

**SÃO PAULO
EDITORA ATLAS S.A. – 2010**

© 2004 by Editora Atlas S.A.

1. ed. 2004; 2. ed. 2008; 3. ed. 2010 (4 impressões)

Capa: Roberto de Castro Polisel

Composição: Lino-Jato Editoração Gráfica



**Dados Internacionais de Catalogação na Publicação (CIP)
(Câmara Brasileira do Livro, SP, Brasil)**

Barbetta, Pedro Alberto

Estatística : para cursos de engenharia e informática / Pedro Alberto Barbetta, Marcelo Menezes Reis, Antonio Cezar Bornia. – 3. ed. – São Paulo : Atlas, 2010.

Bibliografia.

ISBN 978-85-224-5994-0

eISBN 978-85-224-6569-9

1. Estatística I. Reis, Marcelo Menezes. II. Bornia, Antonio Cezar. III. Título.

04-0564

CDD-519.5

Índice para catálogo sistemático:

1. Estatística 519.5

TODOS OS DIREITOS RESERVADOS – É proibida a reprodução total ou parcial, de qualquer forma ou por qualquer meio. A violação dos direitos de autor (Lei nº 9.610/98) é crime estabelecido pelo artigo 184 do Código Penal.



Editora Atlas S.A.

Rua Conselheiro Nébias, 1384 (Campos Elísios)

01203-904 São Paulo (SP)

Tel.: (011) 3357-9144

www.EditoraAtlas.com.br



Sumário

Prefácio, 9

1 INTRODUÇÃO, 11

- 1.1 A estatística, 11
- 1.2 Pesquisas, dados, variabilidade e estatística, 12
- 1.3 A estatística na engenharia, 13
- 1.4 A estatística e a informática, 14
- 1.5 Modelos, 15
- 1.6 Conceitos básicos, 17

2 O PLANEJAMENTO DE UMA PESQUISA, 23

- 2.1 Aspectos gerais, 23
- 2.2 Pesquisas de levantamento, 24
- 2.3 Planejamento de experimentos, 33

3 ANÁLISE EXPLORATÓRIA DE DADOS, 50

- 3.1 Dados e variáveis, 51
- 3.2 Análise de variáveis qualitativas, 53
- 3.3 Análise de variáveis quantitativas, 58
- 3.4 Medidas descritivas, 68
- 3.5 Observações ao longo do tempo, 83
- 3.6 Análise exploratória com apoio do computador, 84
- 3.7 Orientação geral, 85

- 4 PROBABILIDADE, 91
 - 4.1 Espaço amostral e eventos, 93
 - 4.2 Definições de probabilidade, 96
 - 4.3 Probabilidade condicional e independência, 102
 - 4.4 Teorema da probabilidade total, 110
 - 4.5 Teorema de Bayes, 112
- 5 VARIÁVEIS ALEATÓRIAS DISCRETAS, 116
 - 5.1 Variável aleatória, 116
 - 5.2 Principais distribuições discretas, 126
- 6 VARIÁVEIS ALEATÓRIAS CONTÍNUAS, 140
 - 6.1 Caracterização de uma variável aleatória contínua, 140
 - 6.2 Principais modelos contínuos, 147
 - 6.3 A normal como limite de outras distribuições, 159
 - 6.4 Gráfico de probabilidade normal, 164
- 7 DISTRIBUIÇÕES AMOSTRAIS E ESTIMAÇÃO DE PARÂMETROS, 169
 - 7.1 Parâmetros e estatísticas, 169
 - 7.2 Distribuições amostrais, 174
 - 7.3 Estimação de parâmetros, 179
 - 7.4 Tamanho de amostra, 192
- 8 TESTES DE HIPÓTESES, 198
 - 8.1 As hipóteses, 198
 - 8.2 Conceitos básicos, 201
 - 8.3 Tipos de erro, 205
 - 8.4 Abordagem clássica, 206
 - 8.5 Testes unilaterais e bilaterais, 208
 - 8.6 Aplicação de testes estatísticos, 211
 - 8.7 Teste para proporção, 212
 - 8.8 Teste para média, 217
 - 8.9 Teste para variância, 222
 - 8.10 Poder de um teste e tamanho da amostra, 224
- 9 COMPARAÇÃO ENTRE TRATAMENTOS, 232
 - 9.1 Amostras independentes e em blocos, 232
 - 9.2 Teste t para duas amostras pareadas, 235
 - 9.3 Teste t para duas amostras independentes, 238
 - 9.4 Tamanho das amostras, 242

- 9.5 Teste F para duas variâncias, 247
 - 9.6 Comparação de várias médias, 248
 - 9.7 Anova em projetos fatoriais, 258
 - 9.8 Anova em projetos do tipo 2^k , 263
- 10 TESTES NÃO PARAMÉTRICOS, 273
- 10.1 Testes de aderência, 274
 - 10.2 Análise de associação, 287
 - 10.3 Testes para duas populações, 293
- 11 CORRELAÇÃO E REGRESSÃO, 316
- 11.1 Correlação, 316
 - 11.2 Coeficiente de correlação linear de Pearson, 318
 - 11.3 Regressão linear simples, 324
 - 11.4 Introdução à regressão múltipla, 346
- Anexo*, 351
- Respostas de exercícios*, 354
- Apêndice: Tabelas estatísticas*, 368
- Bibliografia*, 409



Prefácio

O presente texto é dirigido ao ensino da Estatística, especialmente para estudantes de Engenharia e Informática. Os métodos são apresentados de forma simples e intuitiva, com exemplos de motivação, aplicações e exercícios típicos da área de estudo do aluno.

Nossa experiência de vários anos ministrando disciplinas de Estatística, para diferentes cursos, sugere a conveniência de um texto de Estatística voltado para a área em que o aluno está cursando, pois isso motiva e melhora o aprendizado. Além disso, permite que o aluno aplique mais facilmente os novos conceitos e técnicas em problemas reais de sua área. Muitas aplicações e exemplos deste livro foram extraídos de dissertações e teses desenvolvidas em programas de pós-graduação de Engenharia e Ciência da Computação, enfatizando a utilização da Estatística em pesquisas nessas áreas.

O conteúdo do livro pode ser desenvolvido em um ou dois semestres. Nos cursos com a matéria de Estatística ministrada em dois semestres, recomendamos a realização de trabalhos práticos e o uso de um pacote computacional de estatística, ou mesmo uma planilha eletrônica. Optamos por não trabalhar um pacote específico, mas apresentamos várias saídas computacionais (que são mais ou menos padronizadas nos programas). Essas saídas são comentadas e interpretadas de acordo com o desenvolvimento dos conceitos. Também descrevemos alguns conceitos que facilitam o trabalho do aluno com apoio do computador.

O livro desenvolve a matéria de Estatística que é usualmente ministrada em cursos de Engenharia, Ciência da Computação e Sistemas de Informação. O Capítulo 1 mostra como a Estatística é importante na vida do Engenheiro ou

do profissional de informática, além de apresentar alguns conceitos básicos. O Capítulo 2 oferece visão geral do planejamento de uma pesquisa científica. O Capítulo 3 enfoca a análise exploratória de dados. Os Capítulos 4 a 6 descrevem uma introdução à probabilidade e os principais modelos probabilísticos. Os Capítulos 7 e 8 apresentam os princípios da inferência estatística. A matéria de aplicação mais interessante é tratada nos Capítulos 9 a 11, os quais tratam de análises estatísticas, como a comparação entre tratamentos e a análise de correlação e regressão.

Não muito comum em outros textos básicos de Estatística, mas que consideramos muito importante, é o planejamento da pesquisa, seja esta um levantamento por amostragem, seja um estudo experimental. Este texto desenvolve o assunto de forma específica no Capítulo 2. Além disso, apresenta comentários sobre o planejamento da pesquisa ao longo dos demais capítulos.

Os leitores podem entrar em contato conosco por meio de nossos endereços eletrônicos: <barbetta@inf.ufsc.br>, <cezar@eps.ufsc.br> e <marcelo@inf.ufsc.br>.

Material didático complementar, como *slides*, arquivos de dados para exercícios e referências de *softwares* livres de análise estatística, pode ser obtido em: <www.inf.ufsc.br/~barbetta/livro2.htm>.



1

Introdução

1.1 A ESTATÍSTICA

No desenvolvimento científico e em nosso próprio dia a dia, estamos sempre fazendo observações de fenômenos, gerando *dados*. Os engenheiros estão frequentemente analisando dados de propriedades dos materiais; os profissionais da informática estão avaliando dados de desempenho de novos sistemas computacionais; e todos nós, ao lermos jornais e revistas, estamos vendo resultados estatísticos provenientes do censo demográfico, de pesquisas eleitorais etc.

Os dados podem provir de estudos observacionais ou de experimentos planejados. Ao acompanhar o desempenho de um processo produtivo em sua forma natural, estamos fazendo um *estudo observacional*; ao alterar de forma proposital as variáveis do processo para verificar seus efeitos nos resultados, estamos realizando um *experimento*.

A *estatística* envolve técnicas para coletar, organizar, descrever, analisar e interpretar *dados*, ou provenientes de experimentos, ou vindos de estudos observacionais.

A análise estatística de dados geralmente tem por objetivo tomadas de decisões, resoluções de problemas ou produção de conhecimento. Mas novos conhecimentos normalmente geram novos problemas de pesquisa, resultando em um processo iterativo (veja a Figura 1.1).

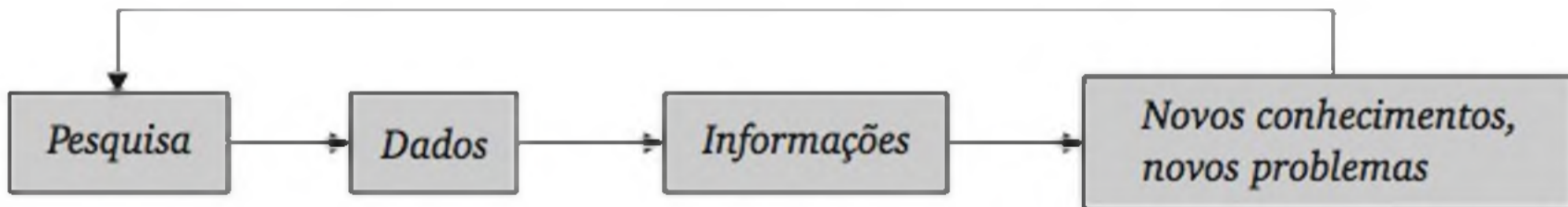


Figura 1.1 Processo iterativo das pesquisas empíricas.

Neste livro, estudaremos *técnicas de amostragens* e de *planejamento de experimentos* (Capítulo 2), as quais permitem que tenhamos observações – ou dados – capazes de responder a um problema. Mas as informações relevantes, que devem estar contidas nos dados, normalmente precisam ser realçadas para que possamos enxergá-las. Isso pode ser feito através da *análise exploratória de dados* (Capítulo 3).

As observações de um experimento costumam vir acompanhadas de *erro experimental*, ou seja, variações aleatórias devidas a uma infinidade de fatores não controláveis. E a tarefa de verificar se alguma variação é real (devida a algum fator em estudo) ou meramente resultado de flutuações aleatórias não é fácil. É por isso que estudaremos a *probabilidade*, parte da matemática preocupada em modelar fenômenos aleatórios (Capítulos 4, 5 e 6).

Inferências estatísticas, ou seja, generalizações de amostras para populações de onde elas foram extraídas, são fundamentais na resolução de problemas de engenharia e nos processos de tomada de decisões. É através de inferências estatísticas que podemos chegar à conclusão de que um material é mais resistente do que outro, que um sistema computacional gera resultados mais precisos do que outro ou, ainda, que um candidato tem intenção de voto no intervalo $30\% \pm 2\%$, com nível de confiança de 95%.

A essência de uma análise estatística é tirar conclusões sobre uma população, ou universo, com base em uma amostra de observações.

Os Capítulos 7 e 8 apresentam as ideias básicas da inferência estatística; nos Capítulos 9 a 11, são expostos métodos estatísticos propriamente ditos, que envolvem planejamento da amostragem, análise exploratória, modelos de probabilidade e inferência.

1.2 PESQUISAS, DADOS, VARIABILIDADE E ESTATÍSTICA

As pessoas normalmente associam o termo *estatística* a números, tabelas e gráficos, mas a importância da estatística fica melhor representada por dois ingredientes comuns em nosso dia a dia: *dados* e *variabilidade*.

Para o engenheiro conhecer as propriedades físicas de um novo material, ele pode medir algumas de suas características, tais como a dureza, a flexibilidade, a densidade, a porosidade etc. Mas se ele medir a dureza em vários corpos de prova do mesmo material com um instrumento de alta precisão, encontrará valores diferentes. Subestimar a presença da variabilidade pode pôr a casa a pique!

Da mesma forma, observações do tempo demandado para transmitir dados através da rede mundial de computadores, ou do número de *bytes* que passam por um servidor, variam uma enormidade ao longo do tempo. O conhecimento desses dados e de sua variabilidade torna-se imprescindível para se projetar um sistema de transmissão de dados, ou mesmo para usar o sistema existente com eficiência.

Em geral, a busca por melhorias na qualidade de um processo produtivo implica a redução da variabilidade. O que você como consumidor pensa quando vê refrigerantes de certa marca com grandes variações de conteúdo nas garrafas? E quando você resolve medir o peso de pacotes de café de 500 g e verifica que alguns têm mais de 520 g e outros têm menos que 480 g? A variabilidade pode ser reduzida com investimentos em pessoal, máquinas e tecnologia, mas muitas vezes ela pode ser acomodada com o conhecimento de relações entre fatores do processo e características funcionais do produto, o que envolve conhecimentos de engenharia, pesquisas, dados e análises estatísticas.

Com a alta competitividade de hoje, para que uma empresa sobreviva, ela tem o desafio de adequar o produto ao cliente. Por exemplo, a demanda exige que certo material tenha um valor específico de dureza. Mas como obter este valor de dureza, com a menor variabilidade possível, alterando fatores do processo, tais como: temperatura do forno, temperatura de têmpera, meio de têmpera, alterações nos componentes do material etc.? A resposta pode ser um *estudo experimental*, em que os fatores do processo são manipulados dentro de uma região operacional de forma planejada. Observações são obtidas e, através de uma análise estatística dos dados, podemos chegar à combinação ideal dos fatores do processo.

Por outro lado, adequar o produto ao cliente envolve saber o que o consumidor deseja. Mas os consumidores têm preferências diferentes, o que exige a realização de *pesquisas observacionais* (ou *de levantamento*) com os consumidores. Essas pesquisas envolvem planejamento, técnicas de amostragem, construção de questionários, organização dos dados, análises estatísticas e interpretação prática dos resultados.

1.3 A ESTATÍSTICA NA ENGENHARIA

Logo após a Revolução Industrial, métodos estatísticos foram incorporados nos processos industriais para garantir a qualidade dos produtos. Amostras de

itens produzidos eram avaliadas sistematicamente para inferir se o processo estava sob controle. Mais recentemente, a avaliação da qualidade passou a ser feita ao longo de todo o processo produtivo como forma de corrigir eventuais falhas no sistema assim que elas aparecessem. Isso levou a um aumento da qualidade do produto final e redução de custos, pois se reduziram drasticamente as perdas por defeitos.

Além do acompanhamento estatístico da qualidade, as indústrias costumam fazer experimentos estatisticamente planejados para encontrar a combinação dos níveis dos fatores do processo que levem a melhor qualidade possível. Na outra ponta, as empresas levantam dados de amostras de consumidores para realizar pesquisas de marketing direcionadas ou para adequar os produtos aos clientes. O planejamento dessas amostras e a análise dos dados necessitam de técnicas estatísticas.

Muitas vezes, a relação entre estatística e engenharia é ainda mais estreita. Os próprios métodos de engenharia costumam incorporar intrinsecamente procedimentos probabilísticos ou estatísticos. Assim, para que um aluno possa entender certos métodos de engenharia, é necessário que tenha conhecimentos de probabilidade e estatística.

1.4 A ESTATÍSTICA E A INFORMÁTICA

Enquanto a Informática é a ciência que trata da informação através de meios eletrônicos, a Estatística procura obter informações relevantes de massas de dados e, nos dias de hoje, isso costuma ser feito com auxílio do computador.

A variabilidade está onipresente nos sistemas computacionais atuais. Você pode observar diferentes tempos de resposta ao carregar um aplicativo num sistema compartilhado, ao transmitir uma mensagem no correio eletrônico etc. Portanto, a análise do desempenho desses sistemas computacionais exige tratamento estatístico.

É comum construir sistemas para simular certas situações reais. Mas, como no mundo real os acontecimentos nem sempre são previsíveis, torna-se necessário incluir no modelo de simulação alguma aleatoriedade, que pode ser feita com base em modelos de probabilidade. Por exemplo, pode ser razoável supor que em uma fila cheguem, em média, cinco indivíduos por minuto, mas o número exato de indivíduos que vão chegar no próximo minuto não é totalmente previsível.

Outra relação importante é o uso conjunto de banco de dados, estatística e inteligência artificial para extrair informações relevantes e não triviais de grandes arquivos de dados, armazenados sob diferentes formatos e em diferentes locais. Por exemplo, as empresas telefônicas têm dados das ligações telefônicas

de seus milhares ou até milhões de clientes. Mas é um grande desafio encontrar, com base nesses dados, possíveis fraudes, tais como as clonagens de telefones celulares. Este é um caso típico da necessidade de usar de forma conjunta técnicas estatísticas e informática.

1.5 MODELOS

Os modelos podem ser considerados como alguma representação da realidade em estudo, destacando aspectos relevantes e desprezando detalhes insignificantes. Em geral, eles servem para simplificar, descrever e facilitar a interpretação daquilo que se está estudando.

Na engenharia, o estudante costuma defrontar com os chamados *modelos determinísticos*, isto é, conhecidas as entradas x_1, x_2, \dots, x_k , o modelo permite chegar ao resultado y , usando uma função $y = f(x_1, x_2, \dots, x_k)$. É o que acontece, por exemplo, com a Lei de Ohm, em que, dadas a tensão (x_1) e a resistência (x_2) de um circuito simples, podemos calcular o fluxo da corrente elétrica (y) por:

$$y = \frac{x_1}{x_2} \quad (1.1)$$

Muitas vezes, porém, as condições do experimento não permitem deduzir qual o resultado, mas somente a chance (ou a *probabilidade*) de possíveis resultados. É o caso da observação da face voltada para cima no lançamento imparcial de uma moeda perfeitamente equilibrada. Antes da realização do experimento não se tem como dizer o resultado, mas é razoável atribuir probabilidade 0,5 para *cara* e 0,5 para *coroa*. É um exemplo de *modelo probabilístico* ou *estocástico*.

Um exemplo menos trivial de modelo probabilístico é a descrição do número de indivíduos que chegam a uma fila, ou do número de pacotes de dados que chegam a um servidor por segundo. Como veremos no Capítulo 4, sob certas condições e admitindo que a taxa média de chegadas por segundo é λ (um valor positivo fixo), a probabilidade de chegar exatamente k pacotes num dado segundo é de, aproximadamente:

$$p(k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad (k = 0, 1, 2, \dots \text{ e } e \cong 2,7183) \quad (1.2)$$

Esse tipo de modelo pode auxiliar o projetista a planejar a capacidade de um sistema computacional.

Todo estudante já deve ter-se defrontado com os *modelos mecânicos*, caracterizados por serem totalmente deduzidos do conhecimento sobre o fenôme-

no físico em questão – a Lei de Ohm é um exemplo. De outro lado estão os chamados *modelos empíricos*, que são construídos com base em observações reais sobre o problema em estudo. Por exemplo, podemos ter interesse em conhecer a relação entre a resistência à compressão de um concreto e seu tempo de hidratação. Para isso, podemos realizar um experimento, que resulta em observações dessas duas variáveis. A Figura 1.2 apresenta os resultados da resistência (MPa) de 11 corpos de prova, com tempos de hidratação entre 10 e 20 dias.

Corpo de prova	Tempo de hidratação (dias)	Resistência (MPa)
1	10	11,3
2	11	12,1
3	12	16,4
4	13	16,3
5	14	20,2
6	15	20,5
7	16	25,0
8	17	26,4
9	18	26,2
10	19	28,4
11	20	30,2

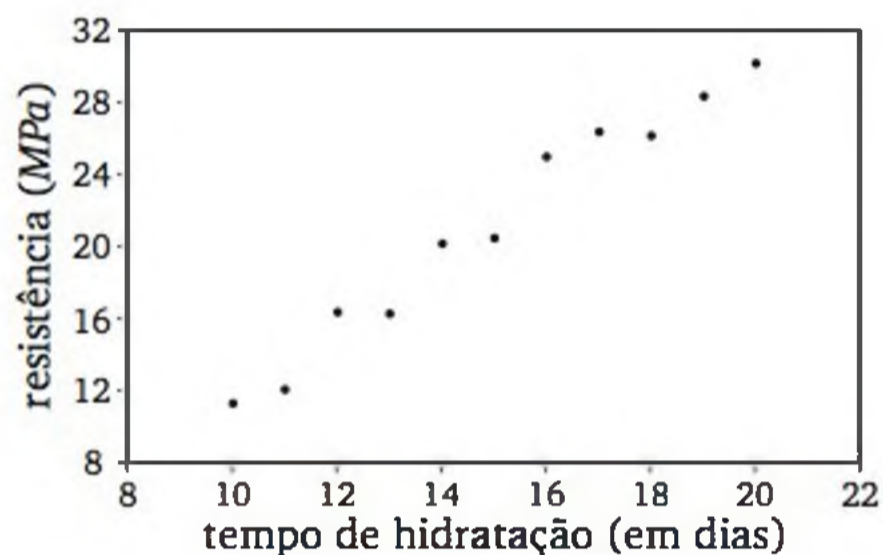


Figura 1.2 Resultados de um experimento sobre resistência à compressão de concreto, em função do tempo de hidratação – dados e gráfico.

A Figura 1.2 mostra que não se tem uma função matemática simples para explicar exatamente a relação entre as duas variáveis em questão. Contudo, o gráfico expõe os pontos *em torno* de uma reta. Ou seja, podemos admitir que a *resistência esperada* do concreto se relaciona linearmente com o tempo de hidratação; e o fato de os pontos observados não estarem exatamente sobre uma reta é porque existem inúmeros fatores não controláveis que agem sobre o processo – o *erro experimental*.

Uma função matemática que *explica aproximadamente* o relacionamento entre duas ou mais variáveis, construída com base em dados observados, pode ser considerada um *modelo de regressão*, um tipo especial de *modelo empírico*. Dado um problema, o conhecimento de engenharia é fundamental para escolher adequadamente as variáveis e, às vezes, a forma funcional (uma reta, uma parábola etc.), mas a construção completa do modelo é feita através dos dados.

No exemplo em questão, a Figura 1.2 sugere que uma reta ($y = \alpha + \beta x$) descreve aproximadamente o relacionamento. As 11 observações da resistência (y) para diferentes tempos de hidratação (x) são usadas para obter valores de α

e β adequados, conforme estudaremos no Capítulo 11. A Figura 1.3 mostra a equação de regressão analítica e graficamente. O chapéu sobre y é para diferenciar o modelo (a reta) dos valores efetivamente observados (os pontos).

$$\hat{y} = 8,01 + 1,95x$$

onde x é o tempo de hidratação e \hat{y} é o valor de resistência predito pelo modelo.

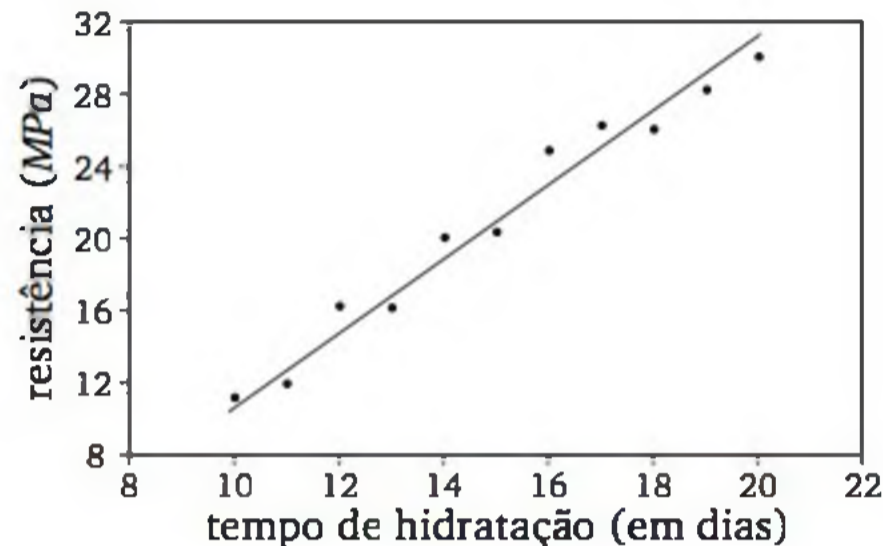


Figura 1.3 Exemplo de um modelo empírico para explicar a resistência de um concreto, em função do tempo de hidratação.

1.6 CONCEITOS BÁSICOS

Apresentaremos alguns conceitos que facilitarão a leitura deste livro. Esses conceitos serão retomados nos capítulos seguintes com definições mais precisas. O exemplo seguinte será usado para ilustrar os principais conceitos.

Exemplo 1.1 Considere uma indústria processadora de suco de frutas. Ao receber um carregamento de laranjas, os técnicos fazem inspeção da qualidade nas frutas. Examinam uma *amostra* de cinco caixas, tomadas de forma *aleatória* dentre toda a *população* de caixas do carregamento.

População: conjunto de elementos que formam o universo de nosso estudo que são passíveis de ser observados, sob as mesmas condições.

Amostra: parte dos elementos de uma população.

Amostragem: processo de seleção da amostra.

Amostragem aleatória simples: o processo de seleção dos elementos é feito por sorteios, fazendo com que todos os elementos da população tenham a mesma chance de ser escolhidos e, além disso, todo subconjunto de n elementos tenha a mesma chance de fazer parte da amostra.

Algumas *características* (ou *variáveis*) podem ser observadas nas cinco caixas de laranjas amostradas, tais como:

- a) uma classificação por um técnico especializado (ótima, boa, regular, ruim ou péssima);
- b) número de laranjas não aproveitáveis por caixa;
- c) peso de cada caixa de laranja etc.

O nível de mensuração de uma variável pode ser *qualitativo*, como no caso (a), em que o resultado é uma qualidade ou atributo; ou *quantitativo*, como nos demais casos, em que o resultado é um valor numa dada escala de medidas. As variáveis mensuradas ao nível qualitativo serão chamadas de *variáveis qualitativas*, e as mensuradas ao nível quantitativo, de *variáveis quantitativas*.

Ao selecionar uma caixa de laranja do carregamento, podemos contar o número de laranjas não aproveitáveis e medir o peso da caixa. Como o resultado de cada variável depende do processo aleatório de seleção da caixa de laranja, preferimos usar a denominação *variável aleatória*.

Uma **variável aleatória** pode ser entendida como uma variável quantitativa, cujo resultado depende de fatores aleatórios.

Ao realizar as observações de certa variável aleatória X , estamos observando uma *amostra de n elementos*, $\{x_1, x_2, \dots, x_n\}$, da variável aleatória X . Por exemplo, ao contar o número de laranjas não aproveitáveis em cada uma das cinco caixas amostradas, temos um conjunto de cinco valores, digamos $\{4, 6, 2, 3, 0\}$, que corresponde à amostra efetivamente observada da variável aleatória $X = \text{número de laranjas não aproveitáveis por caixa}$.¹

Dada uma amostra, é comum calcular medidas descritivas das observações. Em nosso exemplo, podemos dizer que o número *médio* de laranjas não aproveitáveis por caixa é $(4 + 6 + 2 + 3 + 0)/5 = 3$. Esse valor descreve o que se observou na amostra, mas também pode ser interpretado como uma *estimativa* do número médio de laranjas não aproveitáveis por caixa, no carregamento todo.

Parâmetro: uma medida que descreve certa característica dos elementos da *população*.

Estatística: uma medida que descreve certa característica dos elementos da *amostra*.

Estimativa: valor resultante do cálculo de uma estatística, quando usado para se ter uma ideia do parâmetro de interesse.

¹ Observe que, antes de se fazer a seleção das cinco caixas, temos, na verdade, um conjunto de cinco variáveis aleatórias $\{X_1, X_2, \dots, X_5\}$, pois o valor de cada uma vai depender das caixas selecionadas.

Exemplos de *parâmetros* podem ser:

- *média* do número de laranjas não aproveitáveis por caixa;
- *proporção* do número de caixas classificadas como *ótima* etc.

Uma média ou proporção, quando referente a uma *amostra* e não a toda a população, é chamada de *estatística*. Note que as definições de parâmetro e estatística são muito parecidas, só que parâmetro refere-se à população e estatística, à amostra.² Por sua vez, o termo *estimativa* refere-se a um resultado numérico, referente à amostra efetivamente observada.

No Capítulo 3, estudaremos várias medidas descritivas da amostra (estatísticas) ou da população (parâmetros). Mas apresentaremos aqui duas medidas bastante usadas: a média e a variância.

Média

Sejam n observações efetivas de certa variável aleatória X : $\{x_1, x_2, \dots, x_n\}$. Definimos média aritmética por:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i \quad (1.3)$$

Muitas vezes, vamos nos referir ao cálculo da média aritmética antes da observação efetiva da amostra. Nesse caso, temos uma *estatística*, denotada por \bar{X} e chamada de *média amostral*.³

Por outro lado, caso se conheçam todas as possíveis observações de X (por exemplo, a contagem em todas as caixas de laranja do carregamento), dadas por x_1, x_2, \dots, x_N , definimos o *parâmetro* média aritmética (ou *média populacional*) por:⁴

$$\mu = \frac{x_1 + x_2 + \dots + x_N}{N} = \frac{1}{N} \sum_{i=1}^N x_i \quad (1.4)$$

Às vezes, a população não tem tamanho finito. Imagine, por exemplo, realizar n observações do tempo de carga T de certo aplicativo, num sistema com-

² Outra diferença é que os parâmetros são números reais, embora normalmente desconhecidos. Já as estatísticas, antes de efetuar a amostragem, são variáveis aleatórias (dependem da amostra a ser selecionada).

³ Note que a média amostral é uma variável aleatória, pois seu valor depende da amostra a ser selecionada.

⁴ Uma convenção neste livro é denotar as estatísticas por letras latinas e os parâmetros por letras gregas; também usaremos letras maiúsculas para variáveis aleatórias e letras minúsculas para representar observações efetivas.

partilhado, obtendo-se uma amostra $\{t_1, t_2, \dots, t_n\}$. O número de observações que você poderia fazer é ilimitado, mas mesmo assim pode considerar a população como o conjunto (infinito) de possíveis observações de T , sob as mesmas condições. E, também, considerar a média calculada da amostra, \bar{t} , como uma *estimativa* de uma média populacional μ , que representa o *tempo médio* ou *tempo esperado* de carga do aplicativo.

Desvios

A média pode ser considerada como um valor central, uma tentativa de conhecer o valor real daquilo que se está medindo, desconsiderando o erro experimental. Para uma análise mais rigorosa do problema, torna-se fundamental incluir alguma medida de variabilidade, como a *variância*. Esta medida baseia-se nos desvios de cada valor em relação à média: $d_i = x_i - \bar{x}$ ($i = 1, 2, \dots, n$), como ilustra a Figura 1.4.

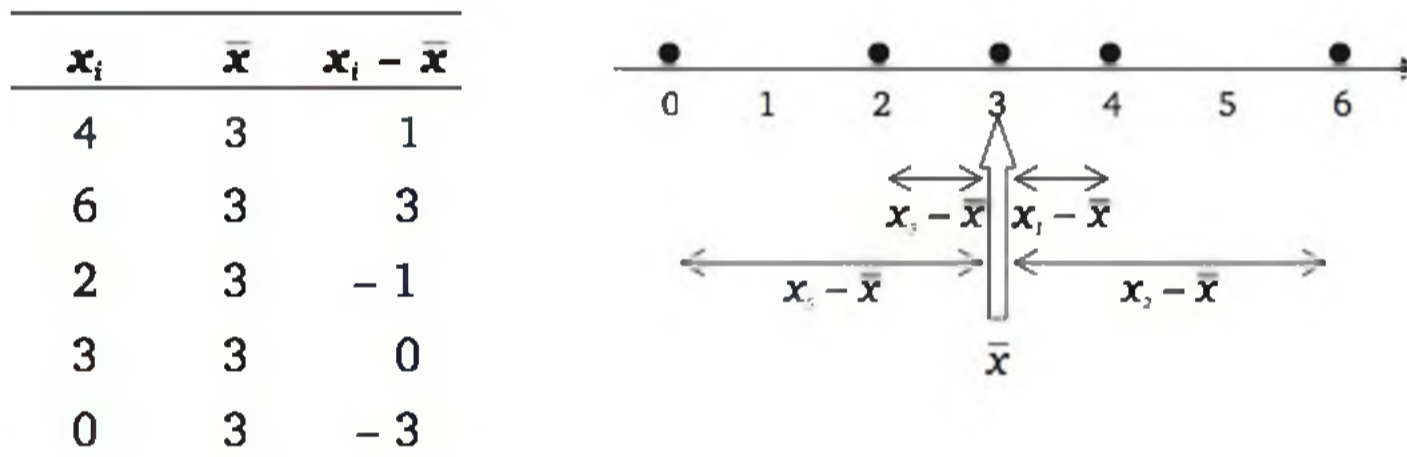


Figura 1.4 Desvios dos valores em relação à média: elementos básicos no cálculo da variância.

Graus de liberdade

Podemos identificar uma observação x como um ponto em \mathcal{R} (pode ser qualquer ponto de um subconjunto de \mathcal{R} formado pelas restrições da mensuração). Da mesma forma, se são realizadas n observações independentes (x_1, x_2, \dots, x_n) , podemos identificá-las como um ponto em \mathcal{R}^n (pode ser qualquer ponto de um subconjunto de \mathcal{R}^n formado pelas restrições da mensuração de cada observação). Dizemos, então, que uma amostra de n observações *tem n graus de liberdade*. Os desvios, $d_i = x_i - \bar{x}$, embora sejam n resultados numéricos, têm soma nula, fazendo com que o último desvio fique completamente determinado pelos demais.⁵ Assim, os desvios têm $(n - 1)$ *graus de liberdade*. Desta forma,

⁵ $\sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n x_i - \sum_{i=1}^n \bar{x} = n\bar{x} - n\bar{x} = 0.$

uma amostra de tamanho n tem n graus de liberdade, os quais podem ser decompostos da seguinte maneira: 1 grau de liberdade para a média \bar{x} e $(n - 1)$ graus de liberdade para os desvios $d_i = x_i - \bar{x}$.

Variância

Para calcular a variância, devemos considerar os desvios de cada valor em relação à média aritmética. Depois, construímos uma espécie de média desses desvios. Para evitar o problema dos desvios negativos, vamos trabalhar com os desvios quadráticos, $d_i^2 = (x_i - \bar{x})^2$. A variância é definida como a média aritmética dos desvios quadráticos. Por conveniência, vamos usar como denominador $(n - 1)$ no lugar de n , ou seja, o número de graus de liberdade associados aos desvios.⁶ Assim, definimos a **variância** de um conjunto de valores por:

$$s^2 = \frac{1}{n - 1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (1.4)$$

Segue o cálculo da variância para os dados relativos ao número de laranjas não aproveitáveis numa amostra de cinco caixas:

x_i	\bar{x}	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
4	3	1	1
6	3	3	9
2	3	-1	1
3	3	0	0
0	3	-3	9

$$s^2 = \frac{1}{n - 1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{20}{4} = 5$$

O valor da variância sozinho não traz informações relevantes, mas se considerarmos que em uma amostra de outro carregamento a variância foi igual a 10, podemos deduzir que a segunda amostra tem mais variabilidade.

Caso se conheçam todas as possíveis observações de X (por exemplo, a contagem em todas as caixas de laranja do carregamento), dadas por (x_1, x_2, \dots, x_N) , definimos o *parâmetro* variância (ou *variância populacional*) por:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 \quad (1.5)$$

onde μ representa a média da população.

⁶ Uma justificativa mais formal dessa conveniência será dada no Capítulo 7, seção 7.3.1.

Desvio padrão

A unidade de medida da variância é a unidade dos dados ao quadrado. No Capítulo 3, são apresentadas formas alternativas para o cálculo da variância e outras medidas de variabilidade, dentre elas o *desvio padrão*, que é definido como a raiz quadrada positiva da variância e, portanto, tem a mesma unidade de medida dos dados. No exemplo precedente, onde a variância foi de $s^2 = 5$ laranjas², o desvio padrão é igual a:

$$s = \sqrt{5 \text{ laranjas}^2} = 2,236 \text{ laranjas}$$

EXERCÍCIOS

1. Dê um exemplo de uma situação prática em que é mais razoável um modelo empírico do que um modelo determinístico.
2. Apresente, em uma situação prática, qual é a população, uma forma de amostragem e uma possível amostra.
3. Dada a seguinte amostra: {7, 8, 6, 5, 9, 4}, calcule:
 - a) a média;
 - b) a variância;
 - c) o desvio padrão.
4. Para avaliar a qualidade de três empacotadoras (A, B e C) de uma indústria de torrefação de café, realizou-se uma amostra de dez pacotes de café de cada empacotadora e mediu-se o peso líquido. O valor declarado é de 500 g. A empacotadora A apresentou peso médio igual a 500,1 g e variância 6,2; a B resultou em peso médio igual a 499,9 g e variância 40,5 e a C, peso médio igual a 530,3 g e variância 5,8. O que se pode dizer sobre as empacotadoras?
5. Ao calcular a variância de um conjunto de valores, encontrou-se o valor $s^2 = 0$. O que se pode dizer sobre o conjunto de valores?

O Planejamento de uma Pesquisa

2.1 ASPECTOS GERAIS

Para que os resultados de uma análise estatística de dados produzam informações úteis, os dados precisam ser coletados de forma planejada. A Figura 2.1 ilustra as principais etapas de uma pesquisa, enfatizando que os métodos estatísticos precisam ser pensados ainda na fase do planejamento da pesquisa.

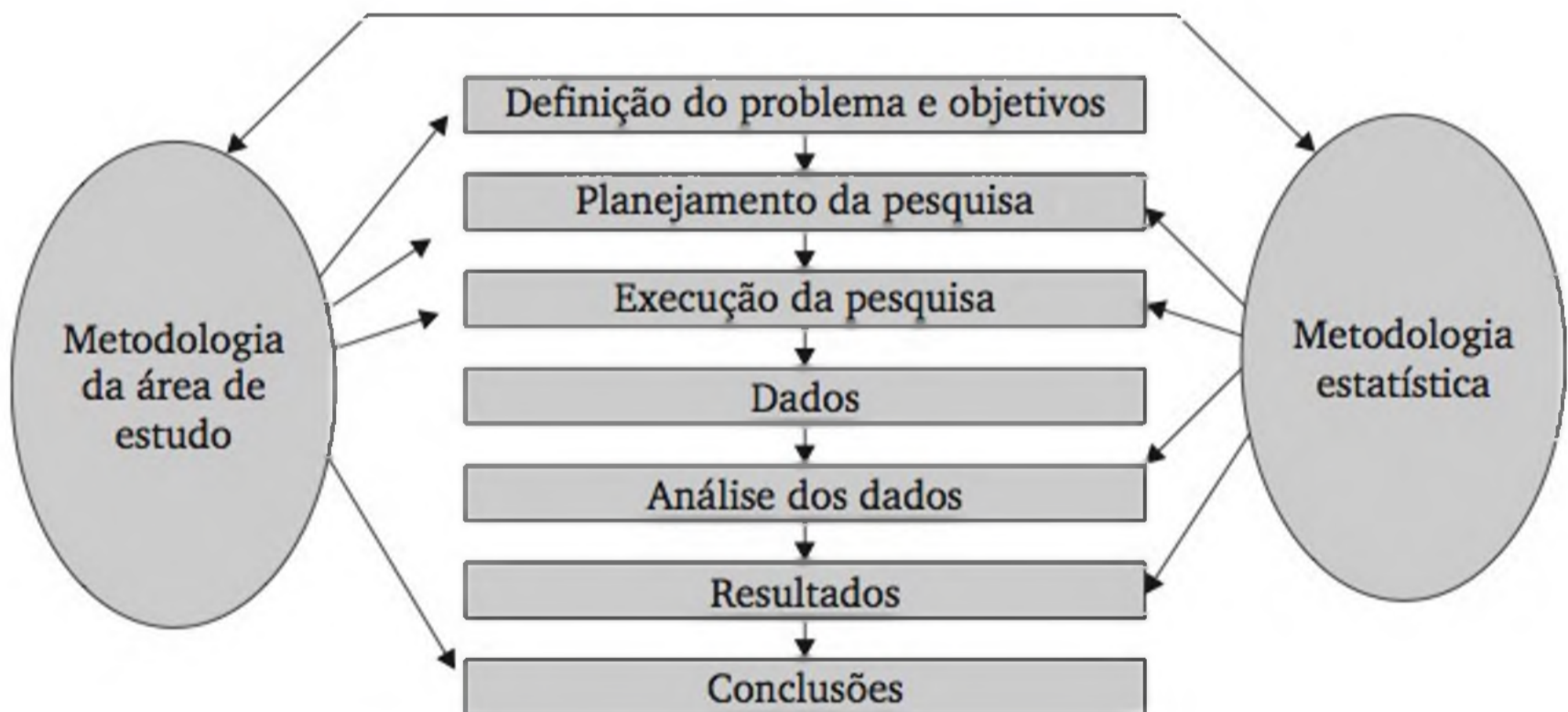


Figura 2.1 *Etapas usuais de uma pesquisa empírica.*

Embora a Figura 2.1 ilustre as etapas da pesquisa em sequência, na fase do planejamento é necessário também pensar na forma de análise dos dados, pois,

dependendo da análise estatística que se deseja fazer, o projeto de pesquisa deve ter suas peculiaridades. Assim, não discutiremos a questão do planejamento de pesquisa somente neste capítulo, mas também quando estivermos discutindo diferentes modelos de análise (Capítulos 7 a 11).

Em função do problema e dos objetivos da pesquisa, devemos decidir entre uma pesquisa observacional e uma pesquisa experimental. Numa *pesquisa observacional* (ou *de levantamento*) as características de uma população são levantadas (observadas ou medidas), mas sem manipulação. É o caso de um censo demográfico, pesquisas eleitorais, pesquisas de mercado, inspeção da qualidade etc. Em todos esses casos, quer-se ter ideia de certa população tal qual ela é na natureza ou no processo.

Nas *pesquisas experimentais*, grupos de indivíduos (ou animais, ou objetos) são manipulados para se avaliar o efeito de diferentes tratamentos. É o caso de se verificar o rendimento de um processo químico para diferentes temperaturas de reação, que são manipuladas de acordo com o interesse prático.

A seção 2.2 discute aspectos relacionados com pesquisas de levantamento, enquanto a seção 2.3 enfatiza o planejamento de pesquisas experimentais.

2.2 PESQUISAS DE LEVANTAMENTO

Delimitação da população

Conforme definimos no capítulo anterior, chamamos de *população* o conjunto de elementos (indivíduos, objetos etc.) que formam o universo de nosso estudo e que são passíveis de ser observados, sob as mesmas condições. Num processo de inspeção da qualidade, a população pode ser considerada como o conjunto de todos os itens que saem da linha de produção; numa pesquisa de mercado, a população é o conjunto de possíveis consumidores; e assim por diante.

Ao desenvolver uma pesquisa, planejamos fazer com que as conclusões sejam válidas para toda a população previamente definida, mesmo que a pesquisa seja feita por amostragem. Contudo, se, numa pesquisa para estudar a dureza de um tipo de aço, só tivermos disponibilidade de trabalhar com corpos de prova de certo lote desse aço, tecnicamente as conclusões só valem para esse lote.

Variáveis a serem levantadas

As características que observamos nos elementos da população são chamadas de *variáveis*. No exemplo anterior, a *dureza* é a variável fundamental de nosso estudo. A cada corpo de prova, temos um e apenas um valor da dureza, que supostamente é medida sempre sob as mesmas condições.

A maioria dos estudos envolve mais que uma variável. A seleção adequada das variáveis é fundamental para o sucesso de uma pesquisa.

Instrumentos para a mensuração de variáveis

Para realizar medidas físicas, normalmente temos instrumentos bem estabelecidos e unidades de medidas padrões. Por exemplo, o comprimento pode ser medido por uma régua, que usa a escala métrica; a temperatura, por um termômetro, que usa a escala de graus Celsius; e assim por diante. Quando desejamos medir variáveis em seres humanos, geralmente precisamos construir um *questionário*. Para variáveis quantitativas, devemos deixar claras as unidades de medida; para variáveis qualitativas, devemos especificar as possíveis categorias ou atributos. Veja o exemplo seguinte:

Há quanto tempo o Sr. (ou Sra.) trabalha nesta empresa?

_____ anos completos.

Qual é seu estado civil? solteiro casado

viúvo desquitado divorciado

Em muitas situações, é comum tentar extrair de um indivíduo o grau de certa característica (satisfação, conhecimento, aptidão etc.). Veja o exemplo seguinte:

Dê uma nota de 1 (um) a 5 (cinco), sendo 1 o grau mínimo e 5 o grau máximo, para as seguintes características relacionadas com você e o curso que está fazendo.

- a) Didática dos professores de seu curso (1 2 3 4 5)
- b) Grau de conhecimento dos professores (1 2 3 4 5)
- c) Bibliografia disponível (1 2 3 4 5)
- d) Laboratórios e outros recursos materiais (1 2 3 4 5)
- e) Conteúdo dos programas das disciplinas oferecidas . . (1 2 3 4 5)
- f) Encadeamento das disciplinas (1 2 3 4 5)

É comum o uso de procedimentos típicos de variáveis quantitativas (cálculo de médias aritméticas, por exemplo) para analisar dados gerados por esse tipo de instrumento, embora seja um pouco forçoso assumir que essas escalas sejam realmente quantitativas.

Censo ou amostragem

A palavra *censo* refere-se à pesquisa de todos os elementos da população. Geralmente, realizamos um censo quando:

- a população é pequena;
- as variáveis são fáceis de ser medidas ou observadas; ou
- necessitamos de resultados exatos.

Grande parte das pesquisas científicas ou de resoluções de problemas de engenharia é feita por *amostragem*, ou seja, observamos apenas um subconjunto de elementos da população. A amostragem é particularmente interessante quando:

- a população é grande ou infinita;
- as observações ou mensurações têm alto custo;
- as medidas exigem testes destrutivos;
- há necessidade de rapidez etc.

Em geral, o uso de amostragem leva à redução de custos e tempo. Mas a amostragem precisa ser feita com critérios, pois pretendemos ter amostras que permitam, a partir de uma análise estatística apropriada, obter conclusões satisfatórias sobre toda a população (veja a Figura 2.2).

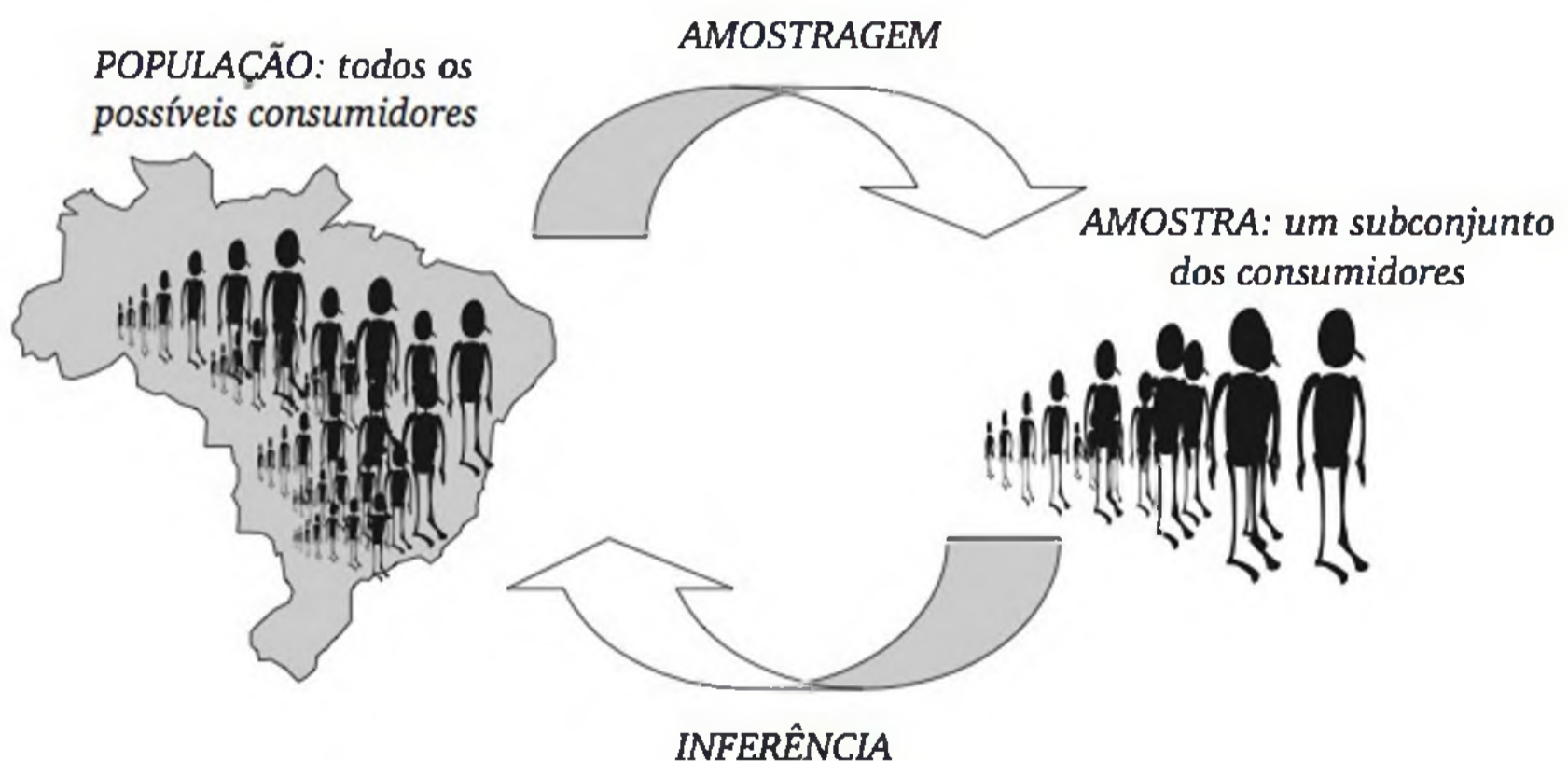


Figura 2.2 Ilustração de um levantamento por amostragem para avaliar a preferência do consumidor.

2.2.1 Técnicas de amostragem

Nesta seção, assumiremos que a população seja finita e composta de N elementos, salvo quando explicitamos o contrário. O número de elementos que serão amostrados será representado por n .

Daremos ênfase às chamadas amostragens *probabilísticas*, que se caracterizam por garantir que cada elemento da população tenha certa probabilidade p conhecida de pertencer à amostra, sendo $0 < p < 1$. Nessas amostragens, sempre ocorre algum *sorteio*.

As amostragens probabilísticas são particularmente importantes nos processos de inferência, pois os métodos estatísticos são construídos sob suas propriedades. Descreveremos quatro tipos de amostragens probabilísticas:

- amostragem aleatória simples;
- amostragem sistemática;
- amostragem estratificada;
- amostragem de conglomerados.

Amostragem aleatória simples

Para a seleção de uma amostra aleatória simples, precisamos ter uma *lista* completa dos elementos da população. Esse tipo de amostragem consiste em selecionar a amostra através de sorteios, sem restrição.

A amostragem aleatória simples tem a seguinte propriedade:

Qualquer subconjunto da população, com o mesmo número de elementos, tem a mesma probabilidade de fazer parte da amostra.

Em particular, temos que cada elemento da população tem a mesma probabilidade de pertencer à amostra e esta probabilidade é dada por n/N .

A seleção de uma amostra aleatória simples pode ser facilitada com o uso de números aleatórios, os seja, números resultantes de sucessivos sorteios do conjunto $\{0, 1, 2, \dots, 9\}$, fazendo com que todo número com a mesma quantidade de algarismos tenha a mesma probabilidade de ocorrência. Existem algoritmos computacionais capazes de gerar dados que satisfazem aproximadamente a esta propriedade, resultando nos chamados números *pseudo-aleatórios*. O quadro a seguir apresenta alguns números pseudo-aleatórios gerados pelo Microsoft Excel:

3820	1007	5964	8990	8845	9584	0145	4074	8632	1386	3002	8021	6960	2715	9040
2450	0455	0324	1641	2196	0171	2850	3431	5536	3573	2913	8021	7889	6759	7553
3718	3556	9102	4660	4261	3039	9756	8066	9911	2562	8503	5570	8730	4410	2177
9516	0534	7050	8164	9724	4663	3002	7501	3514	7756	3297	0860	9768	2855	5343
0743	1984	0641	3583	4870	5112	3734	9858	0407	2307	5745	7060	4014	1110	8973
0050	9261	1003	2567	7756	6796	8090	7243	0850	1323	6568	2584	7651	7002	8587
7561	6265	1736	4048	5523	7114	5551	1811	9702	6869	9120	9542	5943	5576	9681
5287	7966	8056	2622	1779	8667	1148	0595	7615	7383	6680	9268	4517	1681	0619
9862	9255	9038	5449	5007	6749	4898	1458	0380	7962	6018	9300	5339	1320	0823

Como ilustração, considere o problema de extrair uma amostra aleatória simples de tamanho 5 da seguinte população de funcionários de uma empresa, na qual identificamos cada indivíduo com um número:

01. Aristóteles	02. Anastácia	03. Arnaldo	04. Bartolomeu	05. Bernardino
06. Cardoso	07. Carlito	08. Cláudio	09. Ermílio	10. Ercílio
11. Ernestino	12. Endevaldo	13. Francisco	14. Felício	15. Fabrício
16. Geraldo	17. Gabriel	18. Getúlio	19. Hiraldo	20. João da Silva
21. Joana	22. Joaquim	23. Joaquina	24. José da Silva	25. José de Souza
26. Josefa	27. Josefina	28. Maria José	29. Maria Cristina	30. Mauro
31. Paula	32. Paulo César			

Para extrairmos uma amostra aleatória simples de tamanho $n = 5$, basta tomar cinco números aleatórios do conjunto $\{01, 02, \dots, 32\}$. Os funcionários associados aos números selecionados formarão a amostra. Tomando números da primeira linha e desprezando os valores que estiverem fora do conjunto $\{01, 02, \dots, 32\}$ e os valores que se repetirem, temos:

Números aleatórios: 20 10 07 01 32.

Amostra: {João da Silva, Ercílio, Carlito, Aristóteles, Paulo César}

A dificuldade para fazer uma amostragem desse tipo é ter um arquivo com todos os elementos da população, pois a geração de números aleatórios e a identificação dos indivíduos podem ser feitas com auxílio do computador.

A importância da amostragem aleatória simples é que as técnicas estatísticas que serão apresentadas neste livro – e nos livros de estatística em geral – supõem esse tipo de amostragem.

Amostragem sistemática

Um processo mais simples é sortear o primeiro elemento e extrair os demais sistematicamente. Mais especificamente:

- calcula-se o intervalo de seleção, dado por $I = N/n$, desprezando as decimais;
- sorteia-se o primeiro elemento do conjunto $\{1, 2, \dots, I\}$; e
- completa-se a amostra, extraíndo um elemento a cada I elementos.

No exemplo precedente, temos o intervalo de seleção $I = 6$. Devemos, então, sortear um elemento dentre os seis primeiros. Podemos fazer isso extraíndo um número, de um algarismo, do conjunto de números aleatórios apresentado anteriormente. Tomaremos o primeiro número do conjunto que é “3”. Então, o primeiro funcionário da amostra é o “Arnaldo”. Os demais são obtidos pelo intervalo de seleção “6”, a partir do Arnaldo, resultando na seguinte amostra:¹

(3) (9) (15) (21) (27)
 {Arnaldo, Ermílio, Fabrício, Joana, Josefina}

Amostragem estratificada

A técnica da amostragem estratificada consiste em dividir a população em subgrupos, que denominaremos de **estratos**. Esses estratos devem ser internamente mais homogêneos do que a população toda, com respeito às variáveis em estudo. Por exemplo, se para estudar a dureza de certo aço temos corpos de prova de dois fornecedores, então a população dos corpos de prova pode ser dividida em dois estratos.

Sobre os diversos estratos da população são realizadas seleções aleatórias de forma independente. A amostra completa é obtida através da agregação das amostras de cada estrato (veja a Figura 2.3).

Amostragem estratificada proporcional: a proporcionalidade do tamanho de cada estrato da população é mantida na amostra. Por exemplo, se um estrato abrange 20% da população, ele também deve abranger 20% da amostra.

Amostragem estratificada uniforme: selecionamos o mesmo número de elementos em cada estrato. É o processo usual quando se deseja comparar os diversos estratos.

¹ Devido ao uso da parte inteira do número no cálculo do intervalo de seleção, o número n de elementos da amostra pode ficar acima do valor planejado.

POPULAÇÃO

variância σ^2

Estrato 1 variância σ_1^2
Estrato 2 variância σ_2^2
...
Estrato k variância σ_k^2

Suposição: $\sigma_i^2 < \sigma^2$ ($i = 1, 2, \dots, k$)

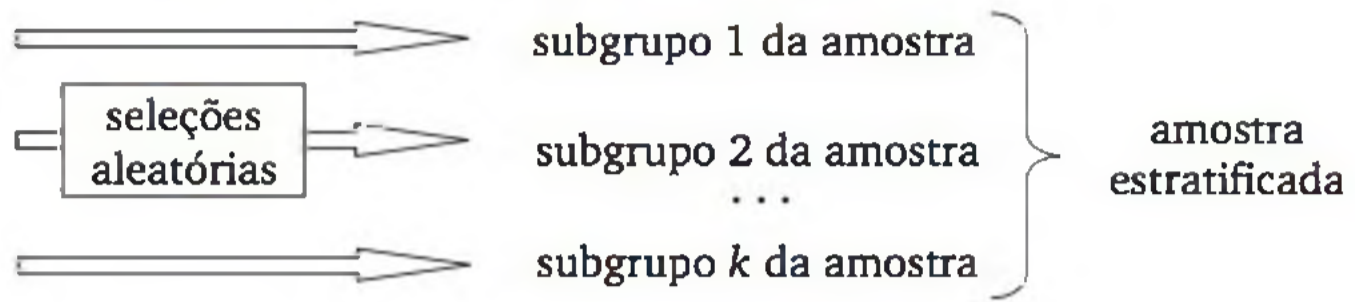


Figura 2.3 Processo de amostragem estratificada.

Amostragem de conglomerados

Ao contrário da amostragem estratificada, a amostragem de conglomerados tende a produzir uma amostra que gera resultados menos precisos, quando comparada com uma amostra aleatória simples de mesmo tamanho. Contudo, seu custo financeiro tende a ser bem menor, especialmente em amostragens de grandes populações.

Chamamos de *conglomerado* um grupamento de elementos da população. Por exemplo, numa população de domicílios de uma cidade, os bairros formam *conglomerados* de domicílios. Ao avaliar a qualidade das laranjas de um carregamento, as caixas de laranjas são os conglomerados.

Esse tipo de amostragem consiste, num primeiro estágio, em selecionar conglomerados de elementos. Num segundo estágio, ou se observam todos os elementos dos conglomerados selecionados no primeiro estágio (*amostragem de conglomerados em um estágio*), ou, como é mais comum, se faz nova seleção, tomando amostras de elementos dos conglomerados extraídos no primeiro estágio (*amostragem de conglomerados em dois estágios*). Todas as seleções devem ser aleatórias.

Outras formas de amostragem

Existem situações em que a seleção de uma amostra aleatória propriamente dita é muito difícil ou até mesmo impossível. Mas uma forma muito comum de extrair amostras em engenharia é através da *amostragem acidental* ou *a esmo*. Exemplos:

- a) para avaliar a qualidade em lotes de 1 kg de pregos, podemos examinar n pregos de cada lote, extraídos a esmo;
- b) para avaliar a qualidade em um carregamento de laranjas, podemos examinar n caixas do carregamento, extraídas a esmo. Se a unidade observacional é a laranja, em cada caixa podemos extrair, a esmo, m laranjas.

A seleção acidental nas situações anteriores parece produzir uma amostra com as mesmas propriedades da amostragem aleatória simples. Contudo, nem sempre isso acontece. Por exemplo, se numa pesquisa de mercado as pessoas forem entrevistadas acidentalmente nas ruas mais movimentadas da cidade, então as pessoas que costumam passar por essas ruas têm maior chance de ser selecionadas e, em consequência, pode haver viés nos resultados da pesquisa. O conhecimento sobre o problema em estudo é fundamental no planejamento de uma amostragem acidental.

Em muitas situações, lidamos com *populações infinitas*. Nesse caso, procuramos realizar as n observações de forma independente e sob as mesmas condições. Exemplos:

- a) para avaliar o desempenho de um sistema computacional em determinada condição, realizam-se n ensaios independentes na condição preestabelecida;
- b) para avaliar a qualidade de itens que saem de uma linha de produção, observa-se um item a cada 15 minutos.²

Quando examinamos características de um material contínuo, procuramos sortear – ou escolher a esmo – as posições em que serão coletadas as amostras. É o caso de se examinar um carregamento de argila que chega em uma fábrica de cerâmicas. Em materiais líquidos, costumamos homogeneizar a solução antes da amostragem.

2.2.2 Tamanho da Amostra

Quando se fala em amostragem, a pergunta natural é qual deve ser o tamanho necessário da amostra. Mas a resposta não é simples. O cálculo do tamanho da amostra, para o caso da amostragem aleatória simples, que é a situação mais fácil, será discutido no Capítulo 7. Contudo, gostaríamos de chamar a atenção desde já para algumas questões básicas.

Um fator importante na determinação do tamanho da amostra é a *variabilidade* da população em termos da variável em estudo. Por exemplo, uma amostra de sangue pode ser bem pequena, pois o sangue é razoavelmente *homogêneo* em nosso corpo. Por outro lado, populações com variâncias grandes exigem amostras maiores.

Outra questão importante é a relação entre tamanho da população (N) e tamanho da amostra (n). Considerando uma precisão desejada para as estimativas de interesse, a relação entre N e n não é linear (veja a Figura 2.4).

² Se os itens forem extraídos em sequência, pode haver dependência entre as características dos itens.

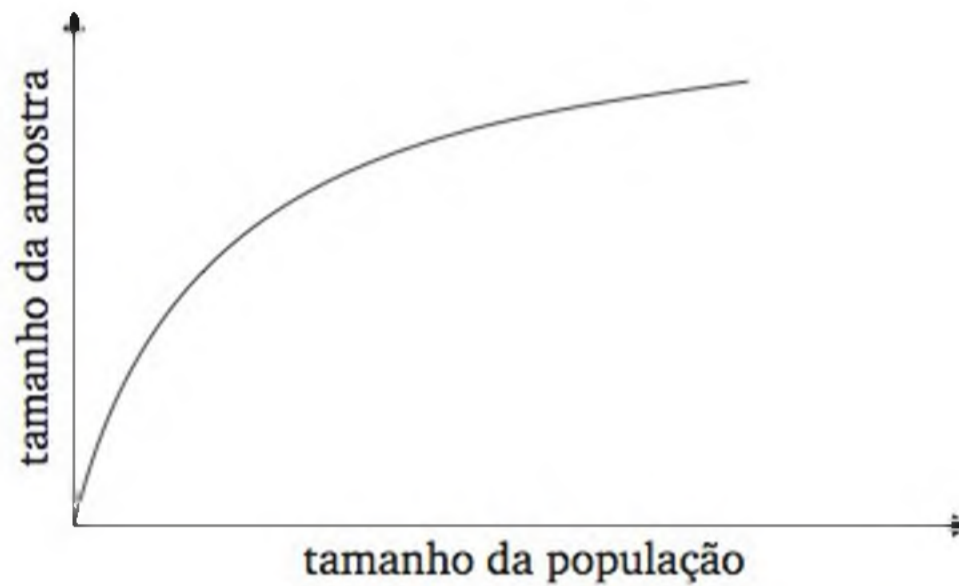


Figura 2.4 *Relação entre tamanho da população e tamanho da amostra.*

Como consequência da não linearidade entre N e n , devemos ter um cuidado adicional quando temos interesse em estudar separadamente certos subgrupos da população. Por exemplo, numa pesquisa eleitoral, podemos estar interessados não só nos resultados de todo o país, mas também em resultados por região demográfica. Nesse caso, as inferências por região normalmente carregam um erro maior do que as inferências de todo o país.

EXERCÍCIOS

1. Considerando a população apresentada a seguir, extraia uma amostra aleatória simples de $n = 6$ funcionários. Use a segunda linha da tabela de números aleatórios.

01. Aristóteles	02. Anastácia	03. Arnaldo	04. Bartolomeu	05. Bernardino
06. Cardoso	07. Carlito	08. Cláudio	09. Ermílio	10. Ercílio
11. Ernestino	12. Endevaldo	13. Francisco	14. Felício	15. Fabrício
16. Geraldo	17. Gabriel	18. Getúlio	19. Hiraldo	20. João da Silva
21. Joana	22. Joaquim	23. Joaquina	24. José da Silva	25. José de Souza
26. Josefa	27. Josefina	28. Maria José	29. Maria Cristina	30. Mauro
31. Paula	32. Paulo César			

2. Usando a terceira linha da tabela de números aleatórios, extraia uma amostra aleatória simples de quatro letras do alfabeto da língua portuguesa.
3. Os elementos de certa população estão dispostos numa lista, cuja numeração vai de 1.650 a 8.840. Descreva como você usaria uma tabela de números aleatórios para obter uma amostra de 100 elementos. Seria necessário efetuar nova numeração?

4. Seja um conjunto de 20 corpos de prova numerados de 1 a 20. Usando uma tabela de números aleatórios, divida aleatoriamente esses corpos de prova em dois grupos de dez elementos.
5. Selecione uma amostra estratificada uniforme, de tamanho $n = 12$, da população do Exercício 1.
6. Considerando a população de funcionários do exercício 1, faça uma amostragem estratificada proporcional de tamanho $n = 8$, usando a variável *sexo* para a formação dos estratos.
7. Comente os seguintes planos de amostragens, apontando suas incoerências, quando for o caso.
 - a) Com a finalidade de estudar o perfil dos consumidores de um supermercado, observaram-se os consumidores que compareceram ao supermercado no primeiro sábado do mês.
 - b) Com a finalidade de estudar o perfil dos consumidores de um supermercado, fez-se a coleta de dados durante um mês, tomando a cada dia um consumidor da fila de cada caixa do supermercado, variando sistematicamente o horário da coleta dos dados.
 - c) Para avaliar a qualidade dos itens que saem de uma linha de produção, observaram-se todos os itens das 14 às 14h30min.
 - d) Para avaliar a qualidade dos itens que saem de uma linha de produção, observou-se um item a cada meia hora, durante todo o dia.
 - e) Para estimar a percentagem de empresas que investiram em novas tecnologias no último ano, enviou-se um questionário a todas as empresas. A amostra foi formada pelas empresas que responderam ao questionário.

2.3 PLANEJAMENTO DE EXPERIMENTOS

Na área tecnológica, são muito comuns pesquisas experimentais, nas quais se manipulam de forma planejada certas *variáveis independentes* ou *fatores* (A , B , C , ...), para verificar o efeito que essa manipulação provoca numa certa *variável dependente* ou *resposta* Y .

Exemplos

- a) Verificar quais são os fatores que mais interferem na resistência à compressão (Y) de um concreto. Os fatores a serem estudados podem ser:

- tempo de hidratação (*A*);
 - dosagem de cimento (*B*);
 - qualidade do cimento (*C*);
 - uso de aditivos (*D*).
- b) Encontrar a melhor condição de operação de um processo químico. A resposta *Y* pode ser o rendimento da reação química e os fatores podem ser:
- tempo de reação (*A*);
 - temperatura da reação (*B*).
- c) Uma empresa de informática quer verificar o tipo de equipamento adequado ao usuário. A resposta *Y* pode ser o tempo de resposta e os fatores podem ser:
- processador (*A*);
 - quantidade de memória RAM (*B*);
 - quantidade de memória fixa (*C*);
 - tipo de carga de trabalho a ser executada (*D*).³

Quando se conhece pouco sobre o problema em estudo, é comum buscar a *caracterização do processo*, isto é, pesquisar quais são os fatores que provocam maiores alterações na resposta. Geralmente, esses estudos iniciam com grande número de fatores e, a partir de um estudo experimental planejado, faz-se uma triagem de fatores, escolhendo os mais significativos. Em outros casos, já se conhecem os fatores mais significativos e busca-se a combinação de níveis dos fatores que levam à melhor resposta possível. É o que chamamos de *otimização do processo*.

Estratégias no planejamento de experimentos

No planejamento de um experimento, devemos:

- reconhecer, estabelecer e delimitar claramente o problema;
- identificar os possíveis fatores que podem afetar o problema em estudo;
- verificar quais fatores poderão ser mantidos fixos e, portanto, não terão seus efeitos avaliados no estudo experimental;

3 Cargas de trabalho são as solicitações feitas pelos usuários de um sistema. Por exemplo, as cargas de trabalho de uma CPU são as instruções a serem executadas; as cargas de trabalho de um banco de dados são as consultas dos usuários etc.

- identificar, para cada fator, o intervalo de variação e os níveis que serão estudados;
- escolher um projeto experimental adequado, isto é, saber como combinar os níveis dos fatores de forma que se possa resolver o problema proposto com o menor custo possível;
- escolher a resposta adequada, ou seja, a variável Y que mede adequadamente o resultado (a qualidade, o desempenho etc.) do processo;
- planejar como será a análise dos dados do experimento.⁴

Exemplo 2.1 Considere o problema de operar adequadamente o protótipo de uma catapulta romana em tamanho reduzido. O objetivo é lançar os projéteis a uma distância especificada. O operador pode controlar três itens da catapulta: altura do pivô, comprimento do braço e ângulo de parada. Veja a Figura 2.5.⁵

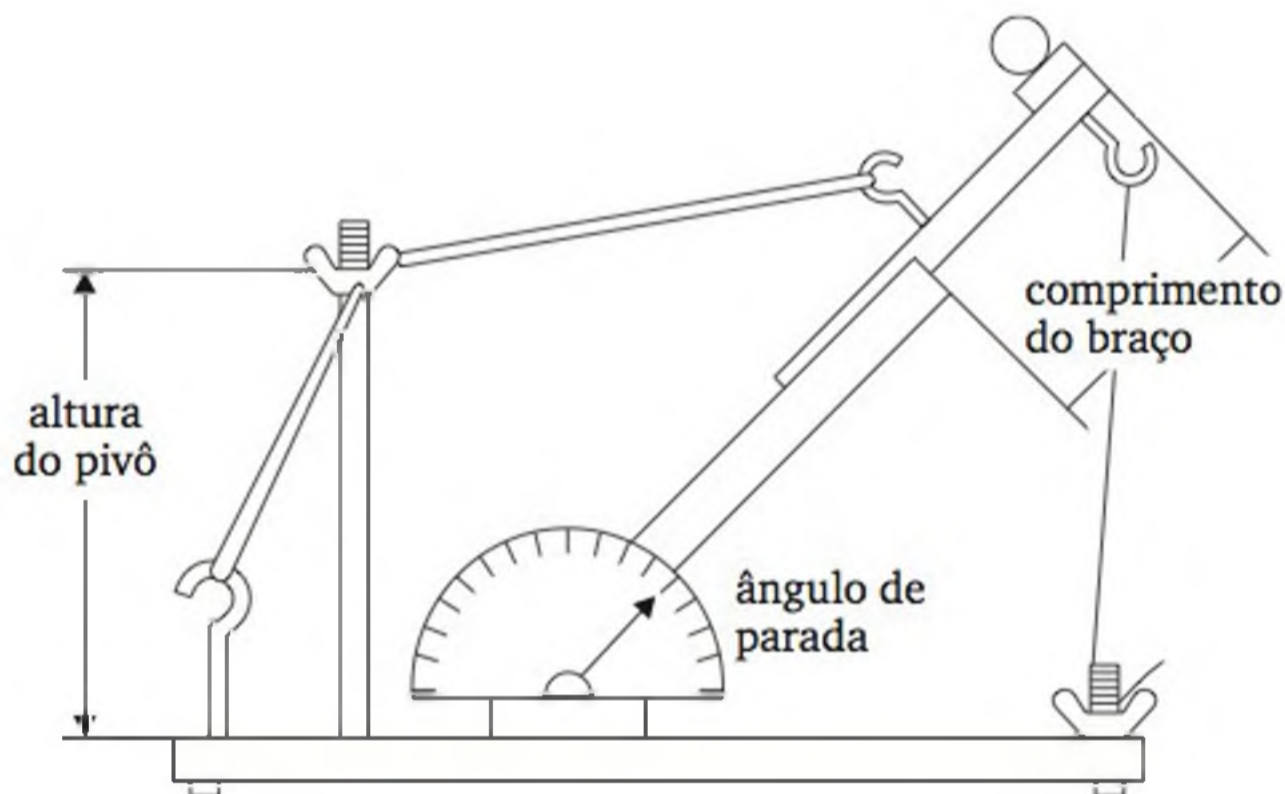


Figura 2.5 Protótipo reduzido de uma catapulta romana (Exemplo 2.1).

Um experimento pode ser planejado para verificar as condições de operação da catapulta que a levam a atingir seu objetivo. Parece natural que os seguintes fatores podem alterar o resultado do processo:

- altura do pivô (A);
- comprimento do braço (B);
- ângulo de parada (C).

⁴ Eventualmente, pode haver mais que uma resposta, mas estudaremos uma por vez.

⁵ Ilustração extraída do artigo de LUNER, J. J. *Quality Engineering*, v. 6, nº 4, p. 691-705, 1994.

A região de variação desses fatores deve ser determinada por conhecimento prévio do processo. Se não houver conhecimento suficiente, podemos optar por uma variação grande e, em experimentos posteriores, reduzir a variação. Também parece razoável supor que, ao variar cada um desses fatores de seu nível mínimo até seu nível máximo, o desempenho do processo deve melhorar até certo ponto e, depois, piorar, caracterizando efeitos não lineares. Assim, é recomendável usar pelo menos três níveis em cada fator, tais como mostrados na tabela a seguir:

Fator	Nível inferior	Nível intermediário	Nível superior
A	6,35 cm	10,16 cm	13,97 cm
B	0,81 cm	5,08 cm	9,35 cm
C	30°	45°	90°

Podem ser identificados outros fatores, tais como pequenas variações no peso do projétil, variações na banda de borracha, fadiga etc. Como esses fatores são muito difíceis de ser controlados na operação natural do processo, eles também não foram controlados no experimento. Devido à presença de fatores não controláveis – o que é comum na grande maioria dos experimentos –, torna-se fundamental usar *replicações* e *aleatorização*. Esses procedimentos serão descritos ainda nesta seção.

Um possível projeto experimental para este exemplo é o projeto fatorial, em que cada nível de um fator é *cruzado* com todos os níveis dos outros fatores, conforme será discutido na seção 2.3.4.

A variável dependente, ou *resposta*, deve avaliar o desempenho ou qualidade do processo. No presente exemplo, ela pode ser a distância entre a queda do projétil e o alvo pretendido.

A Figura 2.6 apresenta um esquema geral de um problema de planejamento de experimento.

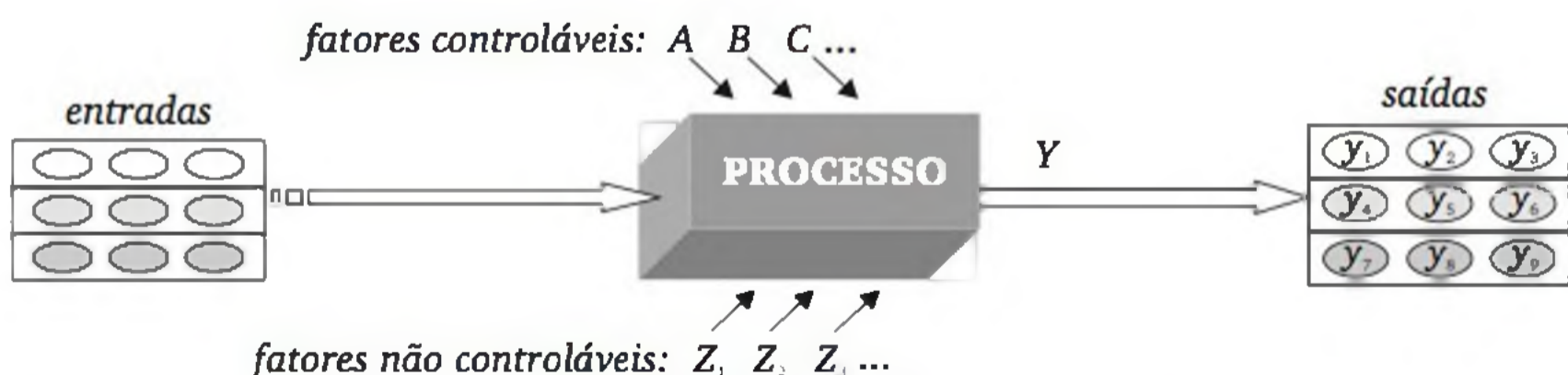


Figura 2.6 Esquema geral de um problema de planejamento de experimento.

Escolha das unidades experimentais e uso de blocos

As *entradas* de um processo são as unidades experimentais que serão estudadas. Em engenharia, as entradas costumam ser *corpos de prova*; em medicina, são os *indivíduos*; em informática, são as *cargas de trabalho*; e assim por diante. É desejável fazer com que as entradas sejam tão homogêneas quanto possível, mas, se isso não é possível ou se o estudo quer extrair conclusões para grupos heterogêneos, devemos tentar construir *blocos* relativamente homogêneos. Por exemplo, se os corpos de prova vêm de diferentes lotes, cada lote pode constituir um bloco. No exemplo da catapulta, se os projéteis têm diferentes formas ou pesos, podemos tentar formar grupos (blocos) de projéteis similares. Veja a Figura 2.7.

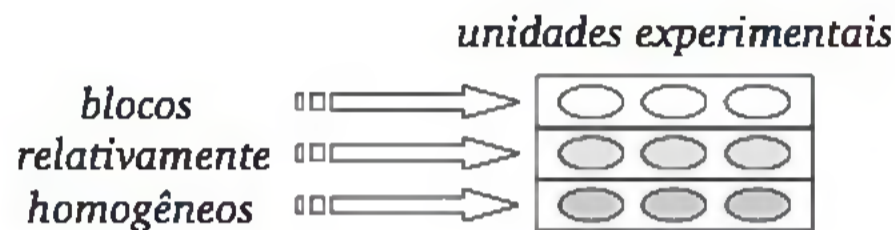


Figura 2.7 Formação de blocos para garantir homogeneidade nas entradas.

Em agronomia, costumamos formar blocos de canteiros homogêneos. Na Ciência da Computação, ao comparar diferentes sistemas computacionais, podemos usar o mesmo conjunto de cargas de trabalho em todos os sistemas em comparação. As observações associadas a uma particular carga de trabalho podem ser consideradas como pertencentes a um mesmo bloco.

Os fatores

Os fatores de um estudo experimental são os fatores controláveis do processo que podem afetar seu desempenho. Algumas vezes, incluímos também fatores não controláveis no processo, mas que podem ser controlados em laboratório. A inclusão de fatores não controláveis é importante quando se busca produzir *produtos robustos*, isto é, produtos pouco sensíveis às variações de fatores ambientais ou de outros ruídos.

Tratamentos

Chamamos de *tratamento* uma particular combinação de níveis dos fatores incluídos no modelo do estudo experimental. No exemplo da catapulta, a combinação de todos os fatores no primeiro nível (altura do pivô em 6,35 cm, altura do braço em 0,81 cm e ângulo de parada em 30°) forma um possível tratamento. Quando usamos blocos, devemos tentar fazer com que em cada bloco sejam realizados todos os tratamentos em estudo.

Replicações

Normalmente, realizamos mais de um ensaio em cada condição experimental (tratamento). Ou seja, realizamos *replicações*. Com as replicações, podemos avaliar o *erro experimental*, isto é, o efeito provocado pelos possíveis fatores que estão agindo no processo, mas que não foram incluídos no estudo.

Aleatorização

Para garantir a validade de um estudo experimental, a alocação dos tratamentos nas unidades experimentais deve ser *aleatorizada*.⁶ Veja o Exemplo 2.2.

Exemplo 2.2 Desejamos estudar a produção por m^2 (Y), de certa cultura, considerando três níveis de dosagens (a , b e c) de certo fertilizante. Dispomos de seis canteiros para o experimento, donde podemos fazer duas replicações. Para aleatorizar o tratamento a ser aplicado em cada canteiro, podemos fazer uso de números aleatórios. A seguir, é reproduzida uma linha do quadro de números aleatórios da seção 2.1:

2450 0455 0324 1641 2196 0171 2850 3431 5536 3573 2913 8021 7889 6759 7553

Excluindo os números fora do conjunto $\{1, 2, 3, 4, 5, 6\}$ e os que se repetem, temos o seguinte projeto experimental para o problema em questão:

Tratamento:	a	a	b	b	c	c
Canteiro (ordem aleatória):	2	4	5	3	1	6

Se for identificado algum fator de heterogeneidade nos canteiros e se estes puderem ser agrupados em dois blocos relativamente homogêneos – digamos, bloco 1 formado pelos canteiros 1, 2 e 3 e bloco 2 pelos canteiros 4, 5 e 6 –, o esquema do projeto experimental ficaria assim:

Bloco:	1	1	1	2	2	2
Tratamento:	a	b	c	a	b	c
Canteiro:	2	1	3	4	5	6

(ordem aleatória em cada bloco)

⁶ Quando os ensaios são realizados em tempos diferentes, a ordem dos ensaios também deve ser aleatorizada.

2.3.1 Projetos com um fator

Embora o objetivo deste capítulo não seja mostrar técnicas de análise de dados experimentais, faremos algumas considerações, que serão usadas em projetos mais elaborados.

O Exemplo 2.2 mostrou um problema com apenas um fator (*dosagem de fertilizante*). Dois projetos experimentais foram ilustrados: *completamente aleatorizado* e *com blocos aleatorizados*. No momento, estudaremos com mais detalhes somente o primeiro caso, que deve produzir dados da seguinte forma:

Tratamento:	<i>a</i>	<i>a</i>	<i>b</i>	<i>b</i>	<i>c</i>	<i>c</i>
Canteiro (ordem aleatória):	2	4	5	3	1	6
Resposta:	y_{11}	y_{12}	y_{21}	y_{22}	y_{31}	y_{32}

onde y_{ij} representa o valor da produção por m^2 , considerando a j -ésima replicação do tratamento i ($i = 1, 2, 3; j = 1, 2$). O cálculo da *média amostral*, calculada para dado tratamento, fornece uma *estimativa* da produção esperada na condição experimental em questão. Em nosso exemplo:

$$\bar{y}_1 = \frac{y_{11} + y_{12}}{2}, \quad \bar{y}_2 = \frac{y_{21} + y_{22}}{2} \quad \text{e} \quad \bar{y}_3 = \frac{y_{31} + y_{32}}{2}$$

Conforme discutimos no Capítulo 1, é fundamental avaliar também a variabilidade do fenômeno em estudo. Assim, podemos calcular a *variância amostral* da produção em cada tratamento (ver seção 1.6), como segue:

$$s_1^2 = \frac{(y_{11} - \bar{y}_1)^2 + (y_{12} - \bar{y}_1)^2}{1}, \quad s_2^2 = \frac{(y_{21} - \bar{y}_2)^2 + (y_{22} - \bar{y}_2)^2}{1}, \quad \text{e}$$

$$s_3^2 = \frac{(y_{31} - \bar{y}_3)^2 + (y_{32} - \bar{y}_3)^2}{1}.$$

Note que, para avaliar a variabilidade, precisamos ter pelo menos duas observações em cada tratamento. Com duas observações, temos apenas 1 *grau de liberdade*; com n observações, temos $n - 1$ *graus de liberdade*. Os graus de liberdade são usados no denominador do cálculo das variâncias.

Geralmente, supomos que a variabilidade não se altera com os diferentes tratamentos. Assim, em nosso exemplo, podemos *agregar* as três variâncias amostrais, produzindo uma única medida de variabilidade com três graus de liberdade (soma dos graus de liberdade das três variâncias amostrais):

$$s_a^2 = \frac{s_1^2 + s_2^2 + s_3^2}{3}$$

No Capítulo 9, veremos que uma estimativa da variância é fundamental para avaliar se diferenças observadas entre as médias dos tratamentos são reais, ou podem ser meramente explicadas por variações casuais associadas ao erro experimental.

De maneira geral, em um projeto completamente aleatorizado com um fator, tendo g tratamentos (unidades experimentais divididas aleatoriamente em g grupos) e n replicações em cada tratamento, temos:

$$\bar{y}_i = \frac{y_{i1} + y_{i2} + \dots + y_{in}}{n} = \frac{1}{n} \sum_{j=1}^n y_{ij} \quad (2.1)$$

$$s_i^2 = \frac{1}{n-1} \sum_{j=1}^n (y_{ij} - \bar{y}_i)^2 \quad (i = 1, 2, \dots, g) \quad (2.2)$$

E a *variância agregada*, que mede a variabilidade devida ao erro experimental, é dada por:

$$s_a^2 = \frac{s_1^2 + s_2^2 + \dots + s_g^2}{g} \quad (2.3)$$

A variância agregada tem $g(n-1)$ graus de liberdade, pois são agregadas g variâncias com $n-1$ graus de liberdade em cada uma.

Se, na realização do experimento do Exemplo 2.4, obtivermos os seguintes dados:

Tratamento:	a	a	b	b	c	c
Respostas:	$y_{11} = 12$	$y_{12} = 14$	$y_{21} = 16$	$y_{22} = 17$	$y_{31} = 15$	$y_{32} = 15$

Temos as seguintes médias e variâncias associadas a essa amostra:

$$\text{Médias:} \quad \bar{y}_1 = 13 \quad \bar{y}_2 = 16,5 \quad \bar{y}_3 = 15$$

$$\text{Variâncias:} \quad s_1^2 = 2 \quad s_2^2 = 0,5 \quad s_3^2 = 0$$

$$\text{Variância agregada: } s_a^2 = \frac{s_1^2 + s_2^2 + s_3^2}{3} = 0,833.$$

2.3.2 Projetos fatoriais

Os experimentos normalmente envolvem vários fatores. Uma forma eficiente de combinar os níveis dos diversos fatores é *cruzando-os*, de tal forma que cada nível de um fator seja combinado com todos os níveis dos outros fatores.

Exemplo 2.3 Considere um processo químico em que se quer pesquisar a influência dos fatores:

- o tipo de catalisador (A), com dois níveis (a_1, a_2) e
- o tempo de reação (B), com quatro níveis (b_1, b_2, b_3, b_4),

sobre o rendimento da reação química (Y).

Um projeto fatorial consiste em considerar $2 \times 4 = 8$ tratamentos, conforme mostrado na Figura 2.8.

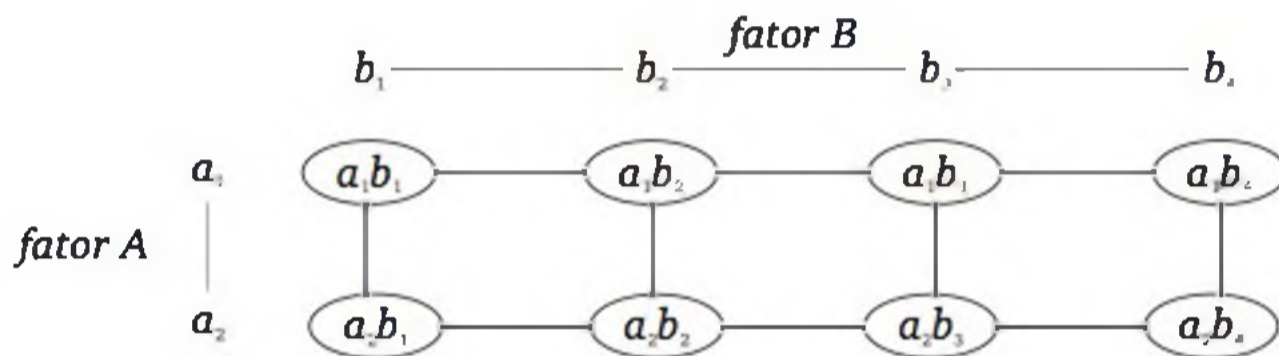


Figura 2.8 Esquema de um projeto fatorial 2×4 .

No exemplo da catapulta (Exemplo 2.1), em que temos três fatores ensaiados a três níveis, o projeto fatorial englobaria $3 \times 3 \times 3 = 3^3 = 27$ ensaios. Veja a Figura 2.9.

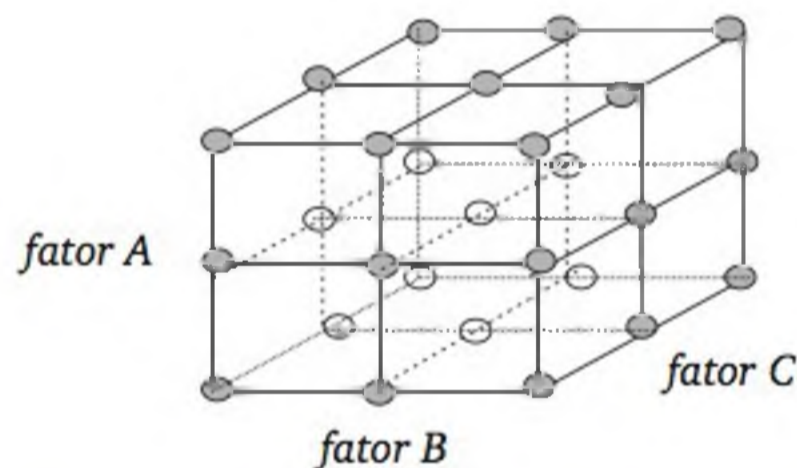


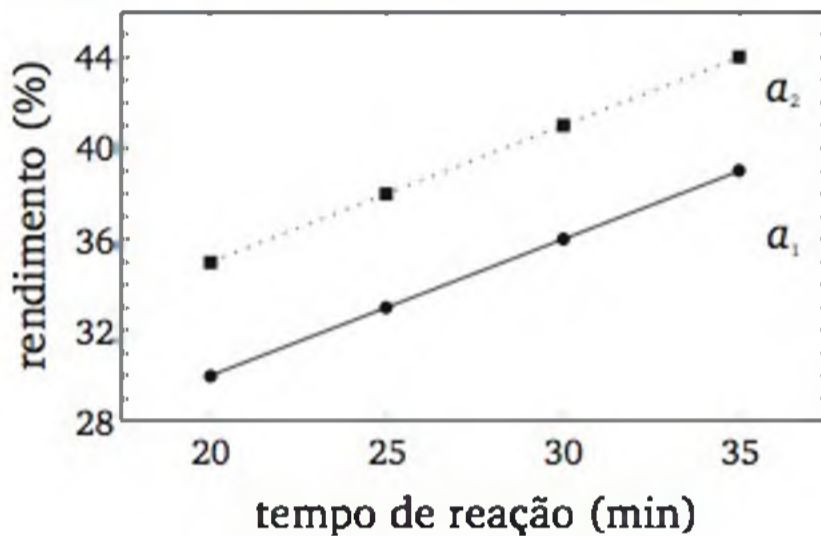
Figura 2.9 Representação geométrica de um projeto fatorial 3^3 .

Nos projetos fatoriais, podemos avaliar não somente o efeito médio de cada fator, mas também possíveis *interações* entre os fatores. Dizemos que existe *interação* entre dois fatores quando a diferença na resposta entre os níveis de um fator não é a mesma para todos os níveis do outro fator.

A Figura 2.10 mostra dois possíveis resultados associados ao experimento descrito no Exemplo 2.3. No primeiro, os fatores tipo de catalisador e tempo de reação agem aditivamente sobre o rendimento da reação química, permitindo analisar cada fator separadamente: a mudança no catalisador altera o rendi-

mento em cinco unidades e, a cada variação de nível do tempo de reação, há alteração em torno de três unidades no rendimento. Já o gráfico (b) mostra uma situação com *interação*. Só é possível analisar o efeito de um fator condicionado ao outro.

(a) Dois fatores agindo aditivamente



(b) Interação entre dois fatores

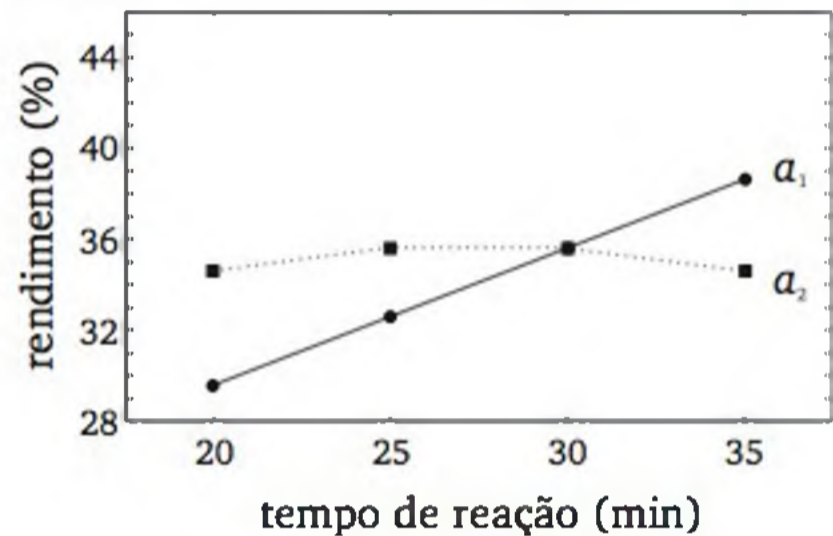


Figura 2.10 Dois possíveis perfis de médias para o experimento do Exemplo 2.3.

Na engenharia, era muito comum realizar experimentos, variando um fator de cada vez. Esse procedimento não permitia verificar possíveis interações entre os fatores e, muitas vezes, levava a conclusões errôneas.

2.3.3 Projetos com muitos fatores

Em muitos problemas, o número de fatores que possivelmente alteram a resposta é grande. Por exemplo, para verificar quais são os fatores que mais interferem na resistência à compressão de um concreto (Y), podemos citar:

- tempo de hidratação (A);
- a dosagem de cimento (B);
- a qualidade do cimento (C);
- o uso de aditivos (D).

Se forem usados três níveis em cada fator, teremos $3^4 = 81$ ensaios, sem contar com replicações adicionais, o que tende a encarecer demasiadamente o experimento.

Projetos fatoriais 2^k

Consiste em ensaiar cada um dos k fatores em apenas dois níveis. Além de evitar grande número de ensaios, esse tipo de projeto é relativamente fácil nos

aspectos computacionais e interpretativos. A Figura 2.11 representa geometricamente os 16 pontos a serem ensaiados em um projeto para avaliar quatro fatores.

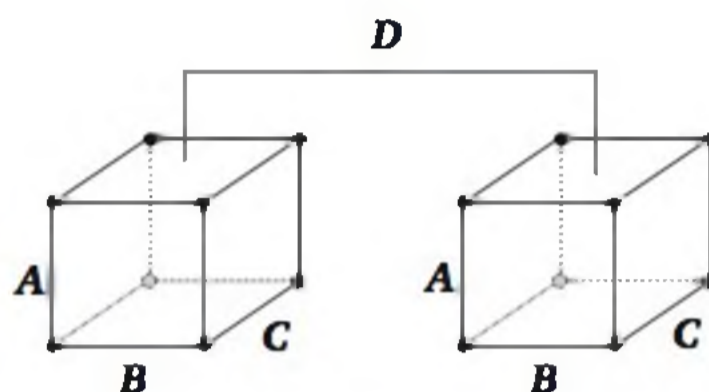
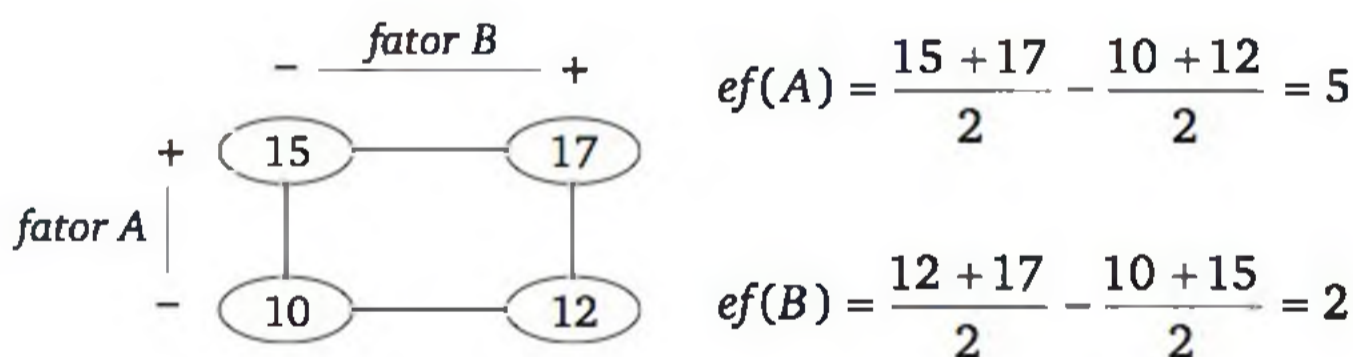


Figura 2.11 Representação geométrica de um projeto fatorial 2^4 .

Projetos do tipo 2^k são largamente utilizados para *caracterizar processos*, ou seja, para selecionar fatores que agem significativamente sobre a resposta. Muitas vezes, quando se deseja *otimizar o processo*, realiza-se, inicialmente, um projeto 2^k para selecionar os fatores mais significativos e, depois, realizam-se novos experimentos.

Denotaremos por “+1”, ou simplesmente por “+”, o nível superior de um fator; e por “-1” ou “-” o nível inferior.⁷ Definimos como *efeito principal* do fator a diferença média na resposta quando se passa do nível -1 para o nível +1. Veja o exemplo a seguir, relativo a um projeto 2^2 , com os resultados de Y em cada condição experimental:



Note que, se o fator B estiver no nível -1, ao variar A de -1 para +1, a variação na resposta é de 5 unidades. Se B estiver no nível +1, ao variar A de -1 para +1, a variação na resposta também é de 5 unidades. Ou seja, o presente exemplo *não tem interação*. De modo geral, podemos calcular o efeito da interação entre dois fatores (A e B) através da diferença da média de quando ambos estão no mesmo nível (-1 ou +1) em relação à média de quando eles estão em níveis trocados. No exemplo em discussão, temos:

⁷ Para fatores qualitativos, tal como o tipo de catalisador (a_1 ou a_2), os códigos + e - podem ser colocados em quaisquer dos níveis dos fatores.

$$ef(AB) = \frac{10 + 17}{2} - \frac{12 + 15}{2} = 0$$

Para facilitar o cálculo dos efeitos, podemos organizar os resultados de Y numa tabela que considera todas as combinações dos fatores, indicando, para cada fator, seu nível (– ou +). Segue a *tabela dos sinais* para um projeto 2^2 , incluindo a coluna de elementos unitários I , associada à média global, em que todos os sinais são positivos; e a coluna da interação AB , em que os sinais são obtidos através da multiplicação de elemento a elemento das colunas A e B .

Condição experimental	I	A	B	AB	Y
1	+	–	–	+	?
2	+	–	+	–	?
3	+	+	–	–	?
4	+	+	+	+	?

Ao realizar o experimento, temos os valores para a resposta Y em cada condição experimental. O cálculo de qualquer efeito (principal ou de interação) pode ser feito por:

$$ef = \bar{y}_+ - \bar{y}_- \quad (2.4)$$

onde:

\bar{y}_+ é a média das observações de Y nas condições experimentais em que o sinal do fator é positivo e

\bar{y}_- é a média das observações de Y nas condições experimentais em que o sinal do fator é negativo.

Exemplo 2.4 Um estudo foi desenvolvido para verificar os fatores que influenciam a qualidade da transmissão de dados através da porta serial de microcomputadores. Propositalmente, foram usados cabos com comprimento bastante superior às especificações técnicas. Observou-se a taxa de falhas de transmissão em função dos fatores: (A) velocidade da transmissão (2.400/9.600 *bauds*), (B) tamanho do arquivo (100/200 *bytes*) e (C) comprimento do cabo serial (15/20 m). Os resultados do experimento, que foi realizado com 2 replicações, foram:⁸

⁸ O sinal negativo indica o nível inferior do fator e o sinal positivo indica o nível superior do fator. Este exemplo é parte de um trabalho de disciplina do acadêmico Maximiliano Pezzin, Curso de Pós-Graduação em Ciência da Computação – UFSC, 2001.

Condição experimental	A	B	C	Rep. 1 y	Rep. 2 y
1	-	-	-	32,5	32,3
2	-	-	+	35,7	35,9
3	-	+	-	33,1	33,4
4	-	+	+	35,9	36,1
5	+	-	-	34,1	34,4
6	+	-	+	36,6	36,9
7	+	+	-	34,2	34,9
8	+	+	+	37,1	36,9

A tabela seguinte inclui os sinais para o cálculo da média global (I) e de todas as interações possíveis (AB , AC , BC e ABC). Os sinais de cada interação correspondem aos sinais da multiplicação de elemento a elemento dos vetores nela envolvidos.

Condição experimental	I	A	B	C	AB	AC	BC	ABC	Rep. 1 y	Rep. 2 y
1	+	-	-	-	+	+	+	-	32,5	32,3
2	+	-	-	+	+	-	-	+	35,7	35,9
3	+	-	+	-	-	+	-	+	33,1	33,4
4	+	-	+	+	-	-	+	-	35,9	36,1
5	+	+	-	-	-	-	+	+	34,1	34,4
6	+	+	-	+	-	+	-	-	36,6	36,9
7	+	+	+	-	+	-	-	-	34,2	34,9
8	+	+	+	+	+	+	+	+	37,1	36,9

Para calcular o efeito do fator A, $ef(A)$, calculamos a média das observações em que A tem sinal positivo e a média das observações em que A tem sinal negativo, ou seja:

$$\bar{y}_+ = \frac{34,1 + 34,4 + 36,6 + 36,9 + 34,2 + 34,9 + 37,1 + 36,9}{8} = 35,6375$$

$$\bar{y}_- = \frac{32,5 + 32,3 + 35,7 + 35,9 + 33,1 + 33,4 + 35,9 + 36,1}{8} = 34,3625$$

Assim,

$$ef(A) = 35,6375 - 34,3625 = 1,275$$

ou seja, quando a velocidade de comunicação passa de 2.400 *bauds* para 9.600 *bauds*, estimamos que a taxa média de falhas aumenta em 1,275 pontos. De forma análoga, podemos calcular todos os demais efeitos, cujos resultados são mostrados abaixo:

$$\begin{aligned}
 \text{média global} &= 35,000 \\
 \text{ef}(A) &= 1,275 & \text{ef}(AB) &= - 0,125 \\
 \text{ef}(B) &= 0,400 & \text{ef}(AC) &= - 0,300 \\
 \text{ef}(C) &= 2,775 & \text{ef}(BC) &= - 0,175 \\
 & & \text{ef}(ABC) &= 0,150
 \end{aligned}$$

Nesse exemplo, os efeitos das interações são bastante pequenos, o que permite interpretar o efeito de cada fator isoladamente.

Projetos fatoriais fracionados 2^{k-p}

Quando temos muitos fatores, mesmo que sejam ensaiados em apenas dois níveis, a quantidade de tratamentos torna-se muito grande. Com oito fatores, por exemplo, temos $2^8 = 256$ tratamentos. A ideia dos projetos fracionados é ensaiar apenas parte das possíveis combinações, mas planejado de tal forma que garanta a possibilidade de estimar os efeitos principais e, às vezes, as interações entre dois fatores.

Os projetos fracionados são particularmente importantes quando se pretende fazer uma *triagem* de fatores para serem usados em estudos posteriores.

Um bom fracionamento pode ser feito realizando-se os ensaios cujo sinal de I é igual ao sinal da interação com todos os k fatores. Ou, equivalentemente, realizando-se os ensaios do grupo complementar. Nesse caso, temos um projeto 2^k com 1 fracionamento, que denotaremos por 2^{k-1} . No projeto 2^3 , por exemplo, podemos realizar os ensaios em que vale a relação $I = ABC$ ou os ensaios da relação $I = -ABC$. Esses dois grupos de condições experimentais são indicados na tabela de sinais, a seguir:

Condição experimental	I	A	B	C	AB	AC	BC	ABC
1	+	-	-	-	+	+	+	-
2	+	-	-	+	+	-	-	+
3	+	-	+	-	-	+	-	+
4	+	-	+	+	-	-	+	-
5	+	+	-	-	-	-	+	+
6	+	+	-	+	-	+	-	-
7	+	+	+	-	+	-	-	-
8	+	+	+	+	+	+	+	+

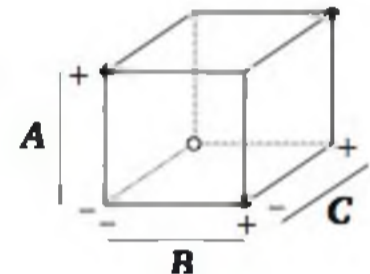
$I = ABC$

$I = -ABC$

Fazendo somente os ensaios que satisfazem à relação $I = ABC$, temos:

Condição experimental	I	A	B	C	AB	AC	BC	ABC
1) 2	+	-	-	+	+	-	-	+
2) 3	+	-	+	-	-	+	-	+
3) 5	+	+	-	-	-	-	+	+
4) 8	+	+	+	+	+	+	+	+

Note, pela representação geométrica ao lado, que, se os pontos ensaiados forem projetados em qualquer dos três planos, passamos a ter um projeto 2^2 completo.



Observando a tabela de sinais do projeto 2^{3-1} , verificamos que algumas colunas são iguais. Em especial: $A = BC$, $B = AC$ e $C = AB$. Logo, com esse projeto, não é possível distinguir os efeitos principais das interações entre dois fatores – existe *confusão* entre esses efeitos. Conseqüentemente, o projeto 2^{3-1} só pode ser usado nas situações em que se supõe não haver interações ou que os efeitos das interações possam ser negligenciados.

Para construir um projeto 2^{3-1} , podemos fazer um projeto 2^2 completo e, depois, incluir o fator C com sinais equivalentes a AB . Em geral, na construção de um projeto 2^{k-1} , fazemos um projeto completo com $k - 1$ fatores e, depois, incluímos o último fator com os mesmos sinais da interação dos $k - 1$ fatores previamente definidos.

Veja, agora, a tabela de sinais de um projeto 2^{4-1} , com a relação $I = ABCD$.⁹ Foram incluídas as colunas relativas aos efeitos principais e às *interações de segunda ordem*, isto é, entre dois fatores. As colunas relativas às *interações de ordem superior* (entre três ou mais fatores) não foram incluídas na tabela.

Condição experimental	I	A	B	C	D	AB	AC	AD	BC	BD	CD
1	+	-	-	-	-	+	+	+	+	+	+
2	+	-	-	+	+	+	-	-	-	-	+
3	+	-	+	-	+	-	+	-	-	+	-
4	+	-	+	+	-	-	-	+	+	-	-
5	+	+	-	-	+	-	-	+	+	-	-
6	+	+	-	+	-	-	+	-	-	+	+
7	+	+	+	-	-	+	-	-	-	-	-
8	+	+	+	+	+	+	+	+	+	+	+

⁹ A construção da tabela dos sinais desse projeto pode ser feita através de todas as combinações dos fatores A , B e C , como no projeto 2^3 completo. Para os sinais de D , faz-se $D = ABC$.

Observamos que as colunas A , B , C e D não se igualam com colunas de interação de segunda ordem, mas existem algumas igualdades entre estas (por exemplo, $AD = BC$).¹⁰ Consequentemente, o projeto 2^{4-1} é uma boa alternativa quando se deseja avaliar apenas os efeitos principais.

Em um projeto 2^{5-1} , não existe confusão entre os efeitos principais nem entre interações que envolvem dois fatores. Ou seja, cinco fatores podem ser estudados tranquilamente com $2^{5-1} = 16$ pontos experimentais. Com seis ou mais fatores, é possível fazer dois fracionamentos (ver o Exercício 11).

EXERCÍCIOS

8. Apresente as 32 combinações de sinais em que os fatores A , B , C , D e E devem ser ensaiados em um projeto 2^5 completo. Anote os ensaios que devem ser realizados em um projeto 2^{5-1} , considerando a relação $I = ABCDE$. Repare que você pode construir o mesmo projeto fazendo, inicialmente, um projeto 2^4 completo e, depois, inserindo a coluna E com a relação $E = ABCD$.
9. Calcule a variância agregada do experimento do Exemplo 2.4. Quantos graus de liberdade estão associados a essa medida?
10. Para avaliar o efeito dos fatores: (A) tempo de hidratação (14 dias/28 dias), (B) relação água/cimento (0,38 e 0,58) e (C) tipo de cimento (comum e pozolânico) na resistência à compressão de um concreto (Y), realizou-se um experimento cujos resultados da resistência (em MPa) são apresentados a seguir:¹¹

Tipo de cimento	Relação água/cimento	Tempo de hidratação	
		14 dias	28 dias
comum	0,38	23,1	42,2
	0,58	12,0	27,9
pozolânico	0,38	24,3	39,5
	0,58	11,1	24,3

Calcule os efeitos principais e as interações de segunda ordem.

¹⁰ No projeto 2^{4-1} , os efeitos principais estão confundidos com interações de terceira ordem (entre três fatores), mas os efeitos destas interações geralmente são negligenciáveis.

¹¹ O exercício é parte de um trabalho de disciplina dos acadêmicos F. Pelisser e L. Matos do Programa de Pós-Graduação em Engenharia Civil – UFSC, 2001. Os dados são hipotéticos, mas foram baseados em HELENE, P. R. L.; TERZIAN, P. *Manual de dosagem e controle do concreto*. São Paulo: PINI, 1995.

11. Em *Applied Statistics*, v. 42, nº 4, p. 671-681 (1993), M. G. Tuck e J. I. L. Cottrell realizaram vários experimentos para obter uma farinha de pão de melhor qualidade. Misturaram à farinha de trigo pequenas quantidades de ingredientes permitidos. Os fatores correspondem à quantidade de cada ingrediente adicionado à farinha, sendo que o nível inferior corresponde à ausência do ingrediente. Parte de um dos experimentos foi realizada sob um projeto 2^{6-2} , em que os fatores *A*, *B*, *C* e *E* formaram um projeto 2^4 completo. Os sinais do fator *D* foram obtidos através da relação $D = ABC$ e os sinais do fator *F* foram obtidos através da relação $F = BCE$. A resposta (*Y*) foi o volume médio dos pães. Em cada condição experimental, realizaram-se quatro ensaios. A tabela, a seguir, apresenta a média \bar{y} e o desvio padrão *s* do volume específico nas quatro replicações em cada combinação dos níveis dos fatores.

Condição experimental	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>	\bar{y}	<i>s</i>
1	-	-	-	-	-	-	429,25	75,39
2	-	-	-	-	+	+	433,00	69,40
3	-	-	+	+	-	+	454,25	88,99
4	-	-	+	+	+	-	456,75	82,24
5	-	+	-	+	-	+	446,75	74,09
6	-	+	-	+	+	-	447,75	80,93
7	-	+	+	-	-	-	455,50	89,58
8	-	+	+	-	+	+	448,25	74,24
9	+	-	-	+	-	-	458,75	79,47
10	+	-	-	+	+	+	449,50	84,58
11	+	-	+	-	-	+	463,75	91,67
12	+	-	+	-	+	-	466,00	88,99
13	+	+	-	-	-	+	449,50	88,88
14	+	+	-	-	+	-	452,75	98,27
15	+	+	+	+	-	-	469,00	82,30
16	+	+	+	+	+	+	471,50	75,11

- a) Calcule os efeitos principais dos seis fatores. Quais provocam maiores variações no nível médio da resposta? Em que níveis eles devem ser fixados para maximizar o volume dos pães?
- b) Calcule os efeitos principais em termos do desvio padrão *s*. Qual fator provoca maior alteração na variabilidade da resposta? Se o objetivo é minimizar a variabilidade, em qual nível este fator deve ser fixado?

Análise Exploratória de Dados

Com o advento da informática, o mundo encheu-se de dados. As empresas têm dados de suas atividades, de seus funcionários, de seus clientes etc. Mas para que esses dados sejam informativos, necessitamos organizá-los, resumí-los e apresentá-los de forma adequada. Este é o papel da *estatística descritiva*.

Como ilustração, considere o problema da pasteurização de leite em um laticínio, em que os engenheiros estão preocupados com a variação da temperatura do pasteurizador. Para se conhecer o nível dessa variação, foram empregadas 1.389 leituras de temperatura, realizadas durante dois meses.¹ A Figura 3.1 mostra parte do conjunto das observações e um gráfico que apresenta a distribuição de frequências das 1.389 observações.² Note que o gráfico é mais informativo do que os dados brutos!

Na *análise exploratória de dados*, além de descrever os dados, buscamos conhecer algumas características do processo, com base nos dados. Com o uso adequado de tabelas, gráficos e medidas, podemos descobrir certas estruturas que não eram evidentes nos dados brutos. Hoje também se usa a expressão *mineração de dados (data mining)*, que significa a busca por relacionamentos não triviais, que podem estar escondidos em grandes massas de dados. Em geral, numa mineração de dados, são aplicadas técnicas estatísticas e computacionais.

1 Extraído da dissertação de mestrado de Luciana S. C. V. da Silva (Programa de Pós-Graduação em Engenharia de Produção, UFSC, 1999).

2 A construção de distribuições de frequências é objeto de estudo deste capítulo.

Temperatura (°C)				
74,8	74,0	74,7	74,4	75,9
73,8	74,4	74,8	76,8	73,6
75,3	73,4	74,7	73,4	74,2
76,4	73,2	76,5	75,6	73,5
76,3	74,1	75,0	76,0	74,7
76,8	74,3	74,9	77,0	75,1
72,9	72,9	74,6	75,0	75,1
74,9	74,5	77,1	74,6	74,8
...

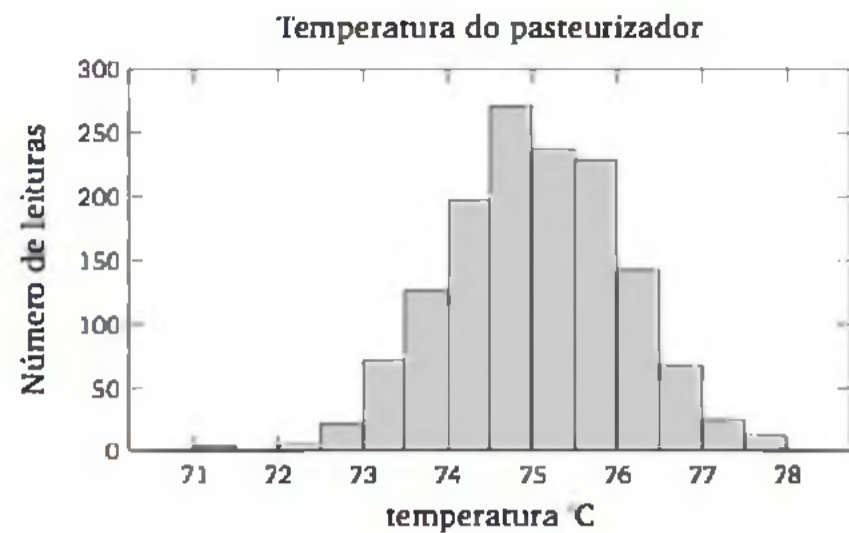
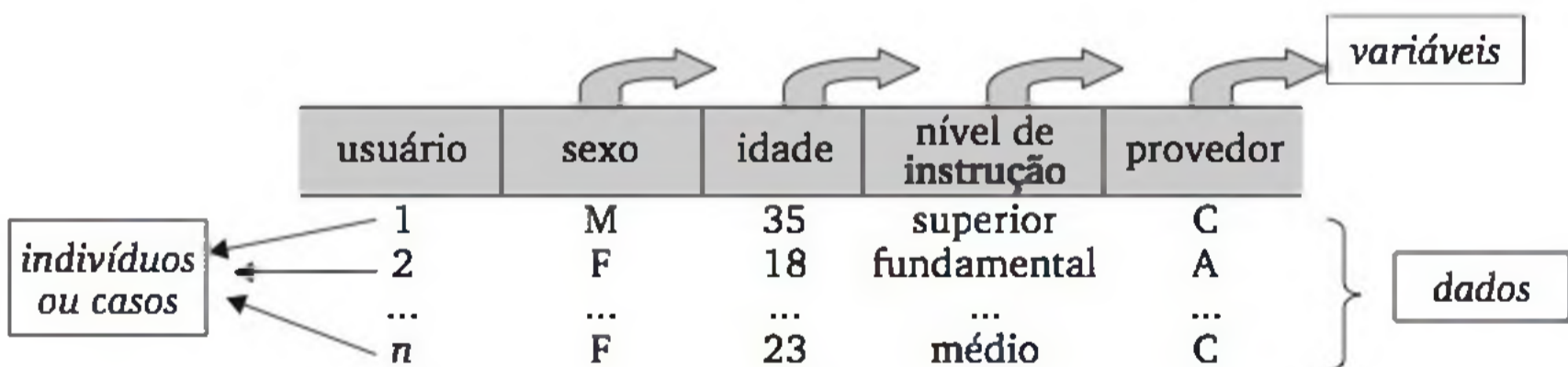


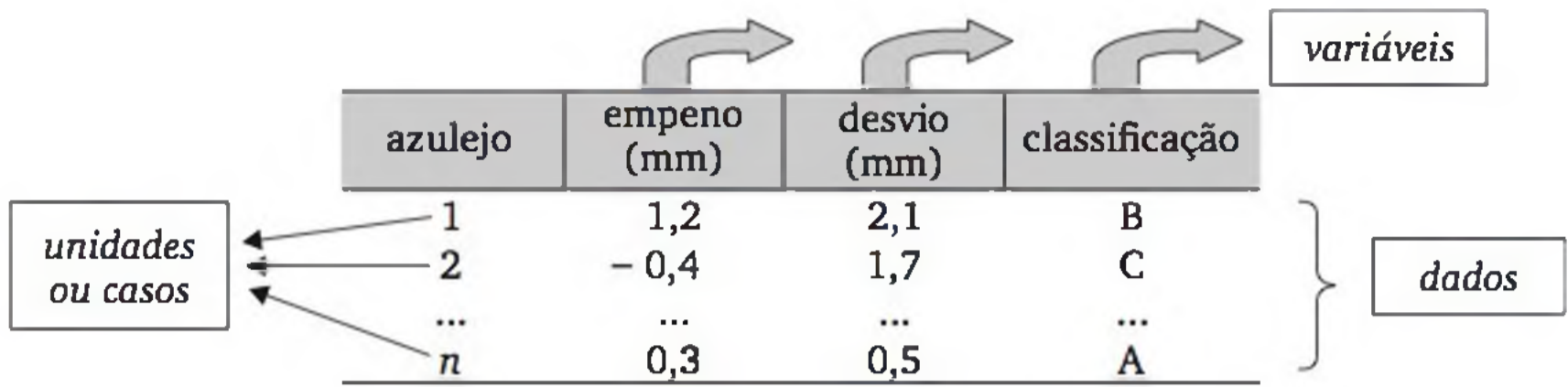
Figura 3.1 Parte dos dados brutos e um gráfico de distribuição de frequências.

3.1 DADOS E VARIÁVEIS

Depois de realizado um levantamento de dados, eles são colocados em arquivos, sob a forma de matrizes. As linhas dessas matrizes correspondem ao que se observou em cada elemento pesquisado, enquanto as colunas correspondem às características (*variáveis*) levantadas. Por exemplo, na atualização das páginas de um *site*, podemos querer avaliar o perfil dos indivíduos que acessam esse *site*. Então, precisamos levantar, junto a cada indivíduo, algumas de suas características, tais como o sexo, a idade, o nível de instrução e o provedor utilizado. Ao realizar a pesquisa, podemos produzir uma matriz de dados, da seguinte forma:



No controle da qualidade de uma linha de produção de azulejos, podemos observar algumas características em azulejos extraídos periodicamente da fase final da linha de produção. Essas características podem ser: *empeno*, *desvio máximo da dimensão ideal*, *variações na tonalidade da cor* etc. Considerando que a última variável produza uma classificação do tipo A, B e C, a matriz de dados pode ser apresentada como:



Os dados podem ser observações de variáveis qualitativas ou de variáveis quantitativas. E as técnicas de análise serão diferentes para cada caso. Quando os possíveis resultados de uma variável são números de certa escala, dizemos que essa variável é quantitativa. Quando os possíveis resultados são atributos ou qualidades, a variável é dita qualitativa (veja a Figura 3.2).

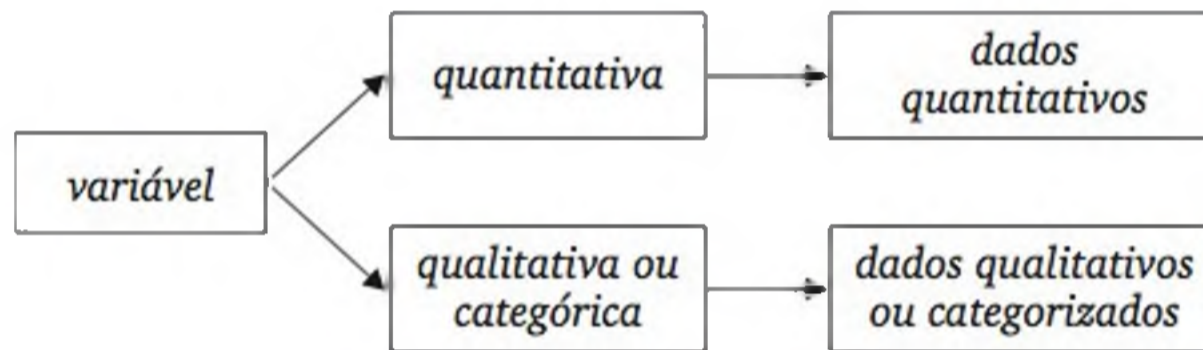


Figura 3.2 Classificação das variáveis e dos dados, em termos do nível de mensuração.

No exemplo dos usuários do site, o sexo, o grau de instrução e o provedor são variáveis qualitativas, enquanto a idade é quantitativa. Nas características dos azulejos, a classificação é uma variável qualitativa, mas o empeno e o desvio são variáveis quantitativas.

Na descrição das variáveis envolvidas na pesquisa, devemos incluir a escala (ou unidade) em que são mensuradas as variáveis quantitativas, e as categorias (possíveis respostas) das variáveis qualitativas. Sempre que uma característica puder ser adequadamente medida sob forma quantitativa, é melhor usarmos esse tipo de mensuração, porque as medidas quantitativas são, em geral, mais informativas do que as qualitativas. Por exemplo, dizer que o empeno é 8,2 mm é mais informativo do que dizer que o empeno é grande.

Distribuição de frequências

Um dos primeiros passos para analisar um arquivo de dados, especialmente quando o número de observações for grande, é a *distribuição de frequências* de cada variável.

A **distribuição de frequências** consiste na organização dos dados de acordo com as ocorrências dos diferentes resultados observados.

Por exemplo, ao observar a variável *sexo*, num conjunto de indivíduos, estaremos classificando cada indivíduo ou na categoria *masculino*, ou na categoria *feminino*. A contagem de quantos elementos existem em cada categoria forma uma *distribuição de frequências* dos dados dessa variável, que pode ser apresentada em uma tabela ou em um gráfico. As frequências podem ser apresentadas de forma absoluta (número de homens e número de mulheres) ou de forma relativa (a porcentagem de homens e a porcentagem de mulheres) ou, ainda, de ambas as formas. Normalmente, frequências absolutas são preferidas quando o número de observações é pequeno. Por outro lado, quando queremos fazer comparações, como estudar as distribuições de homens e mulheres em duas turmas, as frequências relativas são mais informativas.

As duas seções seguintes mostram a construção e a apresentação de distribuições de frequências para variáveis qualitativas e quantitativas, respectivamente.

3.2 ANÁLISE DE VARIÁVEIS QUALITATIVAS

Nesta seção, aprenderemos a descrever e a explorar dados de *variáveis qualitativas*, ou seja, aquelas cujos possíveis resultados são observados na forma de categorias. Para isso, vamos acompanhar o exemplo a seguir:

Exemplo 3.1 Para adequar os produtos às preferências dos clientes, um projetista de páginas de Internet pretende conhecer o perfil dos indivíduos que acessam um de seus *sites*. Pensando nisso, ele fez uma pesquisa e levantou algumas características dos visitantes de seu *site*, tais como o sexo, a idade, o nível de instrução e o provedor utilizado. Geralmente esse tipo de pesquisa envolve centenas ou milhares de respondentes, mas aqui consideraremos apenas uma amostra de 40 pessoas. Os resultados da pesquisa estão apresentados a seguir:

Indiví- duo	Provedor	Indiví- duo	Provedor	Indiví- duo	Provedor	Indiví- duo	Provedor
1	C	11	C	21	B	31	A
2	A	12	A	22	A	32	A
3	B	13	B	23	A	33	B
4	B	14	D	24	B	34	C
5	C	15	A	25	A	35	B
6	B	16	B	26	A	36	D
7	D	17	B	27	B	37	B
8	B	18	C	28	D	38	B
9	B	19	D	29	D	39	B
10	A	20	B	30	C	40	C

Tabelas de frequência

Para construir a distribuição de frequências com os dados de uma variável qualitativa, basta *contar* a quantidade de resultados observados em cada categoria. A Tabela 3.1 mostra a distribuição de frequências da variável *provedor*.

Tabela 3.1 *Distribuição de frequências do provedor usado pelo visitante do site.*

Provedor	Frequência	Porcentagem
A	10	25,0
B	17	42,5
C	7	17,5
D	6	15,0
Total	40	100,0

A primeira coluna da Tabela 3.1 mostra todas as categorias previamente estabelecidas da variável *provedor*. A segunda coluna resulta da *contagem* de quantas observações se identificam com cada categoria: são as *frequências observadas*. Finalmente, a terceira coluna apresenta uma medida relativa da frequência de cada categoria. Essas *porcentagens* são obtidas dividindo-se a frequência de cada categoria pelo número total de observações e, em seguida, multiplicando-se por 100.

As frequências relativas são particularmente importantes para comparar distribuições de frequências. Por exemplo, poderíamos ter interesse em comparar o perfil de visitantes de dois *sites*. Como, em geral, as amostras têm

tamanhos diferentes, a comparação através das frequências absolutas fica prejudicada, enquanto as frequências relativas sempre apresentam o mesmo total.

Representações gráficas

As representações gráficas fornecem, em geral, visualização mais sugestiva do que as tabelas. São formas alternativas de apresentar uma distribuição de frequências.

A Figura 3.3 representa a distribuição de frequências da Tabela 3.1 através de um *gráfico de colunas*, onde cada categoria é representada por uma coluna, e a frequência (absoluta ou relativa) é colocada no eixo vertical.³

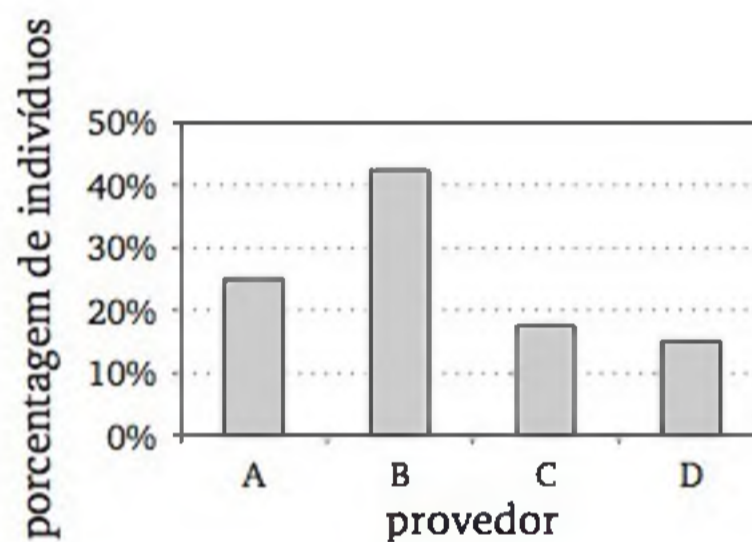


Figura 3.3 *Gráfico de colunas para a apresentação da distribuição de frequências do provedor usado pelo visitante do site.*

Alternativamente, o eixo horizontal poderia representar a escala das frequências e o eixo vertical, as categorias. Estaríamos construindo o chamado *gráfico de barras*.

Um gráfico muito comum para representar distribuições de frequências de variáveis qualitativas é o chamado *gráfico de setores*, que é particularmente útil quando o número de categorias não é grande e as categorias não obedecem a alguma ordem específica. A Figura 3.4 apresenta a distribuição através de um gráfico de setores.⁴

3 Da mesma forma que as tabelas, os gráficos devem conter um título, contendo todas as informações pertinentes. Eles costumam ser referenciados num texto como figuras. A posição do título de uma figura deve ser embaixo da figura.

4 O gráfico de setores é construído através de uma relação linear (*regra de três*) entre as frequências e comprimentos (em graus) de setores de um círculo.

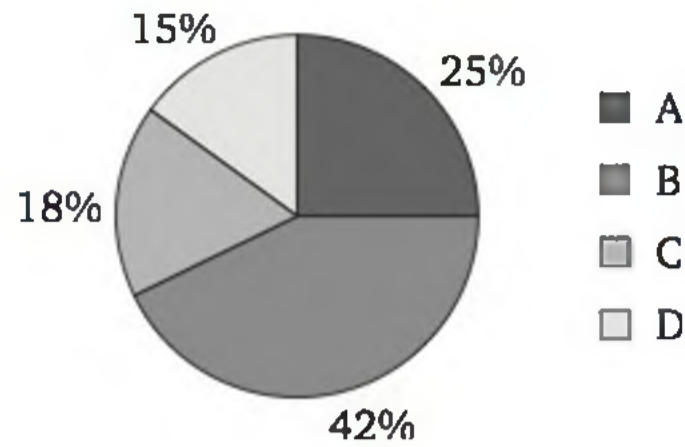
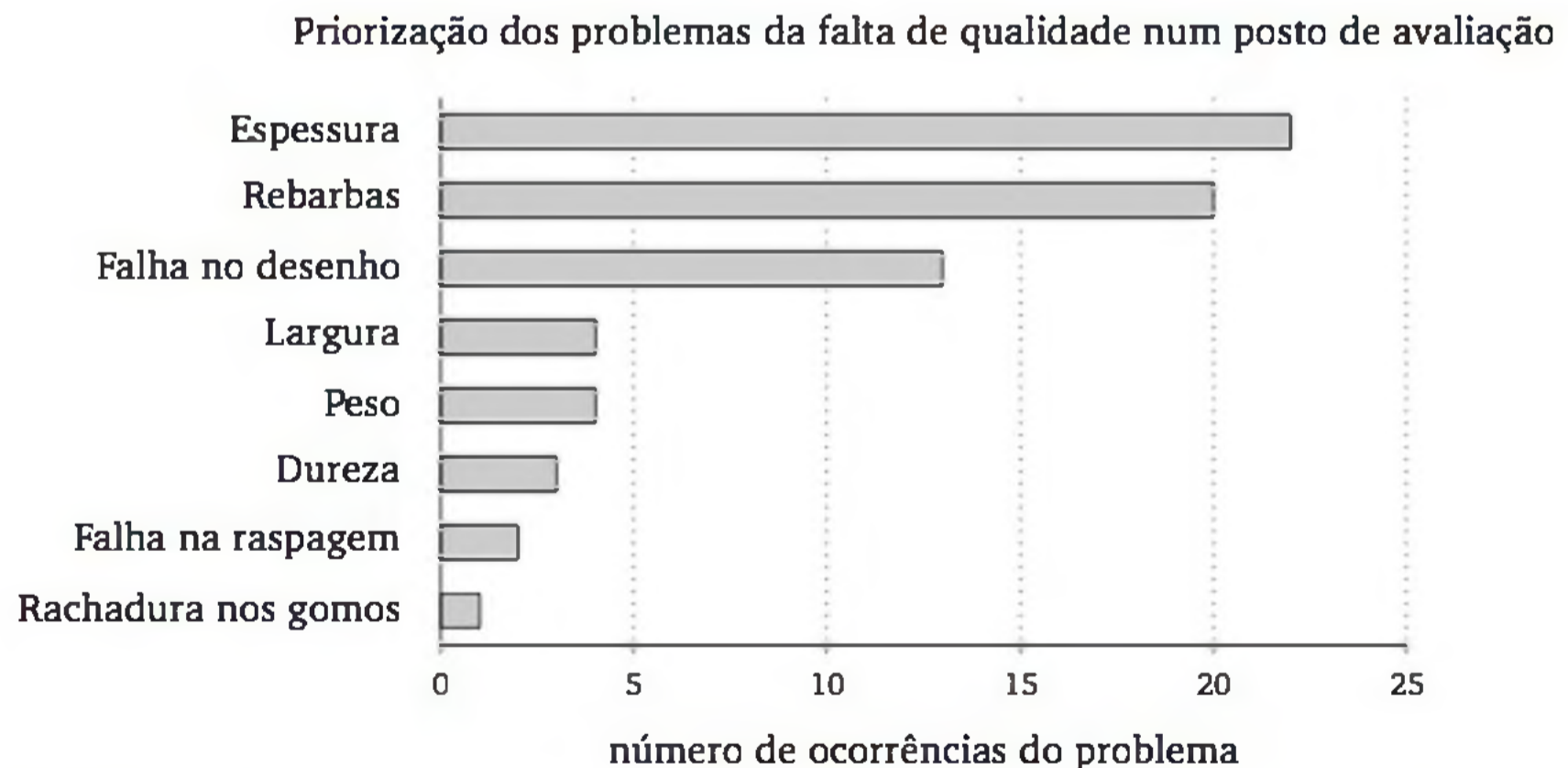


Figura 3.4 Gráfico de setores para a apresentação da distribuição de frequências do provedor usado pelo visitante do site.

Diagrama de Pareto

Uma das ferramentas dos programas de qualidade é o chamado *diagrama de Pareto*. Sua construção em si é bastante simples, pois corresponde ao gráfico de colunas ou ao gráfico de barras, mas com as categorias ordenadas decrescentemente pelas frequências observadas.

A importância desse diagrama é que ele é usado nos processos produtivos, em postos de avaliação da qualidade, colocando hierarquicamente os problemas encontrados pela falta da qualidade. A Figura 3.5 exemplifica um diagrama de Pareto, que mostra os problemas na implantação do controle integrado de processos. Nesse exemplo, a priorização dos problemas foi feita pela frequência de ocorrências, resultando em uma distribuição de frequências.



Fonte: CATEN; RIBEIRO; FOGLIATTO. Revista *Produto & Produção*, v. 4, nº 1, 2000.

Figura 3.5 Ilustração de um diagrama de Pareto – priorização em termos das frequências observadas.

Observando a Figura 3.5, verificamos que as principais características que levam à falta de qualidade no posto de avaliação são a espessura, as rebarbas e as falhas no desenho. Um programa de melhoria da qualidade deve enfrentar prioritariamente essas características.

No diagrama de Pareto, é mais comum priorizar os problemas da qualidade em termos financeiros. A Figura 3.6 ilustra essa situação, colocando também uma curva que indica as perdas acumuladas pela falta de qualidade. O eixo vertical da direita informa o percentual acumulado da perda.

Problemas	Perdas (R\$)	Perdas acumuladas
Manchas	15.000	15.000
Rachaduras	10.000	25.000
Furos	3.000	28.000
Riscos	2.000	30.000

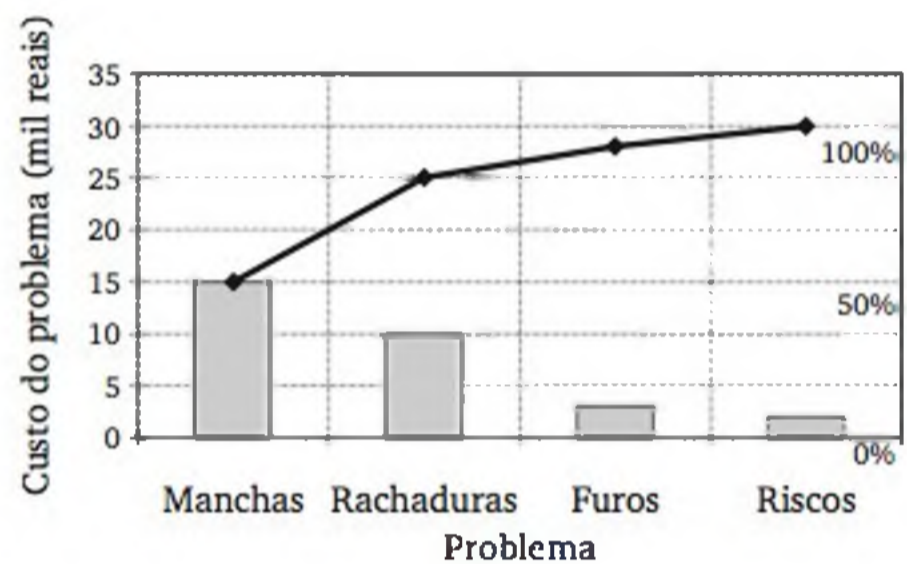


Figura 3.6 Ilustração de um diagrama de Pareto – priorização em termos de perdas financeiras.

EXERCÍCIO

- Considere o objetivo de verificar a demanda da qualidade no desenvolvimento de um *software*. Numa pesquisa de mercado, indagou-se aos clientes potenciais qual dos seguintes itens era considerado mais importante: (a) interface de fácil acesso, (b) desempenho do sistema, (c) métodos de análise avançados, (d) método de custeio, (e) manutenção e suporte, (f) personalização, (g) atualização em tempo real, (h) confiabilidade das informações, (i) segurança dos dados e (j) uso de novas tecnologias de informática.⁵ As frequências de resposta foram (a) 8, (b) 7, (c) 7, (d) 12, (e) 2, (f) 4, (g) 3, (h) 21, (i) 6 e (j) 0, respectivamente.
 - A variável *demanda da qualidade no desenvolvimento de um software*, operacionalizada de acordo com a pergunta feita aos clientes, é qualitativa ou quantitativa?

⁵ Os itens foram extraídos de um artigo de SONDA; RIBEIRO; ECHEVESTE. Revista *Produção*, v. 10, nº 1, 2000.

- b)** Construa um gráfico (ou diagrama) que mostre a distribuição de frequências das respostas, priorizando os itens segundo a frequência de respostas. Qual o nome que se dá a esse diagrama?

3.3 ANÁLISE DE VARIÁVEIS QUANTITATIVAS

Conforme foi discutido anteriormente, uma variável é dita *quantitativa* quando os possíveis resultados são números em certa escala. Por exemplo, numa fábrica de cerâmicas, podemos contar o número de defeitos e medir o grau de empeno (em mm), em cada azulejo. O *número de defeitos* e o *grau de empeno* são variáveis quantitativas. As observações dessas variáveis geram amostras de dados quantitativos.

O *número de defeitos*, cujos possíveis resultados podem ser listados (0, 1, 2, ...), é um exemplo do que chamamos de *variável discreta*. Enquanto o *grau de empeno*, que teoricamente pode assumir qualquer valor num intervalo de números reais (no caso, $[0, \infty)$), é um exemplo de *variável contínua*.

3.3.1 Variáveis discretas

A distribuição de frequências de variáveis discretas pode ser feita de forma análoga à distribuição de frequências de variáveis qualitativas. Porém, como os valores da variável formam uma escala numérica, temos, graficamente, um par de eixos cartesianos. Por convenção, o eixo horizontal representa a variável e o eixo vertical, as frequências.

A análise difere um pouco daquela feita com variáveis qualitativas. Normalmente, três informações principais são procuradas quando estamos explorando uma variável quantitativa:

- a) faixa em que os valores ocorrem com maior frequência (faixa de valores típicos);
- b) valores discrepantes, que podem ter sido originados de erros de mensuração ou digitação, mas também podem corresponder a elementos que apresentam comportamento muito diferente dos demais;
- c) forma da distribuição, a fim de compará-la com modelos probabilísticos, o que nos permite usar técnicas mais avançadas de análise.

A Figura 3.7 ilustra um exemplo de gráfico de distribuição de frequências de uma variável discreta.

Contagem de defeitos em cada unidade de um produto
Amostra de 50 unidades

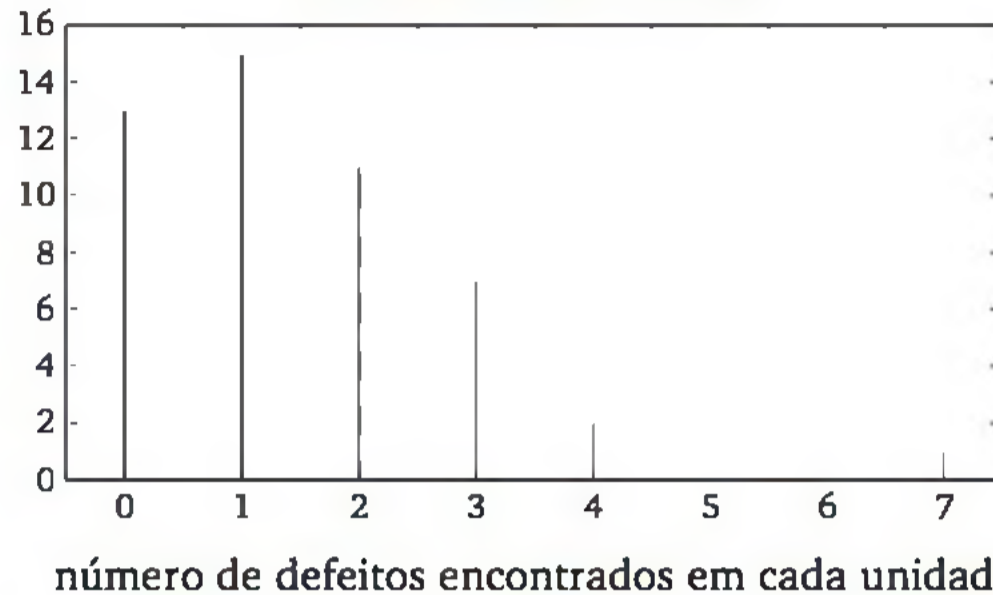


Figura 3.7 *Distribuição de frequências de um conjunto de observações de uma variável discreta.*

Observando o gráfico, notamos que os valores distribuem-se basicamente na faixa de 0 a 3 defeitos. Também é possível observar que há um valor discrepante: um elemento apresentou 7 defeitos. Além disso, vemos maior concentração de observações na parte inferior da escala.

3.3.2 Variáveis contínuas

Nesta seção, apresentaremos algumas formas de construção de distribuições de frequências de dados contínuos (provenientes de variáveis aleatórias contínuas). Esses procedimentos também poderão ser usados para variáveis discretas, quando o número de distintos valores for grande.

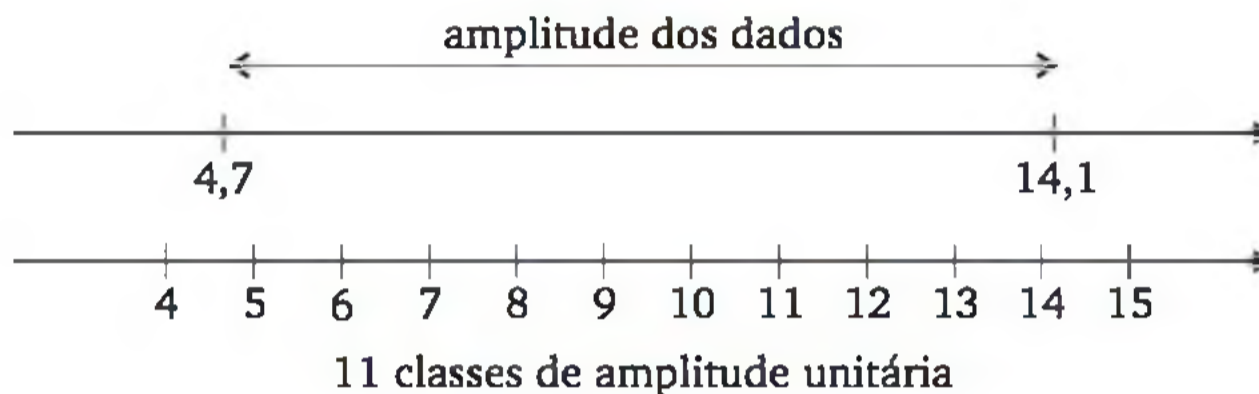
Tabela de frequências

Para construir uma tabela de frequências, dividimos a amplitude total dos dados (diferença entre o maior e o menor valor) em vários intervalos, denominados *classes*. Esses intervalos devem ser mutuamente exclusivos, exaustivos e, de preferência, ter o mesmo tamanho.⁶

⁶ Em alguns casos, não é possível usar intervalos com a mesma amplitude. Por exemplo, se formos analisar os salários em uma empresa, provavelmente encontraremos os valores distribuídos em uma grande amplitude e a maioria deles concentrados na parte inferior da escala. Assim, pode ser conveniente usarmos intervalos menores para os valores iniciais e intervalos maiores (mais amplos) para valores finais.

Exemplo 3.2 Os dados, a seguir, representam o tempo (em segundos) para carga de um aplicativo, num sistema compartilhado (50 observações):

5,2 6,4 5,7 8,3 7,0 5,4 4,8 9,1 5,5 6,2 4,9 5,7 6,3
 5,1 8,4 6,2 8,9 7,3 5,4 4,8 5,6 6,8 5,0 6,7 8,2 7,1
 4,9 5,0 8,2 9,9 5,4 5,6 5,7 6,2 4,9 5,1 6,0 4,7 14,1
 5,3 4,9 5,0 5,7 6,3 6,0 6,8 7,3 6,9 6,5 5,9



O número de classes a ser usado na tabela de frequências é uma escolha arbitrária. Quanto maior o conjunto de dados, mais classes podem ser usadas. Uma tabela com poucas classes apresenta a distribuição de forma bastante resumida, podendo deixar de evidenciar algumas características relevantes. Por outro lado, se usarmos muitas classes, a tabela pode ficar muito grande, não realçando aspectos relevantes da distribuição de frequências.

Em geral, são empregadas de 5 a 20 classes, dependendo da quantidade de dados e dos objetivos. Dentro dessa faixa, uma sugestão é usar, aproximadamente, $c = \sqrt{n}$ classes, onde n é a quantidade de valores observados. Em nosso exemplo, por facilidade e para melhor apresentação, optamos por usar classes com amplitude unitária, totalizando 11 classes.

O passo seguinte é contar quantos valores encontram-se em cada classe previamente estabelecida. Como os dados são arredondados para um número finito de decimais, podem ocorrer valores exatamente no limite entre duas classes. Por convenção, consideraremos sempre o intervalo fechado no limite inferior e aberto no limite superior. A primeira classe, por exemplo, é formada pelo intervalo $[4,0, 5,0)$, ou $4 \mid - 5$, conforme a simbologia comumente usada em Estatística.

A tabela de frequências é construída através da contagem da frequência de observações em cada classe, como mostramos a seguir:

Classes	Contagem	Frequência
4 — 5		7
5 — 6		18
6 — 7		13
7 — 8		4
8 — 9		5
9 — 10		2
10 — 11		0
11 — 12		0
12 — 13		0
13 — 14		0
14 — 15		1

Na apresentação de uma tabela de frequências, é comum colocar também os *pontos médios* das classes, isto é, para cada classe, a média de seus limites. Por exemplo, na classe 4 |— 5, temos o ponto médio 4,5. O ponto médio representa o *valor típico* da classe. A Tabela 3.2 apresenta a distribuição de frequência dos dados em discussão.

Tabela 3.2 *Distribuição de frequências do tempo (em segundos) para carga de um aplicativo, num sistema compartilhado.*

Classes de tempo	Ponto médio	Número de observações n_j	Porcentagem de observações $100f_j$	Porcentagem acumulada $100F_j$
4 — 5	4,5	7	14	14
5 — 6	5,5	18	36	50
6 — 7	6,5	13	26	76
7 — 8	7,5	4	8	84
8 — 9	8,5	5	10	94
9 — 10	9,5	2	4	98
10 — 11	10,5	0	0	98
11 — 12	11,5	0	0	98
12 — 13	12,5	0	0	98
13 — 14	13,5	0	0	98
14 — 15	14,5	1	2	100
Total	—	50	100	—

As duas últimas colunas da Tabela 3.2 apresentam as frequências relativas e as frequências relativas acumuladas, que são obtidas, respectivamente, por:

$$f_i = \frac{n_j}{n} \quad (j = (1, 2, \dots, c)) \quad (3.1)$$

$$F_j = \sum_{i=1}^j f_i \quad (j = 1, 2, \dots, c) \quad (3.2)$$

onde n é o número de observações e c , o número de classes.

Muitas vezes, o maior interesse está nas frequências de observações menores ou iguais a determinados valores. A Figura 3.8 mostra o gráfico das frequências relativas acumuladas. A seta ilustra o percentual aproximado de observações abaixo de 7,0 segundos.⁷

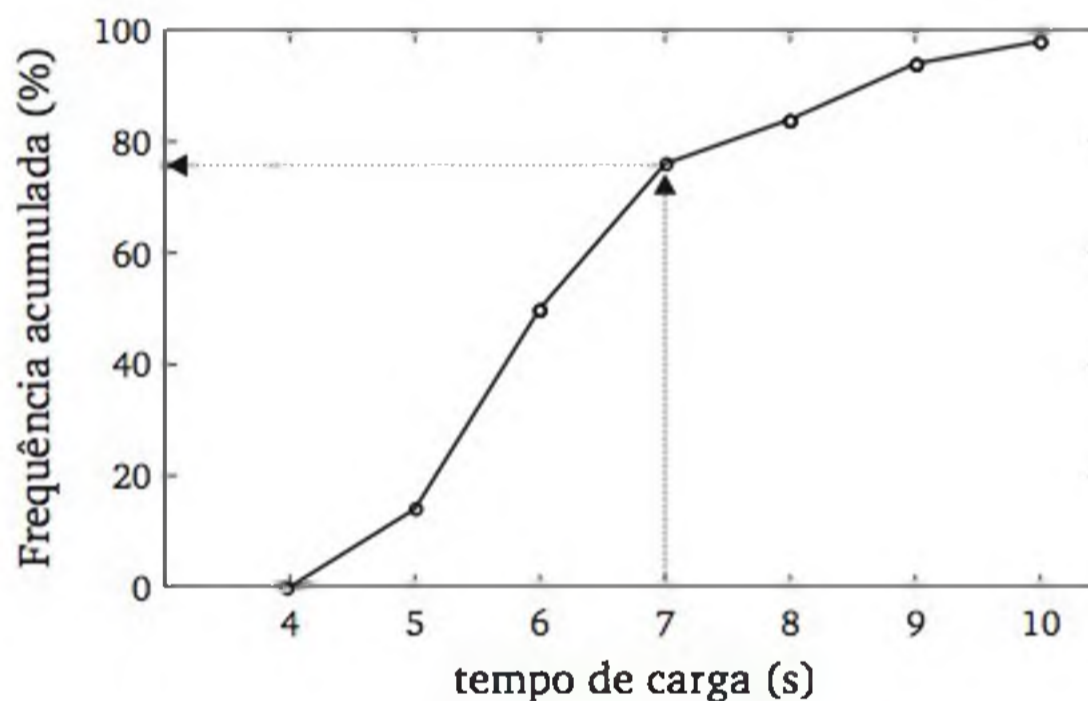


Figura 3.8 Distribuição de frequências acumuladas, em %, de 50 observações do tempo de carga de certo aplicativo.

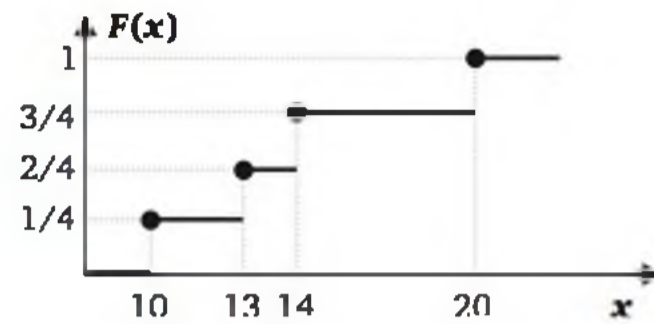
As frequências acumuladas podem ser definidas de maneira mais rigorosa com os dados não agrupados em classes, mas somente ordenados. Chamando de $n(x)$ o número de observações menores ou iguais a x , a frequência de observações até o valor x é dada por:

$$F(x) = \frac{n(x)}{n} \quad (3.3)$$

⁷ Nesta figura, foi retirado o valor discrepante (14,1). Sempre temos de buscar as causas de tais valores discrepantes. Se chegarmos à conclusão de que o valor se deve a algum erro ou alguma situação especial que não deverá ocorrer novamente, esse valor pode ser retirado e as análises posteriores podem desconsiderá-lo. Na maior parte dos casos, um valor discrepante deve ser analisado separadamente dos demais.

Por exemplo, para o conjunto de valores {10, 13, 14, 20}, temos:

$$F(x) = \begin{cases} 0 & \text{para } x < 10 \\ 1/4 & \text{para } 10 \leq x < 13 \\ 2/4 & \text{para } 13 \leq x < 14 \\ 3/4 & \text{para } 14 \leq x < 20 \\ 1 & \text{para } x \geq 20 \end{cases}$$



Histograma

O *histograma* é a forma mais usual de apresentação de distribuições de frequências de variáveis contínuas. A Figura 3.9 mostra um histograma, construído com base na Tabela 3.2. São retângulos justapostos, feitos sobre as classes da variável em estudo. A área de cada retângulo é igual (ou proporcional) à frequência observada da correspondente classe.⁸

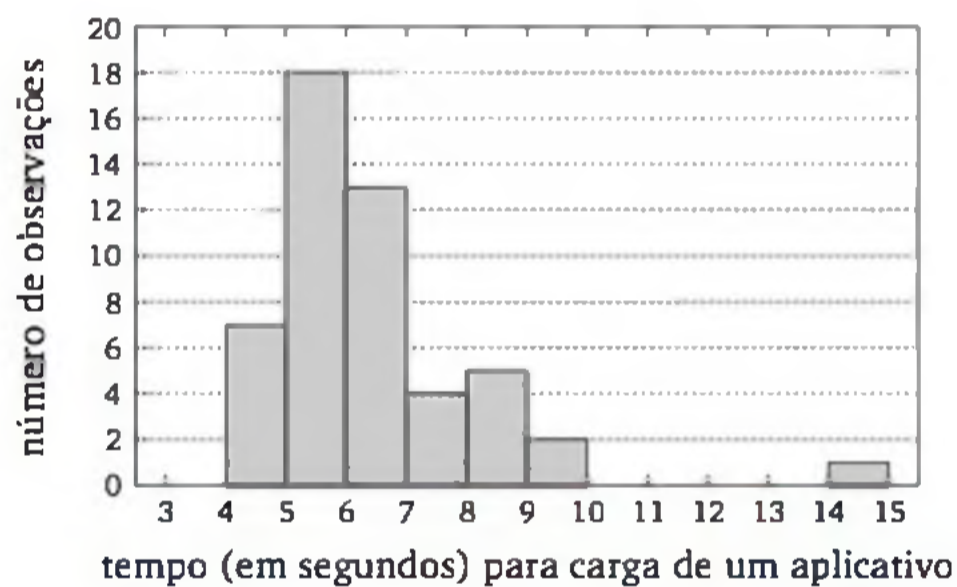


Figura 3.9 Distribuição de frequências de 50 observações do tempo de carga de um certo aplicativo. Apresentação em histograma de frequências.

Note que o histograma (Figura 3.9) permite a mesma análise dos dados que a tabela de frequências (Tabela 3.2), porém de forma mais clara.

Diagrama de pontos

Uma forma simples de observar como poucas observações se distribuem é através do *diagrama de pontos*, onde representamos cada valor como um ponto

⁸ O histograma também poderia ser feito usando percentagens no eixo vertical, mas a sua forma não mudaria. A rigor, o eixo vertical de um histograma deve representar as *densidades de frequências*, isto é, as frequências relativas, f_j , divididas pelos correspondentes intervalos das classes, fazendo com que a soma das áreas dos retângulos se iguale a um. O uso de densidades de frequências é fundamental quando as classes não têm a mesma amplitude.

na reta de números reais. Por exemplo, considere um estudo experimental sobre um processo químico, do qual queremos avaliar o rendimento em dois níveis da temperatura de reação: 60 e 80°C. A Figura 3.10 apresenta os dados e o diagrama de pontos relativos aos oito ensaios realizados com cada temperatura.

Temperatura	Rendimento (%)
60°C	31,0 33,6 32,8 32,2
	31,9 36,2 34,3 34,0
80°C	37,0 34,4 39,8 38,5
	33,9 43,2 35,5 39,0

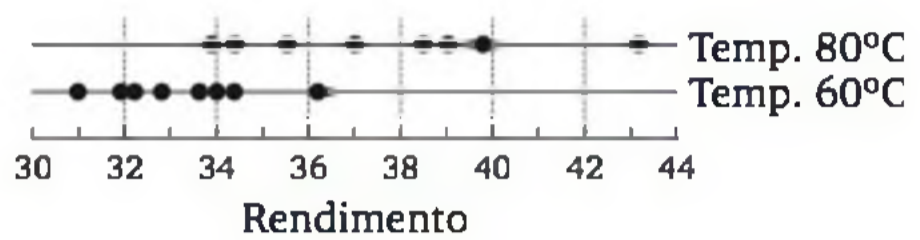


Figura 3.10 Ilustração de um diagrama de pontos.

Podemos observar na Figura 3.10 que a temperatura de 80°C produziu, em geral, rendimentos maiores. Mas os pontos apresentaram-se mais dispersos, o que nos leva a suspeitar que temperaturas mais altas podem provocar maior variabilidade no processo químico.

A estratégia de análise por meio do diagrama de pontos torna-se inadequada quando o número de observações é grande, ao contrário da tabela de frequências e do histograma, que podem ser usados para grandes conjuntos de observações quantitativas.

Diagrama ramo-e-folhas

O diagrama *ramo-e-folhas* consiste em apresentar os dados separando os primeiros dígitos, os quais formarão os ramos, e os demais dígitos, que formarão as folhas. Por exemplo, para os números 10, 15 e 23, as dezenas 1 e 2 ficariam do lado esquerdo de uma linha vertical (os ramos 1 e 2) e as unidades 0, 5 e 3, do lado direito (as folhas), como segue:

$$10, 15, 23 \longrightarrow \begin{array}{c|c} 1 & 0 \ 5 \\ 2 & 3 \end{array}$$

Para construir um *ramo-e-folhas* com os dados do Exemplo 3.2, podemos separar as unidades para formar os ramos e as decimais para formar as folhas, como ilustrado para os cinco primeiros valores.

$$5,2 \ 6,4 \ 5,7 \ 8,3 \ 7,0 \longrightarrow \begin{array}{c|c} 5 & 2 \ 7 \\ 6 & 4 \\ 7 & 0 \\ 8 & 3 \end{array} \quad \text{unidade} = 0,1$$

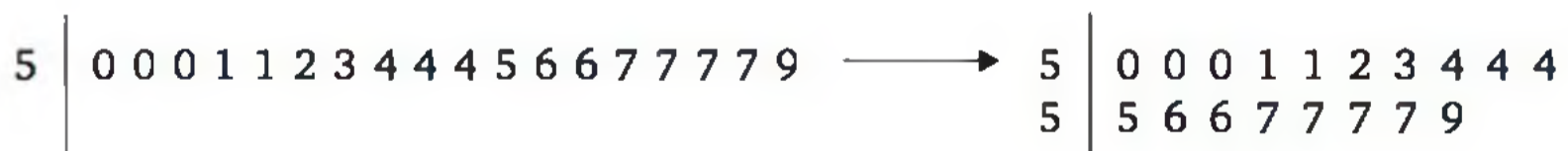
Repetindo o processo para as 50 observações, chegamos ao diagrama da Figura 3.11, onde as folhas (decimais dos números) foram ordenadas para facilitar a leitura.

		Frequência
4	7 8 8 9 9 9 9	7
5	0 0 0 1 1 2 3 4 4 4 5 6 6 7 7 7 7 9	18
6	0 0 2 2 2 3 3 4 5 7 8 8 9	13
7	0 1 3 3	4
8	2 2 3 4 9	5
9	1 9	2
10		0
11		0
12		0
13		0
14	1 unidade = 0,1	1

Figura 3.11 Diagrama ramo-e-folhas para os dados do Exemplo 3.2.

O diagrama *ramo-e-folhas* da Figura 3.11 permite observar que, na maioria das vezes, o tempo de carga fica entre 4 e 7 segundos. O diagrama evidencia um caso atípico (14,1 segundos). Note que em um *ramo-e-folhas* observamos as mesmas características que num histograma (reveja a Figura 3.9).

Muitas vezes, o número de folhas em cada ramo fica muito grande e, em consequência, o diagrama fica concentrado em poucos ramos. É possível, nesses casos, dividir cada ramo (composto de 10 folhas) em 2 ou 5 partes (novos ramos compostos de 5 ou 2 folhas, respectivamente). O esquema a seguir ilustra a divisão do ramo “5” em 2 novos ramos, sendo o primeiro formado pelas folhas de 0 a 4 e o segundo pelas folhas de 5 a 9. É importante que o número de folhas possíveis em cada ramo seja o mesmo em todo o diagrama.



3.3.3 Características de uma distribuição

Na análise exploratória de dados quantitativos, uma das informações usualmente procuradas é a posição central e a forma da distribuição de frequências. Por exemplo, a Figura 3.12 apresenta um histograma construído com 2.000 ob-

servações do tempo de carga de um aplicativo e uma curva contínua, que representa a forma aproximada do que se observou.

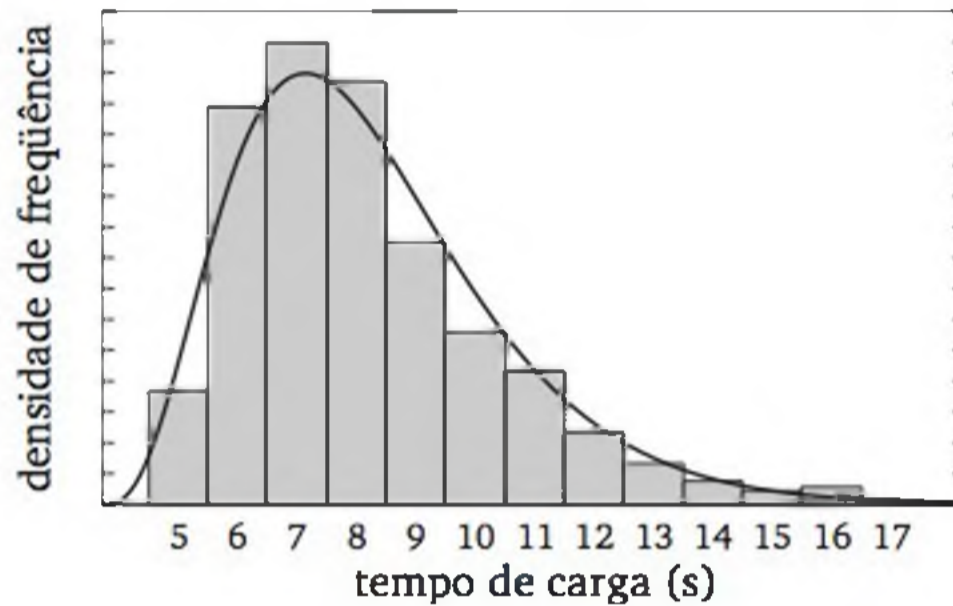
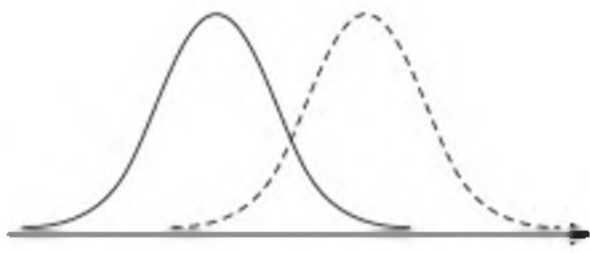
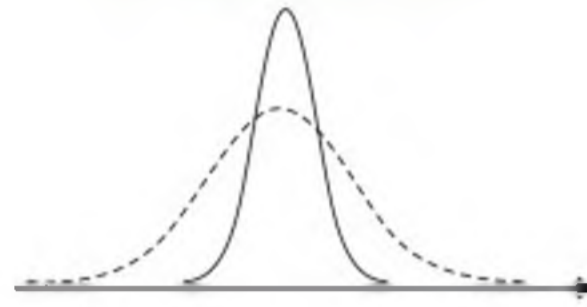
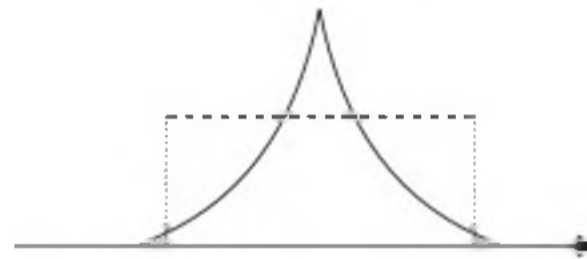


Figura 3.12 *Histograma de 2.000 observações do tempo de carga de um certo aplicativo e uma curva simbolizando a forma da distribuição.*

Ao confrontarmos a distribuição observada com vários modelos teóricos existentes, temos uma ideia de qual modelo seria o mais adequado para explicar o comportamento da variável estudada. Na investigação sobre a forma da distribuição, várias características devem ser observadas. As principais são:

- a) a *posição central*, que informa onde se localiza o centro da distribuição;
- b) a *dispersão*, que se refere à variabilidade dos dados;
- c) a *assimetria*, que representa a concentração dos valores em um dos extremos da distribuição;
- d) a *curtose*, que é o grau de achatamento da distribuição.

A Figura 3.13 ilustra diferentes formas de distribuição de frequências, considerando as características comentadas anteriormente.

(a) Distribuições diferentes em termos da posição central**(b)** Distribuições diferentes quanto à dispersão**(c)** Distribuições diferentes quanto à assimetria**(d)** Distribuições diferentes quanto à curtoseFigura 3.13 *Diferentes formas de distribuição de frequências.*

Na próxima seção, discutiremos algumas medidas que servem para quantificar a posição central e a dispersão.

EXERCÍCIO

2. Os dados abaixo representam 50 leituras de temperatura ($^{\circ}\text{C}$) de um pasteurizador de leite

74,8	74,0	74,7	74,4	75,9	76,8	74,3	74,9	77,0	75,1
73,8	74,4	74,8	76,8	73,6	72,9	72,9	74,6	75,0	75,1
75,3	73,4	74,7	73,4	74,2	74,9	74,5	77,1	74,6	74,8
76,4	73,2	76,5	75,6	73,5	76,2	74,7	76,0	75,8	77,3
76,3	74,1	75,0	76,0	74,7	75,2	77,5	74,7	73,3	74,3

- construa uma tabela de frequências;
- apresente a distribuição em um histograma;
- faça um gráfico da distribuição acumulada. Indique no gráfico a porcentagem aproximada de observações abaixo de 75°C ;
- construa um diagrama ramo-e-folhas.

3.4 MEDIDAS DESCRITIVAS

Quando analisamos uma variável qualitativa, basicamente construímos sua distribuição de frequências. No entanto, ao explorarmos variáveis quantitativas, temos condições de empregar algumas medidas descritivas, que sintetizam as características da distribuição.⁹ Aqui, vamos desenvolver medidas de posição central e dispersão.

Muitas vezes um conjunto de observações quantitativas pode ser bem representado por alguma medida descritiva. Por exemplo, se na primeira avaliação de uma disciplina a média das notas dos alunos foi igual a 8,0 e, na segunda avaliação, foi igual a 9,0, podemos concluir que, em geral, os alunos tiveram melhor aproveitamento na segunda avaliação, mesmo sem nos referirmos às notas de cada aluno individualmente.

Exemplo 3.3 O rendimento de um processo químico é influenciado pelo tempo e pela temperatura de reação. Um experimento é realizado para diferentes níveis do tempo de reação (20, 25 e 30 minutos) e da temperatura de reação (60, 70 e 80°C). Como os ensaios são também afetados por fatores não controláveis, as observações agregam um *erro experimental*. Devido à presença do erro experimental, foram realizados seis ensaios em cada combinação de níveis do tempo e da temperatura. Os resultados do experimento (rendimentos em %) são apresentados a seguir.

Temperatura (°C)	Tempo (minutos)								
	20			25			30		
60	29,7	28,7	30,2	31,0	30,6	32,8	32,9	32,7	34,8
	31,3	31,2	31,7	31,9	31,2	31,2	34,9	33,8	34,9
70	36,6	35,7	35,3	35,7	40,4	41,7	34,8	36,8	37,4
	35,1	30,2	37,2	36,9	34,5	40,0	38,9	38,7	42,5
80	40,2	33,6	33,4	37,0	34,4	29,8	36,0	31,3	36,6
	35,2	38,1	33,0	33,9	43,2	35,5	32,5	39,2	35,9

Observando os dados brutos, é difícil avaliar qual é a influência do tempo e da temperatura de reação sobre o rendimento. Porém, calculando a média aritmética em cada subgrupo da amostra de observações, as relações aparecem de forma mais nítida. A Figura 3.14 apresenta essas médias em forma tabular e gráfica.

⁹ Como vimos na seção anterior, as quatro principais características de uma distribuição são a tendência central, a dispersão, a assimetria e a curtose.

Temperatura (°C)	Tempo (minutos)		
	20	25	30
60	30,5	31,4	34,0
70	35,0	38,2	38,2
80	35,6	35,6	35,3

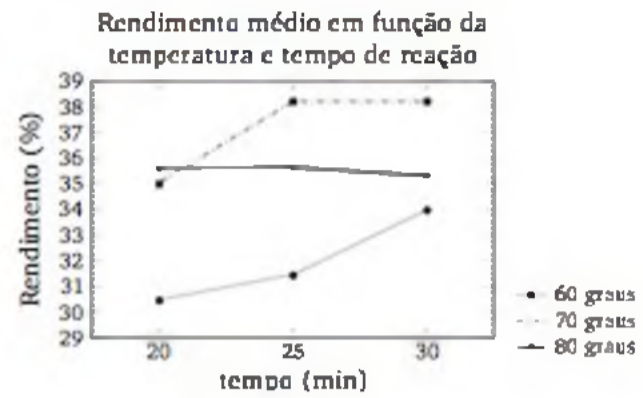


Figura 3.14 Médias aritméticas do rendimento, para diferentes níveis de temperatura e tempo de reação, num processo químico.

Pela Figura 3.14, é possível observar que, se usarmos a temperatura no nível intermediário (70°C) e tempo de reação na faixa de 25 a 30 minutos, então obteremos, em média, melhor rendimento.

3.4.1 Medidas clássicas

A média e o desvio padrão já foram estudadas no Capítulo 1. Aqui as apresentaremos com maiores detalhes.

A média aritmética

O conceito de *média aritmética*, ou simplesmente *média*, é bastante familiar. Seja (x_1, x_2, \dots, x_n) uma amostra de n observações de certa variável aleatória X . A média aritmética dessas observações é definida por

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i \quad (3.4)$$

No Exemplo 3.3, mostramos como a média aritmética resume os dados de forma a torná-los mais informativos. Já o Exemplo 3.4 ilustra alguns problemas ao tentarmos resumir vários valores por uma média.

Exemplo 3.4 Considere as notas finais, relativas aos alunos de três turmas, representadas em diagramas de pontos (Figura 3.15). As setas apontam para as posições das médias aritméticas.

Turma	Notas dos alunos								Média da turma
A	4	5	5	6	6	7	7	8	6,00
B	1	2	4	6	6	9	10	10	6,00
C	0	6	7	7	7	7,5	7,5		6,00

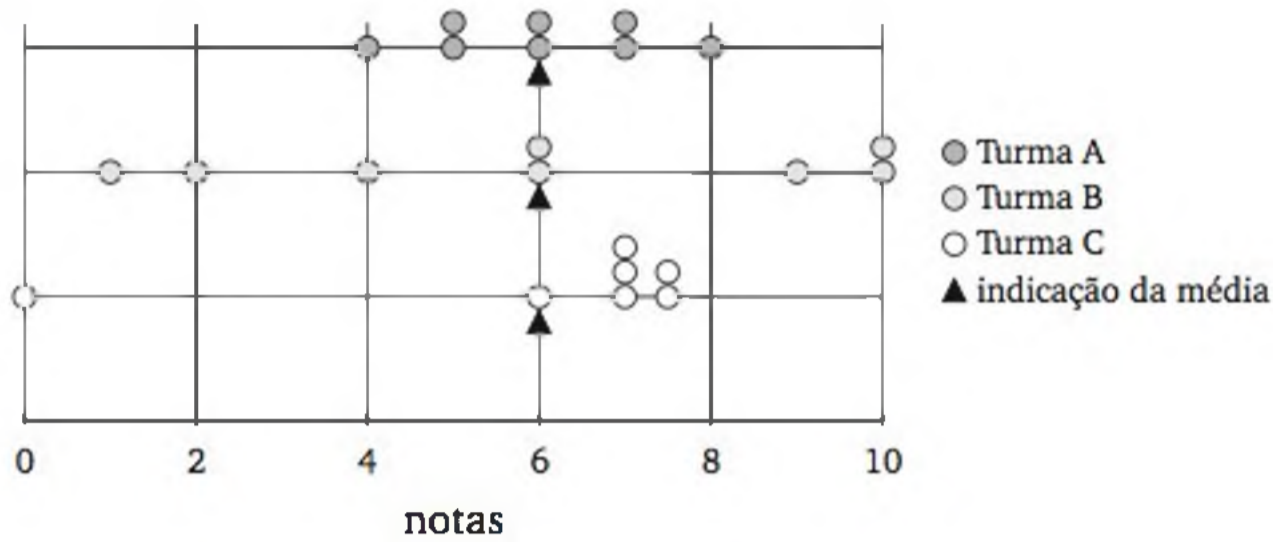


Figura 3.15 Representação das distribuições das notas de três turmas e correspondentes posições das médias aritméticas.

Observando a Figura 3.15, percebemos que, em cada diagrama de pontos, a média aritmética apresenta-se, de alguma forma, na posição central dos valores observados. Mais precisamente, a média aritmética indica o *centro* de um conjunto de valores, considerando o conceito físico de *centro de gravidade*. Se imaginarmos os pontos como pesos sobre uma tábua, a *média* é a posição em que um suporte equilibraria essa tábua.

Na Figura 3.15, também observamos que os três conjuntos de valores, apesar de estarem distribuídos sob diferentes formas, apontam para uma mesma média aritmética. Isso mostra que a média aritmética *resume* o conjunto de dados, em termos de posição central, ou de *valor típico*, mas não fornece qualquer informação sobre outros aspectos da distribuição. Comparando, por exemplo, as notas da turma A com as notas da turma B, verificamos que o segundo conjunto de notas é bem mais *disperso*, indicando que a turma B é mais heterogênea em termos de notas obtidas. No conjunto de notas da turma C, observamos um ponto discrepante dos demais: uma nota extremamente baixa, fazendo com que a média fique abaixo da maioria das notas da turma.¹⁰

Medidas de dispersão

Para melhorar o resumo dos dados, podemos apresentar, ao lado da média aritmética, uma medida da dispersão desses dados. Uma forma simples de medir a dispersão é através da *amplitude*, isto é, a diferença entre o maior e o menor valor. Matematicamente:

$$a = \text{máx.}(x_1, x_2, \dots, x_n) - \text{mín.}(x_1, x_2, \dots, x_n) \quad (3.5)$$

¹⁰ Podemos observar no diagrama de pontos referente à turma C que a presença de um valor discrepante *arrasta* a média para seu lado. Assim, a média deixa de representar propriamente um *valor típico* do conjunto de dados. Um tratamento mais adequado para dados que contenham valores discrepantes será visto na seção 3.4.2.

Como (3.5) é calculada usando apenas os dois valores mais extremos, ela pode levar a conclusões errôneas quando existirem valores discrepantes. Medidas mais apropriadas são a variância e o desvio padrão.

Tanto a variância quanto o desvio padrão são medidas que fornecem informações complementares à informação contida na média aritmética. Essas medidas avaliam a *dispersão* do conjunto de valores em análise. Para calcularmos a variância ou o desvio padrão, devemos considerar os desvios de cada valor em relação à média aritmética. Depois, construímos uma espécie de média desses desvios. Ilustramos, a seguir, as etapas de cálculo, usando o conjunto de notas da turma A (Exemplo 3.4).

Descrição	Notação	Resultados numéricos							
Valores (notas dos alunos)	x_i	4	5	5	6	6	7	7	8
Média	\bar{x}	6							
Desvios em relação à média	$x_i - \bar{x}$	-2	-1	-1	0	0	1	1	2
Desvios quadráticos	$(x_i - \bar{x})^2$	4	1	1	0	0	1	1	4

Para evitar o problema dos desvios negativos, vamos trabalhar com os desvios quadráticos, $(x_i - \bar{x})^2$ ($i = 1, 2, \dots, n$). A variância é definida como a média aritmética dos desvios quadráticos. Se estivermos trabalhando com uma amostra, devemos calcular essa média usando como denominador $(n - 1)$ no lugar de n . Assim, definimos a **variância** de um conjunto de valores pela expressão

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (3.6)$$

Em relação ao conjunto de notas da Turma A, a variância é

$$s^2 = \frac{4 + 1 + 1 + 0 + 0 + 1 + 1 + 4}{8 - 1} = 1,71$$

Como a variância de um conjunto de dados é calculada em função dos desvios quadráticos, sua unidade de medida equivale à unidade de medida dos dados ao quadrado. Nesse contexto, é mais comum trabalhar com a *raiz quadrada positiva* da variância. Essa medida é conhecida como *desvio padrão*, o qual é expresso na mesma unidade de medida dos dados em análise. Então, o **desvio padrão** de um conjunto de valores pode ser calculado por:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (3.7)$$

Em termos do conjunto de notas da turma A (Exemplo 3.4), temos o seguinte desvio padrão: $s = \sqrt{1,71} = 1,31$.

Ao compararmos os desvios padrões de vários conjuntos de dados, podemos avaliar quais se distribuem de forma mais (ou menos) dispersa. O desvio padrão será sempre *não negativo* e será tão maior quanto mais dispersos forem os valores observados. A Tabela 3.3 mostra o desvio padrão das notas de cada uma das três turmas de alunos, referente aos dados do Exemplo 3.4.

Tabela 3.3 *Medidas descritivas das notas finais dos alunos de três turmas.*

Turma	Número de alunos	Média	Desvio padrão
A	8	6,00	1,31
B	8	6,00	3,51
C	7	6,00	2,69

Analisando a Tabela 3.3, verificamos, através das médias, que os alunos das três turmas *tenderam* a ter as notas em torno de seis, mas, pelos desvios padrões, concluímos que os alunos da turma A obtiveram notas relativamente próximas umas das outras, quando comparados aos alunos das outras turmas.¹¹ Por outro lado, as notas dos alunos da turma B foram as que se apresentaram mais heterogêneas.

Retomando o Exemplo 3.3, onde buscávamos verificar o rendimento de um processo químico para diferentes níveis do tempo de reação (20, 25 e 30 minutos) e da temperatura de reação (60, 70 e 80°C), calculamos a média e o desvio padrão para cada subconjunto de seis observações realizadas nas diferentes combinações de níveis do tempo e temperatura de reação. Os resultados são colocados em gráficos, conforme mostramos na Figura 3.16.

Analisando as médias, observamos que, se usarmos a temperatura no nível intermediário (70°C) e o tempo de reação na faixa de 25 a 30 minutos, então teremos, em média, maior rendimento. Porém, analisando os desvios padrões, é possível verificar que, para temperaturas maiores, temos um aumento na variabilidade. Uma das preocupações da engenharia é reduzir a variabilidade de processos. Assim, devemos evitar temperaturas superiores a 70°C, pois, além de reduzir o rendimento médio, aumenta a variabilidade.

¹¹ Observe, pela Figura 3.15, que as notas da turma C estão mais concentradas do que as da turma A. Porém, o valor discrepante, além de deslocar a média, aumenta o desvio padrão. Se o valor discrepante fosse desconsiderado, o desvio padrão das notas da turma C seria o menor de todos – a média seria 7 e o desvio padrão 0,55.

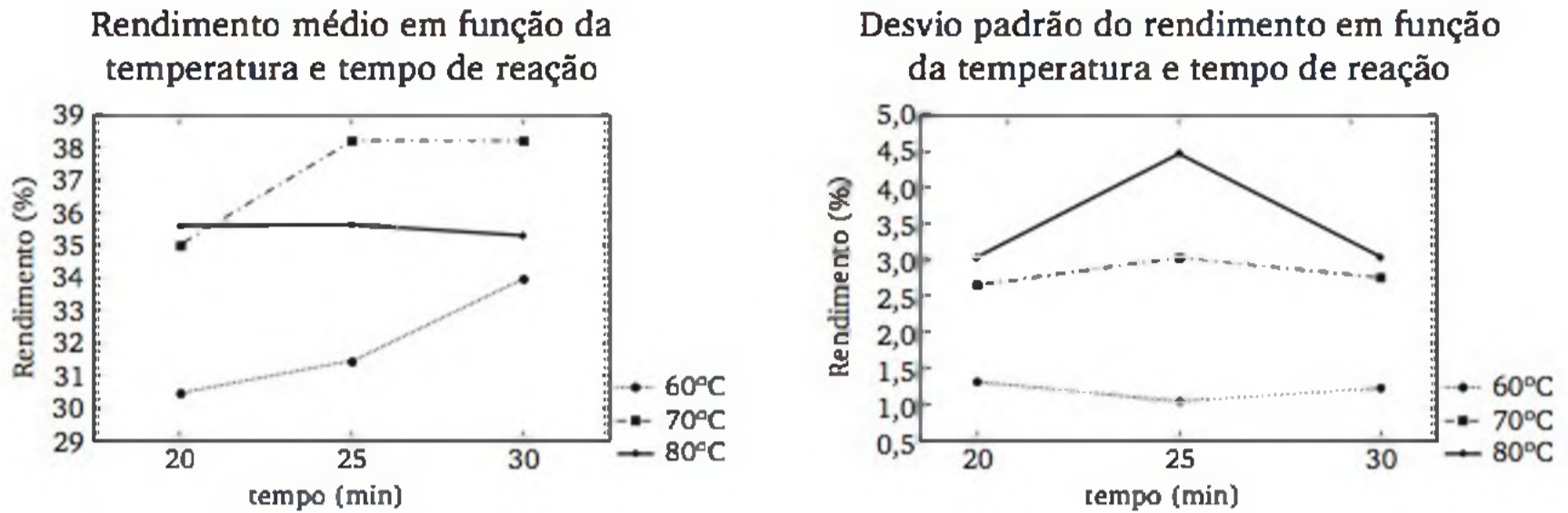


Figura 3.16 Médias aritméticas e desvios padrões do rendimento (%), para diferentes níveis de temperatura e tempo de reação (Exemplo 3.3).

O cálculo da média e do desvio padrão

Ao calcularmos o desvio padrão nos casos em que a média, \bar{x} acusar um valor fracionário, os desvios $(x_i - \bar{x})$ acumularão erros de arredondamento, que poderão comprometer o resultado final. Para evitar esse inconveniente, podemos usar a seguinte fórmula alternativa para o cálculo do desvio padrão, que é matematicamente equivalente à expressão (3.7):¹²

$$s = \sqrt{\frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right)} \tag{3.8}$$

Ilustraremos o uso desta nova formulação com as notas obtidas pelos alunos da Turma A (Exemplo 3.4):

Valores x_i : 4 5 5 6 6 7 7 8 $\therefore \sum_{i=1}^n x_i = 48$ e $\bar{x} = 6$

Valores ao quadrado x_i^2 : 16 25 25 36 36 49 49 64 $\therefore \sum_{i=1}^n x_i^2 = 300$

Então:

$$s = \sqrt{\frac{300 - 8 \cdot (6)^2}{7}} = \sqrt{\frac{300 - 288}{7}} = \sqrt{\frac{12}{7}} = 1,31$$

12 A igualdade no numerador das duas fórmulas de s^2 é demonstrada a seguir:

$$\begin{aligned} \sum_{i=1}^n (x_i - \bar{x})^2 &= \sum_{i=1}^n (x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + \sum_{i=1}^n \bar{x}^2 = \\ &= \sum_{i=1}^n x_i^2 - 2\bar{x}n\bar{x} + n\bar{x}^2 = \sum_{i=1}^n x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2 \end{aligned}$$

Como era de se esperar, chegamos ao mesmo resultado encontrado anteriormente pela Expressão (3.7).

Outro aspecto relativo ao cálculo da média e do desvio padrão refere-se à soma de valores repetidos. Por exemplo, ao calcularmos a média das notas da Turma A, fizemos a seguinte soma:

$$\sum_{i=1}^n x_i = 4 + 5 + 5 + 6 + 6 + 7 + 7 + 8$$

que é equivalente a $4(1) + 5(2) + 6(2) + 7(2) + 8(1) = \sum_{j=1}^k x_j n_j$

onde: x_j ($j = 1, 2, \dots, k$) representam os k valores distintos de X ;

n_j ($j = 1, 2, \dots, k$) são os números de ocorrências desses valores; e

$n = n_1 + n_2 + \dots + n_k$ é o número total de valores.

Analogamente, podemos calcular a soma quadrática dos valores de X por

$$\sum_{j=1}^k x_j^2 n_j = 4^2 + 5^2(2) + 6^2(2) + 7^2(2) + 8^2 = 300$$

Com essa nova notação, as formulações de média e desvio padrão são apresentadas a seguir:

$$\bar{x} = \frac{x_1 n_1 + x_2 n_2 + \dots + x_k n_k}{n} = \frac{1}{n} \sum_{j=1}^k x_j n_j \quad (3.9)$$

$$s = \sqrt{\frac{1}{n-1} \sum_{j=1}^k n_j (x_j - \bar{x})^2} = \sqrt{\frac{1}{n-1} \left\{ \sum_{j=1}^k (x_j^2 n_j) - n \bar{x}^2 \right\}} \quad (3.10)$$

A Tabela 3.4 mostra a sequência de cálculos para a obtenção da média e do desvio padrão, usando as notas finais dos alunos da Turma A.

Tabela 3.4 Cálculos auxiliares para a obtenção de \bar{x} e s .

Nota x_j	Frequência n_j	$x_j n_j$	$x_j^2 n_j$
4	1	4	16
5	2	10	50
6	2	12	72
7	2	14	98
8	1	8	64
Total	8	48	300

$$\text{Assim: } \bar{x} = \frac{48}{8} = 6 \quad \text{e} \quad s = \sqrt{\frac{300 - 8(6)^2}{7}} = 1,31$$

Nas situações em que existem muitas repetições de valores, as expressões (3.9) e (3.10) não só facilitam o cálculo de \bar{x} e s , como também reduzem a possibilidade de erros computacionais.

O coeficiente de variação

Embora o desvio padrão seja a medida de dispersão mais usada, ela mede a dispersão em termos absolutos. O coeficiente de variação, definido por

$$cv = \frac{s}{\bar{x}} \quad (3.11)$$

mede a variação em termos relativos. Veja, por exemplo, os três conjuntos de valores apresentados na Tabela 3.5 e seus respectivos desvios padrões.

Tabela 3.5 Média, desvio padrão e coeficiente de variação de três conjuntos de valores.

Conjunto de valores	\bar{x}	s	cv
1) 1 2 3	2	1	0,5
2) 101 102 103	102	1	0,01
3) 100 200 300	200	100	0,5

Os conjuntos (1) e (2) têm o mesmo desvio padrão, pois os intervalos entre os valores são iguais. Por outro lado, os intervalos entre os valores do conjunto (3) são 100 vezes maiores que nos outros conjuntos. Portanto, o desvio padrão no conjunto (3) é 100 vezes maior que os dos outros. Note, porém, que os níveis de variabilidade nos conjuntos (1) e (3) são proporcionalmente iguais; logo, eles têm o mesmo coeficiente de variação.

Ao dividirmos o desvio padrão pela média, a unidade de medida é cancelada. Logo, o coeficiente de variação é adimensional (não tem unidade de medida), tornando-se útil quando queremos comparar a variabilidade de observações com diferentes unidades de medidas. Cabe observar que o cv não faz sentido em variáveis que assumem valores com ambos os sinais (positivo e negativo).

3.4.2 Medidas baseadas na ordenação dos dados

A média e o desvio padrão são as medidas mais usadas para avaliar a posição central e a dispersão de um conjunto de valores. Contudo, essas medidas são fortemente influenciadas por *valores discrepantes*. Por exemplo, nas notas da turma C (Exemplo 3.4), o valor discrepante 0 (zero) *puxa* a média para baixo, como ilustra a Figura 3.17. Apesar de a média aritmética ser 6, o diagrama de pontos sugere que o valor 7 é um valor *mais típico* para representar as notas da turma, pois, além de ser o valor *mais frequente*, ele é o *valor do meio*, deixando metade das notas abaixo dele, e a outra metade acima dele.

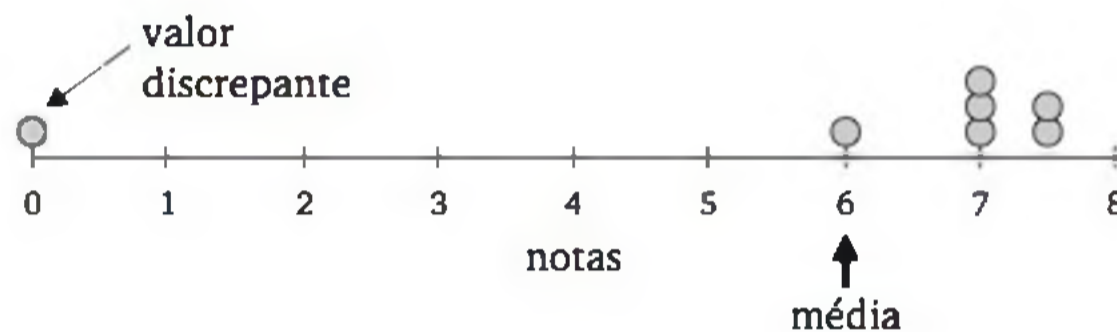


Figura 3.17 *Influência de um valor discrepante no cálculo da média aritmética.*

Nesta seção, apresentaremos algumas medidas que são menos afetadas por valores discrepantes e, em consequência, são mais recomendadas para a análise de dados que possam conter valores discrepantes.

A mediana

A mediana avalia o centro de um conjunto de valores, sob o critério de ser o valor que divide a distribuição ao meio, deixando os 50% menores valores de um lado e os 50% maiores valores do outro lado. Por exemplo, o conjunto de valores $\{2, 3, 4, 5, 8\}$ tem como mediana o valor 4, pois a quantidade de valores menores que 4 é igual a 2, a mesma quantidade de valores superiores a 4.

De forma mais precisa, podemos definir a *mediana* como o valor que ocupa a posição $(n + 1)/2$, considerando os dados ordenados crescente ou decrescentemente. Se $(n + 1)/2$ for fracionário, a mediana é definida como a média dos dois valores de posições mais próximas a $(n + 1)/2$. Vamos representar a mediana por m_d .

Exemplos:

- a) Conjunto de notas da Turma C: $\{0; 6; 7; 7; 7; 7,5; 7,5\}$
 \Rightarrow posição $(n + 1)/2 = 4 \Rightarrow m_d = 7$

- b) $\{5, 3, 2, 8, 4\} \xrightarrow{\text{ordenando}} \{2, 3, 4, 5, 8\}$, posição $(n + 1)/2 = 3 \Rightarrow \Rightarrow m_d = 4$
- c) $\{3, 5, 6, 7, 10, 11\} \Rightarrow$ posição $(n + 1)/2 = 3,5 \Rightarrow m_d = (6 + 7)/2 = = 6,5$

Quando os dados estão apresentados num *ramo-e-folhas*, é muito fácil obter a mediana, pois, neste caso, os valores já estão ordenados. O esquema seguinte ilustra a obtenção da mediana no *ramo-e-folhas* dos dados do Exemplo 3.2.

	Tempo de carga (s)	Frequência acumulada	
4	7 8 8 9 9 9 9	7	$n = 50$ \Rightarrow posição $\frac{n + 1}{2} = \frac{51}{2} = 25,5$ $\Rightarrow m_d = \frac{5,9 + 6,0}{2} = 5,95$
5	0 0 0 1 1 2 3 4 4 4 5 6 6 7 7 7 7 9	25	
6	0 0 2 2 2 3 3 4 5 7 8 8 9	38	
7	0 1 3 3	42	
8	2 2 3 4 9	47	
9	1 9	49	
10		0	
11		0	
12		0	
13		0	
14	1	50	

unidade = 0,1

Comparação entre média e mediana

A Figura 3.18 mostra os valores da média e da mediana num histograma. Note que o valor discrepante da classe 14|— 15 *puxa* mais a média do que a mediana.

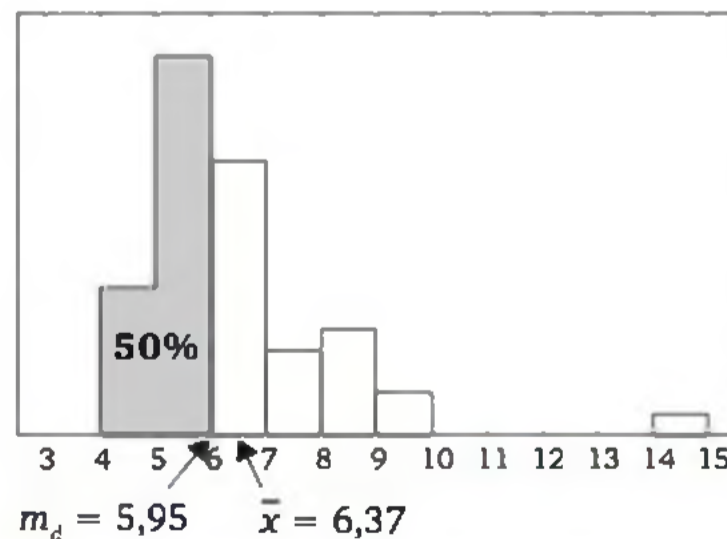


Figura 3.18 Posição da média e da mediana no histograma do tempo de carga de um aplicativo.

Em distribuições simétricas, a média e a mediana são iguais. Em distribuições assimétricas, a média tende a deslocar-se para o lado da cauda mais longa (ver Figura 3.19).

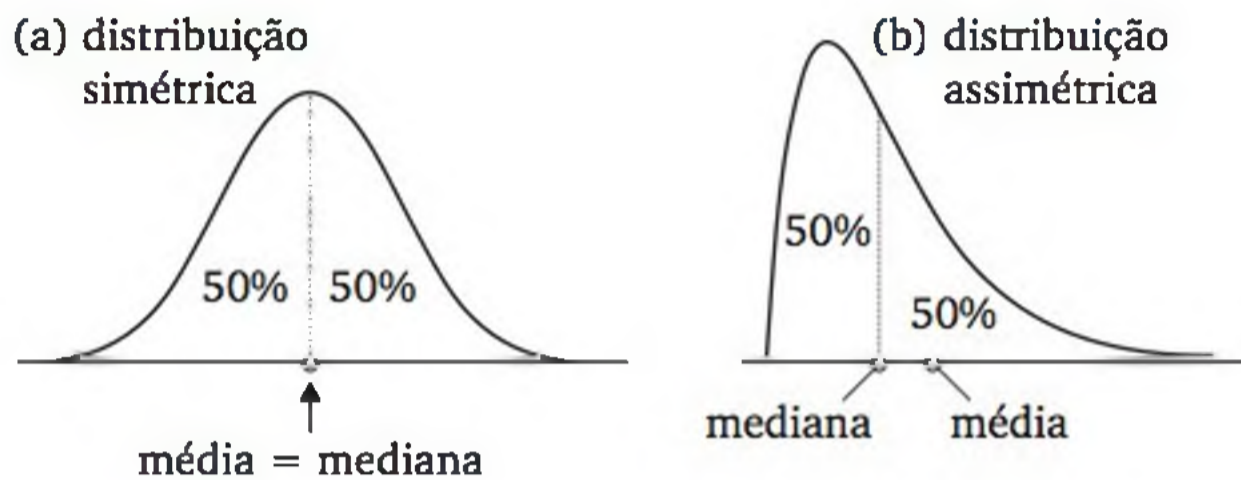


Figura 3.19 Posições da média e mediana, segundo a forma (simétrica ou assimétrica) da distribuição.

Em geral, dado um conjunto de valores, a média é a medida de posição central mais adequada, quando se supõe que esses valores têm uma distribuição razoavelmente simétrica, enquanto a mediana surge como uma alternativa para representar a posição central em distribuições muito assimétricas.¹³ Muitas vezes, calculamos ambas as medidas para avaliar a posição central sob dois enfoques diferentes, além de obtermos uma primeira avaliação sobre a assimetria da distribuição.

Quartis e extremos

Na maioria dos casos práticos, o pesquisador tem interesse em conhecer outros aspectos relativos ao conjunto de valores, além de um valor central, ou valor típico. Algumas informações relevantes podem ser obtidas através do conjunto de medidas: *mediana*, *extremos* e *quartis*, como veremos a seguir.

Chamamos de *extremo inferior* ao menor valor do conjunto de valores, isto é, $\text{mín.}(x_1, x_2, \dots, x_n)$, e de *extremo superior* ao maior valor, isto é $\text{máx.}(x_1, x_2, \dots, x_n)$. Por exemplo, dado o conjunto de valores $\{5, 3, 6, 11, 7\}$, temos $\text{mín.} = 3$ e $\text{máx.} = 11$.

¹³ Mesmo para variáveis que supostamente tenham distribuições razoavelmente simétricas, a média e a mediana podem não se igualar, já que, em geral, estamos observando apenas alguns valores (amostra) dessas variáveis. Para variáveis com distribuições razoavelmente simétricas, a média é a medida de posição central mais adequada, porque usa o máximo da informação contida nos dados. A média é calculada usando propriamente a magnitude dos valores, enquanto a mediana utiliza somente a ordenação dos valores.

Chamamos de *primeiro quartil ou quartil inferior* (q_i) o valor que delimita os 25% menores valores; de *terceiro quartil ou quartil superior* (q_s) o valor que separa os 25% maiores valores. O *segundo quartil, ou quartil do meio*, é a própria mediana, que separa os 50% menores dos 50% maiores valores (ver a Figura 3.20).

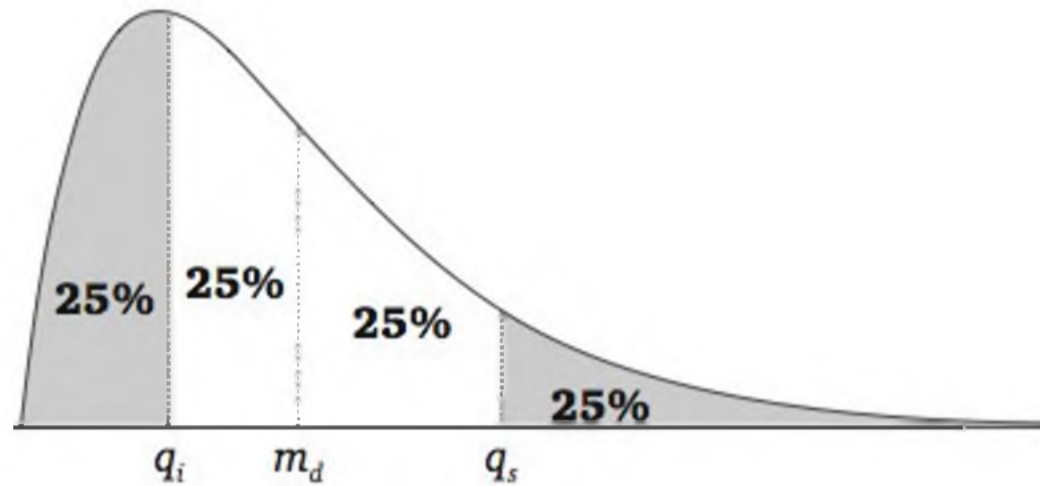


Figura 3.20 Os quartis dividem a distribuição em quatro partes iguais.

Com os dados ordenados crescentemente, temos:

posição de q_i : $\frac{n+1}{4}$	posição de m_d : $\frac{n+1}{2}$	posição de q_s : $\frac{3(n+1)}{4}$
------------------------------------	------------------------------------	---------------------------------------

Quando os resultados das operações acima são fracionários, fazemos uma interpolação linear com os valores de posições vizinhas ao resultado da fração.

Exemplos:

a) Observações: 15, 18, 5, 7, 9, 11, 3, 5, 6, 8, 12. Ordenando:

3 5 5 6 7 8 9 11 12 15 18

$$n = 11 \longrightarrow \text{posição de } q_i: \frac{n+1}{4} = 3 \longrightarrow q_i = 5$$

$$\text{posição de } m_d: \frac{n+1}{2} = 6 \longrightarrow m_d = 8$$

$$\text{posição de } q_s: \frac{3(n+1)}{4} = 9 \longrightarrow q_s = 12$$

b)

Tempo de carga (s)		
4	7 8 8 9 9 9 9	$n = 50$
5	0 0 0 1 1 2 3 4 4 4 5 6 6 7 7 7 7 9	$\Rightarrow \frac{n+1}{2} = 25,5 \Rightarrow m_d = 5,95$
6	0 0 2 2 2 3 3 4 5 7 8 8 9	$\Rightarrow \frac{n+1}{4} = 12,75 \Rightarrow q_i = 5,175$
7	0 1 3 3	$\Rightarrow \frac{3(n+1)}{4} = 38,25 \Rightarrow q_s = 6,925$
8	2 2 3 4 9	
9	1 9	
10		
11		
12		
13	unidade = 0,1	
14	1	

Com a mediana, quartis e extremos, podemos ter informações sobre a posição central, dispersão e assimetria da distribuição de frequências, como ilustra a Figura 3.21.

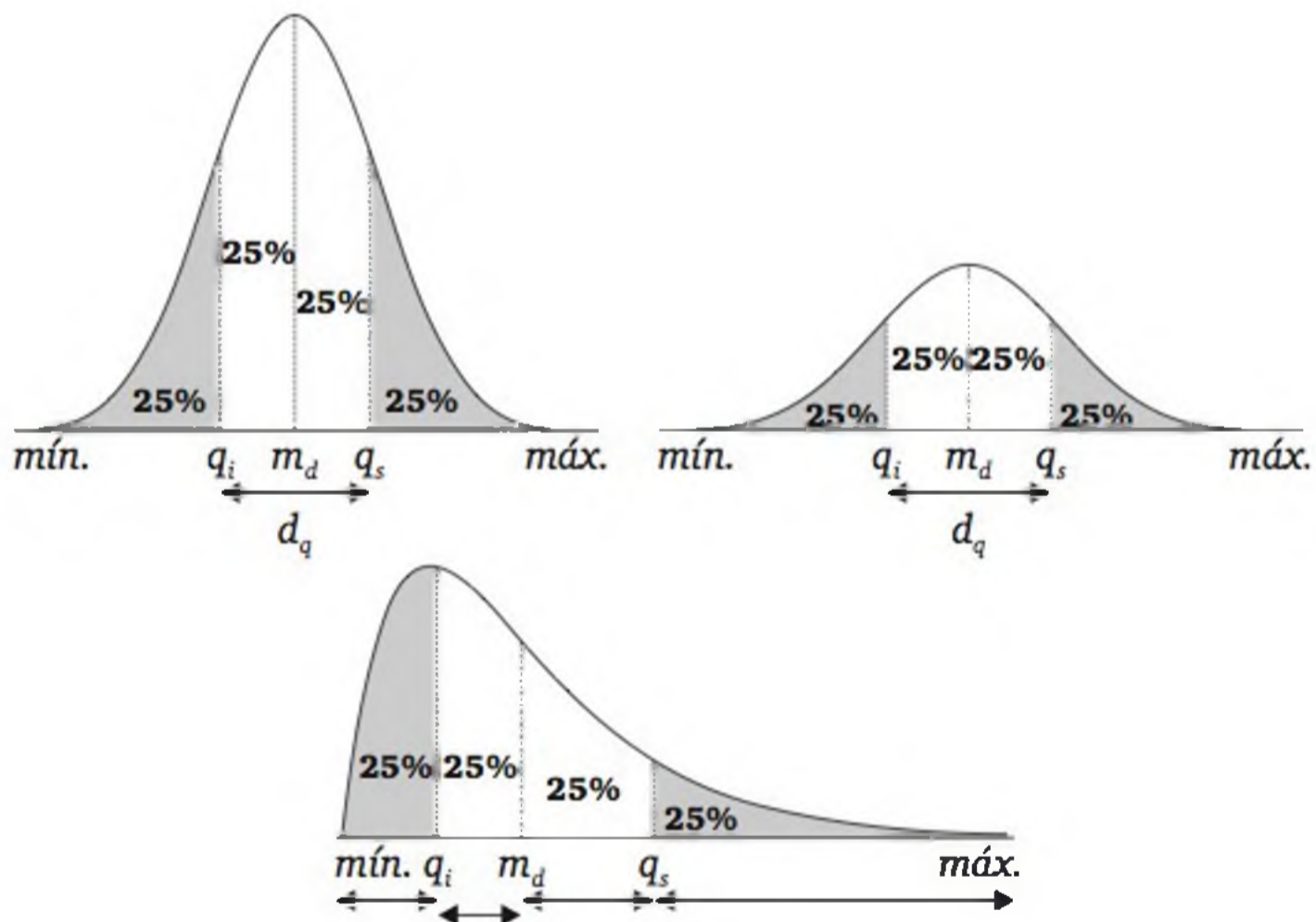


Figura 3.21 Posição dos quartis e extremos em distribuições diferentes quanto à dispersão e assimetria.

O desvio interquartilício ($d_q = q_s - q_i$) é muitas vezes usado como uma medida de dispersão. Veja na Figura 3.21 que, quanto mais dispersa a distribuição,

maior será o valor de d_q . Em distribuições mais dispersas, os valores dos quartis (e dos extremos) ficam mais distantes. Em distribuições simétricas, a distância entre o quartil inferior e a mediana é igual à distância entre a mediana e o quartil superior, enquanto em distribuições assimétricas essas distâncias são diferentes.

Diagrama em caixas

Uma forma de apresentar graficamente os conceitos discutidos é através do *diagrama em caixas*. Trata-se de um retângulo que representa o desvio interquartil. Esse retângulo representa, portanto, a faixa dos 50% dos valores mais típicos da distribuição. O retângulo é dividido no valor correspondente à mediana; assim, ele indica o quartil inferior, a mediana e o quartil superior. Entre os quartis e os extremos, são traçadas linhas. Caso existam valores discrepantes (além de $1,5 d_q$), a linha é traçada até o último valor não discrepante, e os valores discrepantes são indicados por pontos. Eventuais pontos muito distantes (além de $3 d_q$) normalmente são representados por símbolos diferentes para serem bem destacados (veja a Figura 3.22).

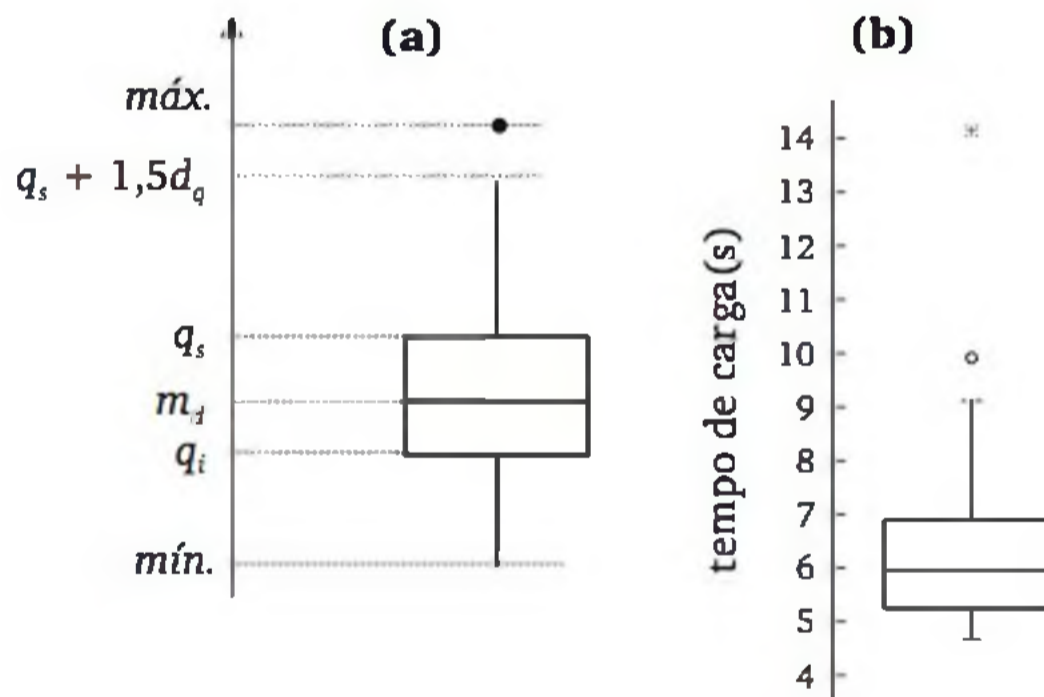


Figura 3.22 (a) Construção de um diagrama em caixas. (b) Diagrama em caixas das 50 observações do tempo de carga de um aplicativo (Exemplo 3.2).

A Figura 3.23 mostra a forma do *diagrama em caixas* para uma distribuição simétrica e para uma distribuição assimétrica. Note as diferenças e imagine como ficaria um *diagrama em caixas* se tivéssemos uma distribuição mais dispersa.

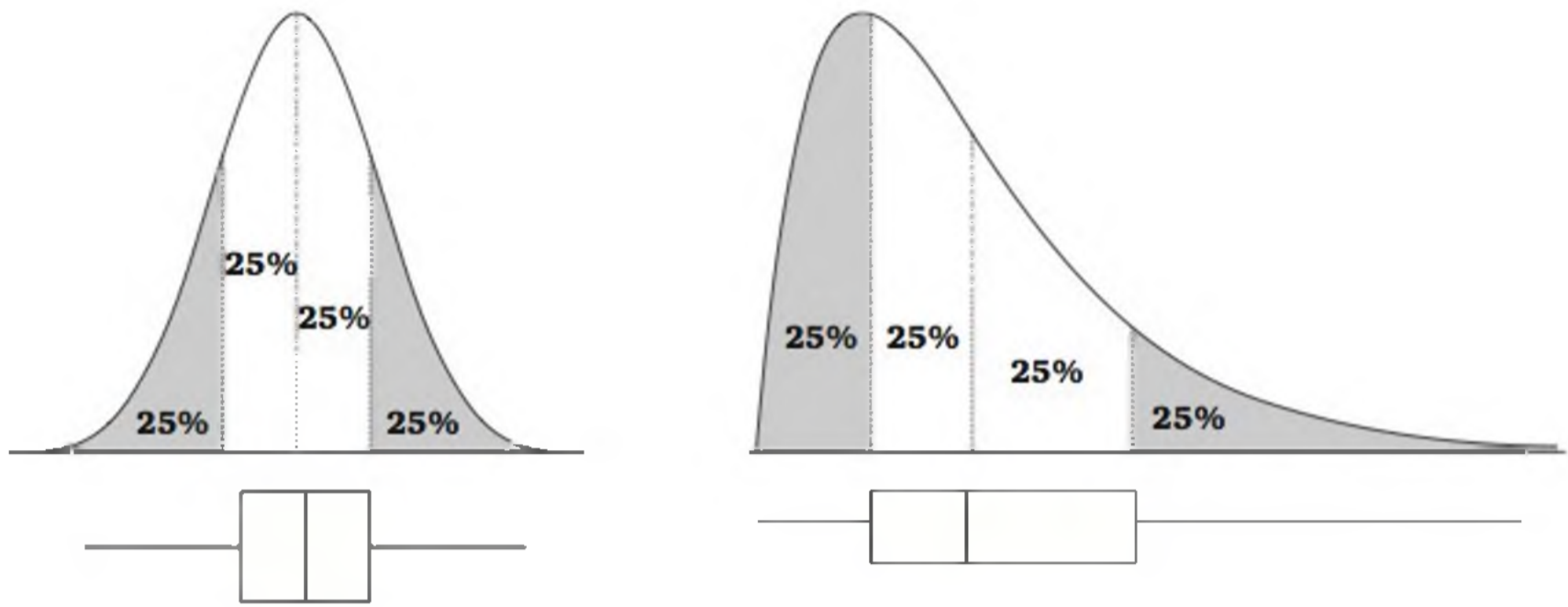


Figura 3.23 Diagrama em caixas e forma da distribuição.

Exemplo 3.5 Para avaliação da qualidade, foram pesados 228 sacos de leite tipo C, em cada boca de ensacamento, durante um mês.¹⁴ Os diagramas em caixas das amostras são apresentados na Figura 3.24.

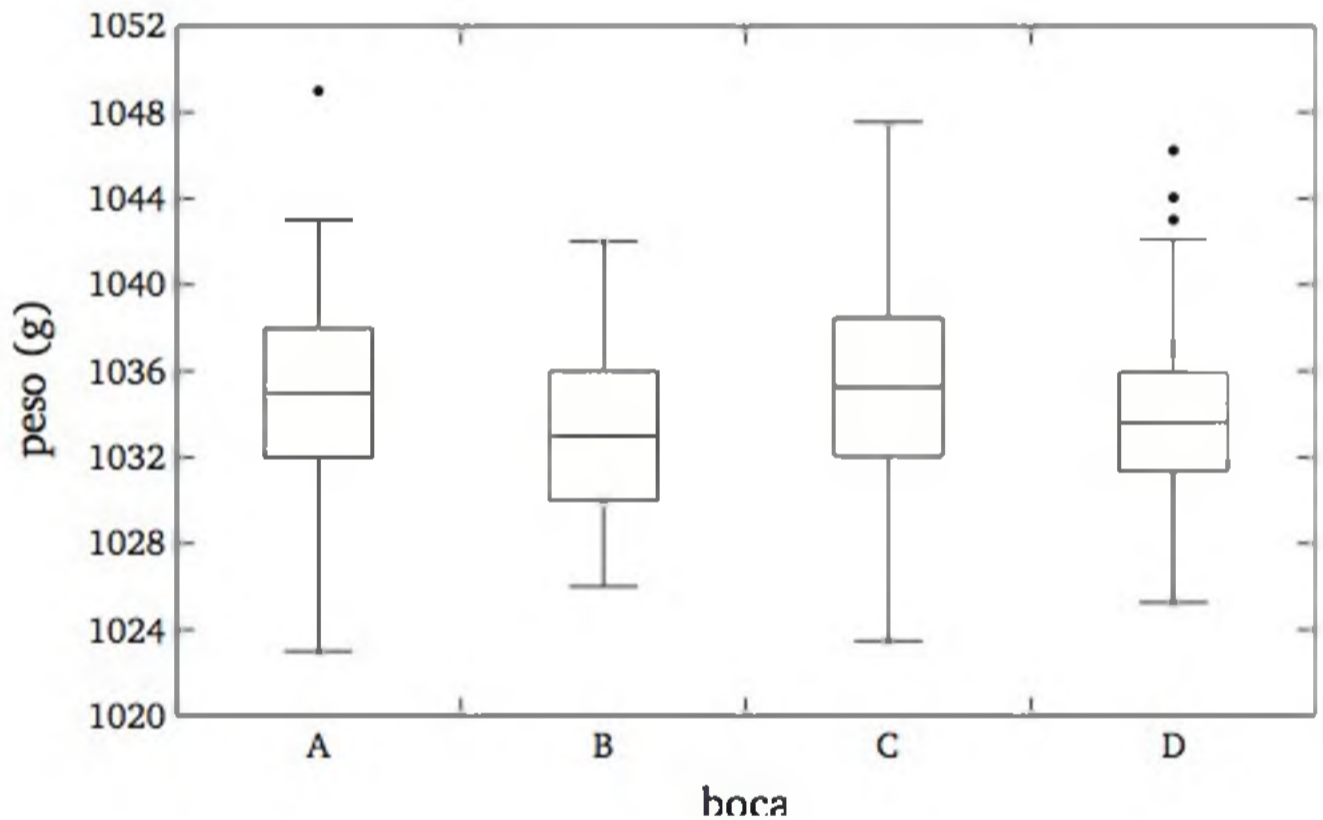


Figura 3.24 Representação de distribuições do peso de litros de leite que saem de um laticínio, por boca de ensacamento.

Através da Figura 3.24, observamos pouca diferença entre as bocas de ensacamento. Aparentemente, as bocas B e D apresentam nível de peso e variabilidade levemente menores do que as bocas A e C. A busca por melhoria da qualidade passa pela redução de variabilidade do processo.

¹⁴ A amostra de observações foi extraída da dissertação de mestrado de Luciana S. C. V. da Silva (Programa de Pós-Graduação em Engenharia de Produção/UFSC, 2001).

3.5 OBSERVAÇÕES AO LONGO DO TEMPO

No acompanhamento da qualidade de um processo, coletamos sistematicamente pequenas amostras do produto que está sendo manufaturado, em diversas fases da produção. Retomando o Exemplo 3.5, suponha que em todos os dias sejam coletadas amostras de dez sacos de leite, os quais têm seu peso medido. Embora diagramas em caixas ou histogramas possam fornecer uma ideia da variabilidade do processo, eles não permitem avaliar certas tendências do processo.

Um procedimento mais adequado é calcular medidas descritivas das amostras e apresentá-las ao longo do tempo. No presente exemplo, calculamos a média (\bar{x}) e o desvio padrão (s) de amostras de 10 sacos de leite, durante 23 dias, conforme ilustrado a seguir:

Amostra 1	Amostra 2	... Amostra 23
1.034	1.035	... 1.031
1.029	1.028	... 1.031
1.032	1.031	... 1.029
1.030	1.035	... 1.036
1.032	1.030	... 1.028
1.033	1.035	... 1.034
1.037	1.029	... 1.031
1.031	1.035	... 1.030
1.033	1.040	... 1.030
1.035	1.036	... 1.031
$\bar{x}_1 = 1.032,60$	$\bar{x}_2 = 1.033,34$	$\bar{x}_{23} = 1.030,99$
$s_1 = 2,39$	$s_2 = 3,79$	$s_{23} = 2,34$

A Figura 3.25 apresenta as medidas descritivas \bar{x}_i e s_i ($i = 1, 2, \dots, 23$) ao longo do tempo, num gráfico de linhas.

A linha horizontal cheia representa a média aritmética da medida descritiva em análise (\bar{x} no primeiro gráfico e s no segundo) das 23 amostras. Os pontos ligados por linhas tracejadas são os resultados das medidas descritivas. As linhas horizontais tracejadas podem ser interpretadas como limites das variações estatisticamente toleráveis. Esses limites baseiam-se na teoria que veremos no Capítulo 7.

Observando os gráficos da Figura 3.25, verificamos que, apesar de as médias das amostras estarem sempre dentro dos limites de controle, a amostra 4 apresentou variabilidade acima do tolerado. Além disso, entre as amostras 1 a 4, as médias amostrais apresentaram tendência crescente. Provavelmente, após termos observado a amostra 4, ocorreu uma intervenção no processo produtivo para corrigir a tendência aparentemente crescente do peso médio de leite e da alta variabilidade. A amostra 7 também merece algum cuidado. Embora esteja dentro dos limites toleráveis, a média amostral apresenta-se bastante baixa, próxima do limite inferior.

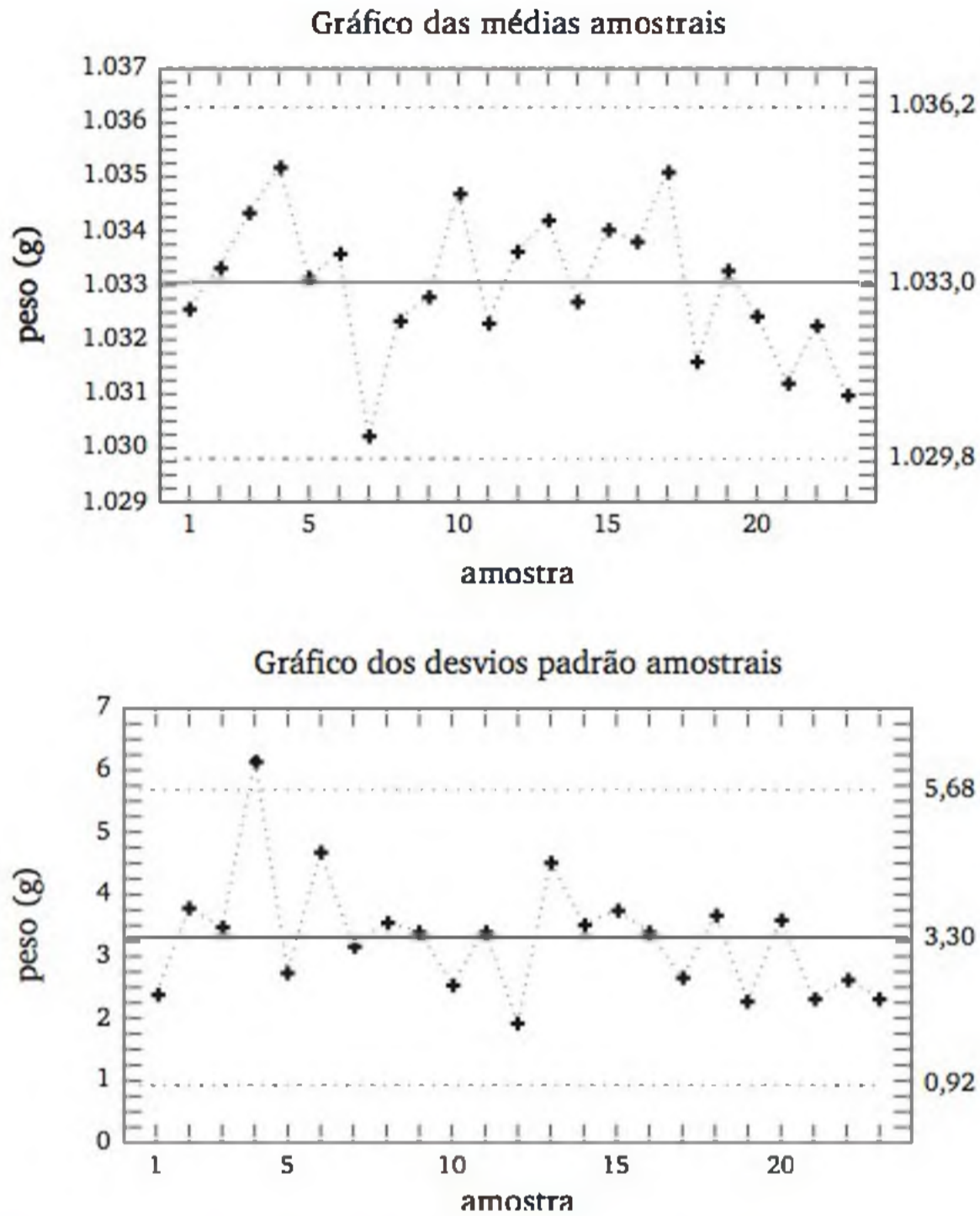


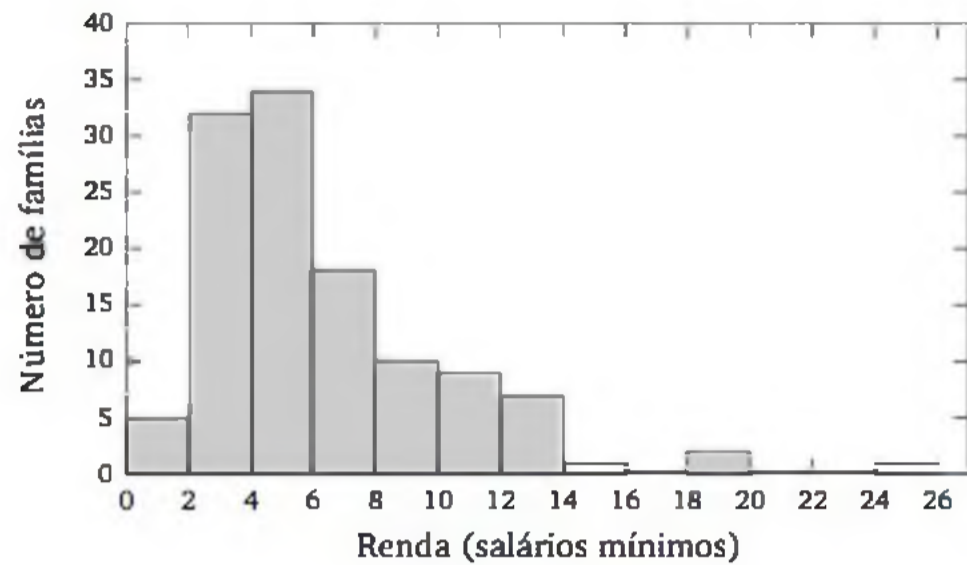
Figura 3.25 Gráficos de controle do peso de sacos de leite que saem de um laticínio, numa boca de ensacamento.

3.6 ANÁLISE EXPLORATÓRIA COM APOIO DO COMPUTADOR

Em geral, nos pacotes computacionais de estatística, ou mesmo em planilhas eletrônicas, é bastante simples obter um conjunto de medidas descritivas dos valores de uma variável quantitativa. A seguir, apresentamos medidas descritivas da renda, em salários mínimos, de uma amostra de famílias de um bairro de Florianópolis. As medidas descritivas foram obtidas através da planilha eletrônica *Excel*. Ao lado, é apresentado o histograma de frequências para facilitar a interpretação.¹⁵

¹⁵ No *Microsoft Excel*, várias técnicas estatísticas podem ser feitas acionando no menu principal “ferramentas”, “suplementos” e solicitando que se instalem as “ferramentas de análise”. Para obter as medidas descritivas, acionar “ferramentas”, “análise de dados” e “estatísticas descritivas”. O histograma foi construído com o apoio do *STATISTICA 5.1*. Ver <www.statsoft.com.br>.

Renda	
Média	6,34
Erro padrão	0,37
Mediana	5,40
Moda	3,90
Desvio padrão	4,03
Variância da amostra	16,26
Curtose	4,55
Assimetria	1,71
Intervalo	25,60
Mínimo	0,10
Máximo	25,70
Soma	754,50
Contagem	119



Em termos de posição central, temos a *média*, a *mediana* e a *moda*. Esta última medida apresenta o valor mais frequente do conjunto de dados. O fato de a média apresentar um valor maior que a mediana e a moda sugere uma distribuição assimétrica, com cauda mais longa para o lado direito, o que é confirmado pelo gráfico. Aliás, na lista de medidas, aparece o chamado *coeficiente de assimetria*, com valor igual a 1,71. Em distribuições simétricas, esse coeficiente aproxima-se de zero. Coeficiente de assimetria positivo (especialmente quando superior à unidade) indica cauda mais longa para o lado direito. Por outro lado, quando negativo (especialmente quando inferior a -1), indica cauda mais longa para o lado esquerdo.

A medida *erro padrão* será apresentada no Capítulo 7. A *curtose* é pouco usada e, por isso, não será discutida neste texto. O *intervalo* é o que definimos, em (3.5), como *amplitude*; e a *contagem* é o número de valores (n) usado no cálculo das medidas descritivas.

3.7 ORIENTAÇÃO GERAL

Em grandes conjuntos de dados, é comum, inicialmente, construir a distribuição de frequências de cada variável e, depois, explorar possíveis associações entre pares de variáveis. Os procedimentos estatísticos dependem do tipo de variável em estudo (ver Figura 3.26).

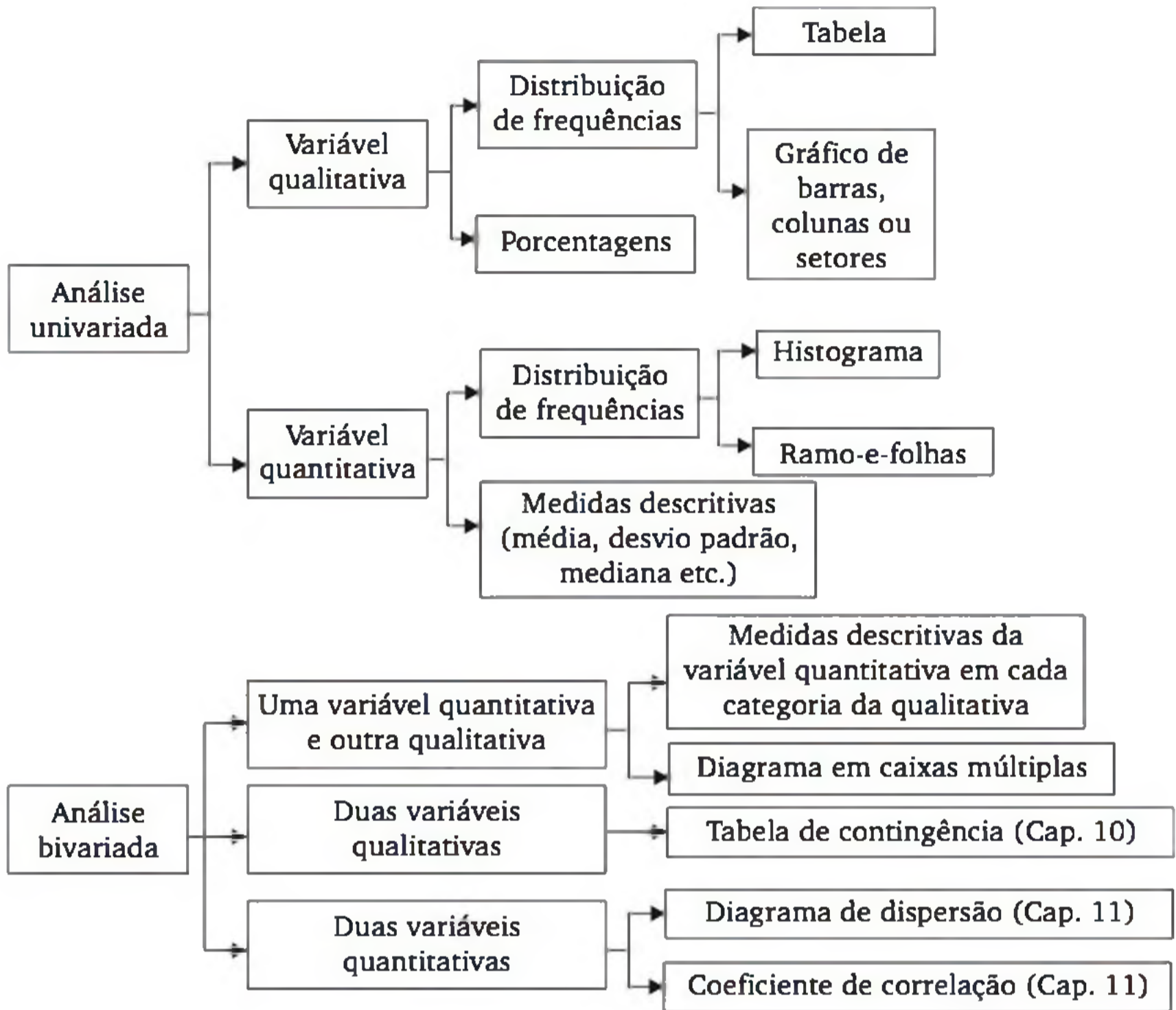


Figura 3.26 Esquema geral para a análise exploratória de dados.

EXERCÍCIOS

3. Dado o seguinte conjunto de dados: {7, 8, 6, 10, 5, 9, 4, 12, 7, 8}, calcule:
 - a) a média e
 - b) o desvio padrão.
4. Calcule a média e o desvio padrão da seguinte distribuição de frequências, a qual se refere ao número de defeitos encontrados em placas de circuito integrado.

Número de defeitos	Frequência
0	30
1	25
2	10
3	5
4	2

5. Considerando o Exercício 2 (seção 3.3), obtenha a mediana e os quartis.
6. Com o objetivo de direcionar campanhas de *marketing*, uma livraria virtual está registrando o número de acessos diários em algumas de suas páginas da *Web*, nos últimos três meses. A tabela, a seguir, mostra medidas descritivas desses registros, em páginas de três categorias de livros.

Livro	Média	Desvio padrão	Quartil inferior	Mediana	Quartil superior
Romance	910	690	412	650	1.500
Ficção	220	180	145	398	1.023
Técnico	630	480	115	190	1.500

- a) Quais as diferenças das três distribuições em termos de posição central e dispersão?
- b) As medidas sugerem distribuições simétricas?
7. Os dados a seguir são leituras da pressão do homogeneizador de um latifício.

Leite tipo C	Leite UHT
3,0 3,1 3,0 3,0 3,0 2,9 2,9 3,0	2,2 2,2 2,3 2,2 2,2 2,2 2,4 2,4
3,1 2,9 3,0 3,0 3,0 3,0 3,0 3,0	2,2 2,4 2,6 2,6 2,4 2,2 2,2 2,8
3,0 3,0 3,0 3,0 2,9	2,6 2,2 2,6 2,4 2,0

Para cada conjunto de dados, calcule as medidas descritivas que você conhece. Com base nessas medidas, comente as principais diferenças entre os dois conjuntos de valores.

EXERCÍCIOS COMPLEMENTARES

8. Bernardin (Mestrado Engenharia Mecânica/UFSC, 1994) realizou um experimento que tinha o objetivo de melhorar a qualidade do processo de

formulação de massa cerâmica para pavimento. Os corpos de prova eram “biscoitos” que saíam do processo de queima e a qualidade era avaliada por três variáveis, a saber: $X_1 = \text{retração linear (\%)}$, $X_2 = \text{resistência mecânica}$ e $X_3 = \text{absorção de água (\%)}$. O experimento foi realizado sob 8 condições diferentes (no estudo original eram 18). Foram feitos 5 ensaios em cada uma das 8 condições experimentais. Os dados são apresentados a seguir:

C^1	X_1	X_2	X_3	C^1	X_1	X_2	X_3	C^1	X_1	X_2	X_3	C^1	X_1	X_2	X_3
1	8,9	41,1	5,5	3	9,4	50,0	0,8	5	13,4	60,6	0,5	7	12,9	41,1	0,2
1	9,2	39,0	4,8	3	9,9	48,3	0,6	5	13,4	60,0	0,5	7	12,4	39,0	0,4
1	8,0	36,9	6,2	3	9,6	50,1	0,6	5	13,6	68,4	0,2	7	12,6	36,9	0,5
1	8,7	39,2	5,7	3	9,2	49,9	0,7	5	13,4	60,8	0,7	7	12,6	39,2	0,4
1	8,7	35,9	5,5	3	9,4	56,2	0,5	5	12,4	51,4	1,0	7	12,9	35,9	0,3
2	12,6	52,7	0,9	4	6,6	31,2	9,0	6	9,6	41,2	3,9	8	8,2	40,8	4,4
2	13,6	53,5	0,4	4	6,4	25,3	10,2	6	10,6	53,0	4,5	8	9,2	43,8	3,9
2	11,6	47,0	1,3	4	5,9	22,8	10,5	6	8,9	37,0	3,3	8	9,2	48,6	4,0
2	10,1	31,1	1,8	4	5,9	27,5	10,6	6	7,5	30,1	3,0	8	8,5	46,9	4,3
2	12,1	50,9	1,1	4	6,8	31,9	9,3	6	8,9	41,6	3,5	8	8,7	46,2	4,1

¹ C = condição experimental.

- Como as variáveis X_1 , X_2 e X_3 podem ser classificadas (qualitativas, quantitativas discretas ou quantitativas contínuas)?
- Apresente a distribuição de frequências de X_1 através de um diagrama ramo-e-folhas. Comente a forma da distribuição.
- Apresente as distribuições de frequências de X_2 e X_3 através de histogramas. Comente as formas das distribuições.
- Calcule a média e o desvio padrão de X_3 para cada condição experimental (por simplicidade, considere apenas as condições 1, 4 e 8). Quais as informações que podem ser extraídas com estas medidas?
- Construa diagramas de pontos para X_3 nas condições experimentais 1, 4 e 8. As informações fornecidas por esses diagramas são iguais às obtidas no item anterior?
- Calcule a mediana e quartis de X_1 (sugestão: use o diagrama ramo-e-folhas do item (b)).
- Construa um diagrama em caixas para X_1 .

- h)** Considere o objetivo de verificar qual das variáveis (X_1 , X_2 e X_3) apresenta maior variabilidade. Qual medida de dispersão você deve usar?
- 9.** Com respeito ao exercício anterior, o estudo da variabilidade natural do processo, em termos das 50 observações de X_1 , X_2 e X_3 , fica prejudicado, pois os ensaios foram feitos sob 8 condições experimentais diferentes. Considerando, porém, que, para cada variável, a média \bar{x}_j corresponde a uma estimativa da j -ésima condição experimental ($j = 1, 2, \dots, 8$), então os desvios $d_{ij} = x_{ij} - \bar{x}_j$ ($i = 1, 2, \dots, 5$) fornecem informações da variabilidade natural do processo. A tabela seguinte apresenta esses desvios.

d_1	d_2	d_3	d_1	d_2	d_3	d_1	d_2	d_3	d_1	d_2	d_3
0,20	2,68	-0,04	-0,10	-0,90	0,16	0,16	0,36	-0,08	0,22	2,68	-0,16
0,50	0,58	-0,74	0,40	-2,60	-0,04	0,16	-0,24	-0,08	-0,28	0,58	0,04
-0,70	-1,52	0,66	0,10	-0,80	0,04	0,36	8,16	0,38	-0,08	-1,52	0,14
0,00	0,78	0,16	-0,30	-1,00	0,06	0,16	0,56	0,12	-0,08	0,78	0,04
0,00	-2,52	-0,04	-0,10	5,30	-0,14	-0,84	-8,84	0,42	0,22	-2,52	-0,06
0,60	5,66	-0,20	0,28	3,46	-0,92	0,50	0,62	0,26	-0,56	-4,46	0,26
1,60	6,46	-0,70	0,08	-2,44	0,28	1,50	12,42	0,86	0,44	-1,46	-0,24
-0,40	-0,04	0,20	-0,42	-4,94	0,58	-0,20	-3,58	-0,34	0,44	3,34	-0,14
-1,90	-15,94	0,70	-0,42	-0,24	0,68	-1,60	-10,48	-0,64	-0,26	1,64	0,16
0,10	3,86	0,00	0,48	4,16	-0,62	-0,20	1,02	-0,14	-0,06	0,94	-0,04

Apresente as distribuições de frequências de d_1 , d_2 e d_3 através de histogramas ou ramo-e-folhas. O que você pode dizer da dispersão dessas distribuições comparadas às distribuições construídas no exercício anterior, itens (b) e (c)?

- 10.** Os dados abaixo apresentam a distância (em km) entre a residência e o local de trabalho dos funcionários da empresa AAA.

1,8 2,5 0,4 1,9 4,4 2,2 3,5 0,2 0,9 1,4
 1,1 1,7 1,2 2,3 1,9 0,8 1,5 1,7 1,4 2,1
 3,2 15,1 2,1 1,4 0,5 0,9 1,7 0,5 0,8 3,7
 1,4 1,8 2,0 1,1 1,0 0,8

- a)** Apresente esses dados em *ramo-e-folhas*.
- b)** Na empresa BBB, a distância (em km) até a residência de seus 300 funcionários apresenta as seguintes medidas descritivas:

Mediana = 2,8 Quartil inferior = 1,6 Quartil superior = 4,2

Extremo inferior = 0,4 Extremo superior = 8,8

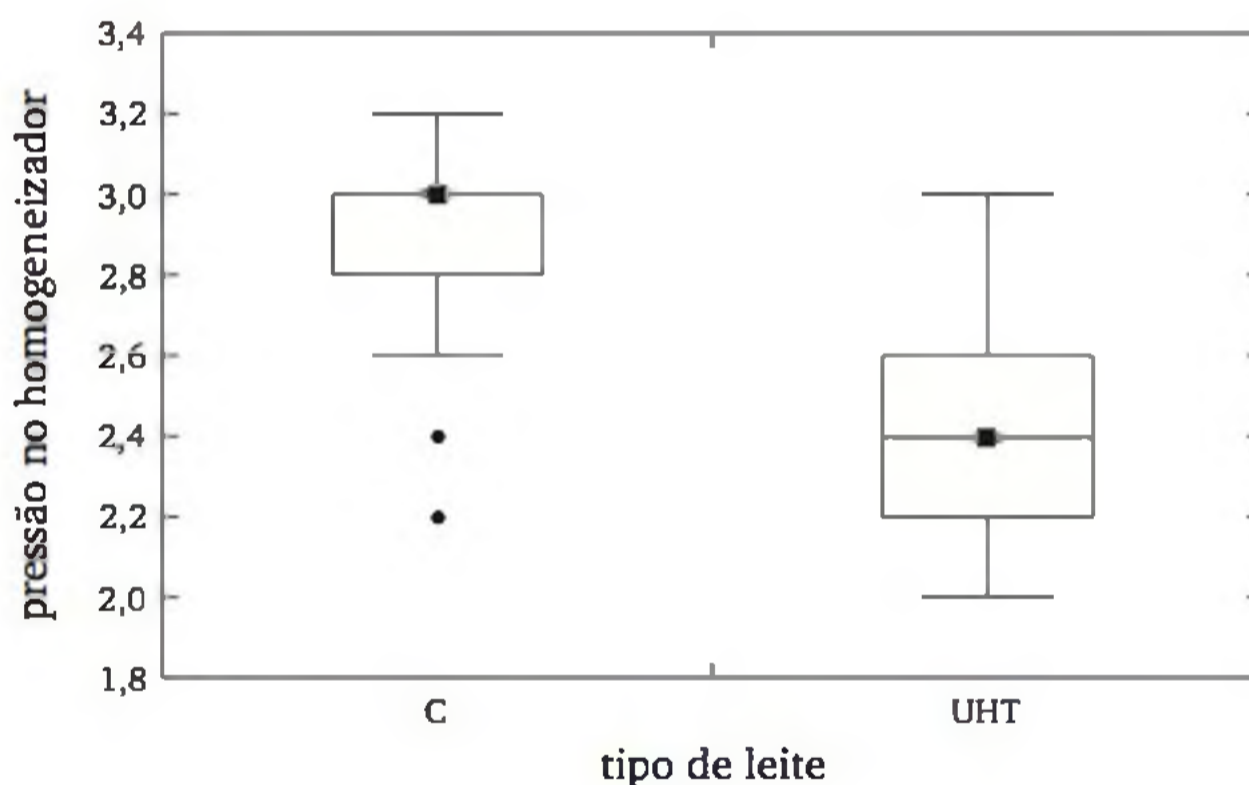
Quais as principais diferenças entre as empresas AAA e BBB, em termos da distância entre a residência e o local de trabalho dos funcionários?

11. Apresentam-se, abaixo, algumas medidas descritivas da distribuição de salários, em R\$, de três empresas do mesmo ramo.

Empresa	Média	Desvio padrão	Extremo inferior	Quartil inferior	Mediana	Quartil superior	Extremo superior
A	300	100	100	200	302	400	510
B	400	180	100	250	398	550	720
C	420	350	100	230	300	650	10.000

O que se pode dizer sobre a distribuição dos salários nas três empresas? Quais as diferenças em termos da posição central, dispersão e assimetria?

12. Cada diagrama em caixas da figura a seguir foi construído com 95 leituras da pressão do homogeneizador. Discuta as diferenças.



A decorative graphic consisting of several curved lines in shades of gray, with the number '4' prominently displayed in the center.

Probabilidade

No capítulo anterior, procuramos conhecer a variabilidade de algum processo com base em observações das variáveis pertinentes. Nestes três próximos capítulos, continuaremos a estudar os processos que envolvem variabilidade, aleatoriedade ou incerteza, mas procuraremos construir modelos matemáticos para facilitar a análise. Esses modelos normalmente são construídos a partir de suposições sobre o processo, mas podem também basear-se em dados observados no passado.

O leitor já deve ter ouvido falar de *probabilidade* ou de *modelos probabilísticos*. Há dois aspectos a considerar. O primeiro é que intuitivamente as pessoas procuram tomar decisões em função dos fatos que têm *maior probabilidade* de ocorrer. Veja os seguintes exemplos:

- a) se o céu está nublado, então há chance considerável de chover. Deve-se levar um guarda-chuva ao sair de casa!;
- b) se um inspetor de qualidade está observando as peças produzidas por uma máquina, e verifica que elas estão saindo fora do padrão, então ele pode deduzir que existe alta chance de essa máquina continuar produzindo peças fora do padrão. Logo, a máquina deverá receber atenção especial;
- c) se em determinada família há muitos casos de doença cardíaca, então há maior chance de pessoas daquela família serem afetadas; portanto, os exames preventivos precisam ser feitos mais frequentemente.

O segundo aspecto é a *incerteza* inerente às decisões que podem ser tomadas sobre determinado problema.

- a) por mais nublado que o céu esteja, pode não chover, ao menos durante o período de tempo em que a pessoa estiver fora de casa;
- b) algumas peças poderiam estar fora do padrão por motivos meramente casuais. O processo pode estar funcionando bem;
- c) apesar dos vários precedentes familiares, uma pessoa pode viver a vida inteira sem ter problemas cardíacos.

Se for possível quantificar a incerteza associada a cada fato, algumas decisões tornam-se mais fáceis. Veja os casos a seguir:

- a) Qual deve ser a capacidade instalada de uma usina hidrelétrica, em função da vazão e da precipitação pluviométrica?
- b) Qual deve ser a capacidade do servidor de comércio eletrônico de uma empresa, em função da demanda prevista?

No caso (a), se for possível prever as variações na quantidade de chuva e, no caso (b), se houver previsão da demanda, podemos responder melhor às questões que foram colocadas. A teoria do cálculo de probabilidades permite obter uma quantificação da incerteza associada a um ou mais fatos e, portanto, é extremamente útil no auxílio à tomada de decisões.

Os modelos probabilísticos são aplicados em situações que envolvem algum tipo de *incerteza* ou *variabilidade*. Mais especificamente, consideraremos a presença de algum *experimento aleatório* como princípio para a construção de modelos probabilísticos.

Exemplo 4.1 São exemplos de experimentos aleatórios:

- a) o lançamento de um dado e a observação da face voltada para cima; não sabemos exatamente qual face vai ocorrer, apenas que será uma das seis existentes. Além disso, se o dado for não viciado e o lançamento imparcial, todas as faces têm a mesma chance de ocorrer;
- b) a observação dos diâmetros, em mm, de eixos produzidos em uma metalúrgica; sabemos que as medidas devem estar próximas de um valor nominal, mas não sabemos exatamente qual é o diâmetro de cada eixo antes de efetuar as mensurações;
- c) o número de mensagens que são transmitidas corretamente por dia em uma rede de computadores; sabemos que o mínimo possível é zero, mas não sabemos nem sequer o número máximo de mensagens que serão transmitidas.

Nos casos em que os possíveis resultados de um experimento aleatório podem ser listados (caso discreto), um **modelo probabilístico** pode ser entendido como a listagem desses resultados, acompanhados de suas respectivas probabi-

lidades. A Figura 4.1 ilustra as etapas para a construção de um modelo probabilístico.

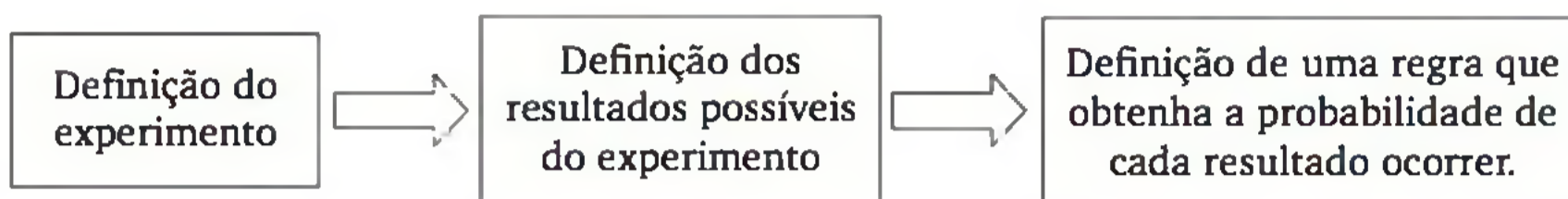


Figura 4.1 Passos para a construção de um modelo probabilístico (caso discreto).

4.1 ESPAÇO AMOSTRAL E EVENTOS

Seja um experimento aleatório qualquer.

O conjunto de *todos* os possíveis resultados do experimento é chamado de **espaço amostral** e é denotado pela letra grega Ω .

Exemplo 4.2 Seguem alguns experimentos aleatórios com os respectivos espaços amostrais:

- a) lançamento de um dado e observação da face voltada para cima: $\Omega = \{1, 2, 3, 4, 5, 6\}$;
- b) retirada de uma carta de um baralho comum (52 cartas) e observação do naipe: $\Omega = \{\text{copas, espadas, ouros, paus}\}$;
- c) o número de mensagens que são transmitidas corretamente por dia em uma rede de computadores: $\Omega = \{0, 1, 2, 3, \dots\}$;¹
- d) a observação do diâmetro, em mm, de um eixo produzido em uma metalúrgica: $\Omega = \{d, \text{ tal que } d > 0\}$.²

O espaço amostral pode ser:

1. *finito*, formado por um número limitado de resultados possíveis, como nos casos (a) e (b);
2. *infinito enumerável*, formado por um número infinito de resultados, os quais podem ser listados, como no caso (c); ou

¹ Note que não há um limite superior conhecido, mas somente é possível a ocorrência de valores inteiros.

² Não há um limite superior e, teoricamente, pode haver uma infinidade de valores.

3. *infinito*, formado por intervalos de números reais, como no caso (d).

Um espaço amostral é dito **discreto** quando for finito ou infinito enumerável; é dito **contínuo** quando for infinito, formado por intervalos de números reais.

Os elementos para se tomar alguma decisão podem corresponder a um conjunto de resultados (ou *evento*) associados ao experimento aleatório. Por exemplo, se o diâmetro D de um eixo, em mm, que sai da linha de produção, pertencer ao conjunto (ou *evento*) $A = \{49,0 \leq D \leq 51,0\}$, então se decide que ele é adequado.

Chamamos de **evento** qualquer subconjunto do espaço amostral:
 A é um evento $\Leftrightarrow A \subseteq \Omega$

Exemplo 4.3 Seja o experimento do lançamento de um dado. Temos: $\Omega = \{1, 2, 3, 4, 5, 6\}$. São exemplos de eventos:

$$A = \text{número par do dado} = \{2, 4, 6\};$$

$$B = \text{número maior que 2 do dado} = \{3, 4, 5, 6\};$$

$$C = \text{número 6} = \{6\}.$$

Dizemos que um evento ocorre quando um dos resultados que o compõem ocorre. Com respeito ao Exemplo 4.3, se o dado for lançado e ocorrer o número 4, então ocorrem os eventos A e B , mas não ocorre o evento C .

Como um evento é um subconjunto do espaço amostral, então todos os conceitos da teoria de conjuntos podem ser aplicados a eventos. Considerando A e B eventos quaisquer, veja as principais operações na Figura 4.2.

Operação	Notação	Conjunto	Evento
a) união	$A \cup B$	reúne os elementos de ambos os conjuntos	ocorre quando ocorrer pelo menos um deles (A , B ou ambos)
b) interseção	$A \cap B$	formado somente pelos elementos que estão em A e B	ocorre quando ocorrer ambos os eventos (A e B)
c) complementar	\bar{A}	formado pelos elementos que não estão em A	ocorre quando não ocorrer o evento A (<i>não</i> A)

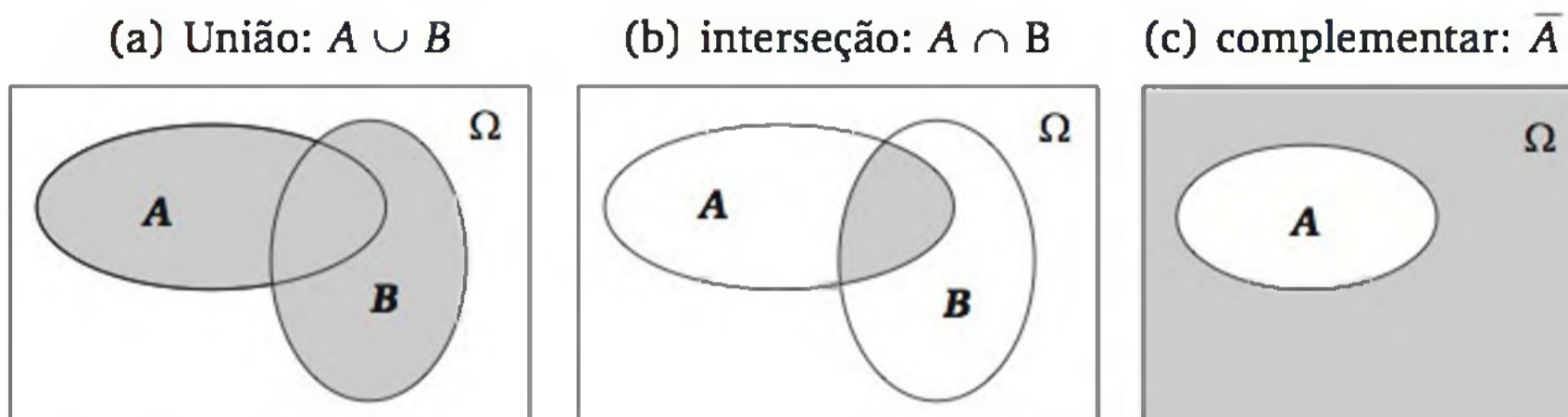


Figura 4.2 Principais operações entre eventos e representações gráficas.

Exemplo 4.3 (continuação). Sejam:

$A = \text{número par do dado} = \{2, 4, 6\};$

$B = \text{número maior que 2 do dado} = \{3, 4, 5, 6\};$

$C = \text{número 6} = \{6\}.$

Eventos complementares:

$\bar{A} = \text{número ímpar do dado} = \{1, 3, 5\};$

$\bar{B} = \text{número menor ou igual a 2 do dado} = \{1, 2\};$

$\bar{C} = \text{não 6} = \{1, 2, 3, 4, 5\}.$

Algumas uniões:

$$A \cup B = \{2, 3, 4, 5, 6\}; \quad A \cup C = \{2, 4, 6\}; \quad \bar{A} \cup A = \Omega$$

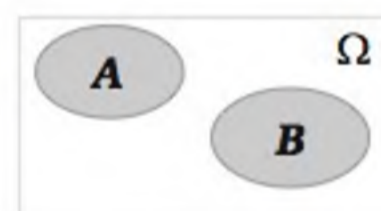
Algumas interseções:

$$A \cap B = \{4, 6\}; \quad A \cap C = \{6\}; \quad \bar{A} \cap A = \{\} = \emptyset.$$

Eventos são ditos **mutuamente exclusivos** se e só se eles não puderem ocorrer simultaneamente. Então, para dois eventos quaisquer, A e B , temos:

A e B são mutuamente exclusivos $\Leftrightarrow A \cap B = \emptyset.$

No Exemplo 4.3, os eventos $\bar{A} = \{1, 3, 5\}$ e $C = \{6\}$ são mutuamente exclusivos, pois eles não podem ocorrer simultaneamente (observe que $\bar{A} \cap C = \emptyset$).



EXERCÍCIOS

1. Apresente os espaços amostrais dos seguintes experimentos aleatórios:
 - a) Lançamento de uma moeda honesta e observação da face voltada para cima.
 - b) Observação da qualidade de peças produzidas, registrando o número de peças defeituosas.
 - c) Contagem do número de clientes numa fila única de banco, que chegam durante uma hora.
 - d) Medição da velocidade do vento, em km/h, na pista de um aeroporto.
 - e) Medição da temperatura, em graus Celsius, numa estação meteorológica da cidade de Florianópolis.

2. Considere que você vai cronometrar o tempo, em segundos, para carregar uma página da *web*.
 - a) Represente, em forma de conjuntos, os seguintes eventos:

$A = \text{mais do que 5 e, no máximo, 10 segundos};$
 $B = \text{mais do que 10 segundos};$
 $C = \text{mais do que 8 segundos};$
 $D = A \cup B; \quad E = A \cap B, \quad F = A \cap C, \quad G = \bar{A}$
 - b) Represente geometricamente (como intervalos na reta dos reais) os conjuntos do item anterior.

4.2 DEFINIÇÕES DE PROBABILIDADE

Intuitivamente, as pessoas sabem como calcular algumas probabilidades para tomar decisões. Observe os seguintes exemplos.

Exemplo 4.4

- a) Vamos supor que você fez uma aposta com um amigo. O vencedor será aquele que acertar a face que ficar para cima, no lançamento de uma moeda *honest*.³ Qual é a probabilidade de você ganhar?

Intuitivamente, você responderia que a probabilidade de ganhar é igual a 50% (ou $\frac{1}{2}$).

³ Usaremos a expressão *moeda honesta* para referenciar uma moeda perfeitamente equilibrada e lançamentos imparciais. De forma análoga, usaremos o adjetivo *honesto* para dado, baralho etc.

- b)** Você continua apostando com o mesmo amigo. O vencedor será aquele que acertar o naipe de uma carta que será retirada, ao acaso, de um baralho comum de 52 cartas. Qual é a probabilidade de você ganhar?

Novamente, de forma intuitiva, você responderia que é de 25% (ou $\frac{1}{4}$).

O que há em comum entre as situações (a) e (b) do Exemplo 4.4? Refletindo um pouco, você observará que em ambas as situações temos experimentos aleatórios. A cada realização do experimento *apenas* um dos resultados possíveis pode ocorrer. Além disso, como se supõe que a moeda e o baralho são *honestos*, cada um dos resultados possíveis tem a *mesma* probabilidade de ocorrer.

4.2.1 Definição clássica de probabilidade

Se um experimento aleatório tem n resultados *igualmente prováveis*, e n_A desses resultados pertencem a certo evento A , então a probabilidade de ocorrência do evento A será:

$$P(A) = \frac{n_A}{n} \quad (4.1)$$

Exemplo 4.4 (continuação)

- a)** No caso da moeda, há apenas dois resultados possíveis e igualmente prováveis, resultando que a probabilidade de ocorrência de uma das faces será igual a $\frac{1}{2}$ (ou 50%).
- b)** No caso dos naipes do baralho, há quatro resultados possíveis e igualmente prováveis, resultando que a probabilidade de ocorrência de um deles será igual a $\frac{1}{4}$ (ou 25%). Analogamente, podemos considerar cada carta do baralho como um resultado. Nesse caso, a probabilidade de ocorrer certo naipe é de $\frac{13}{52} = \frac{1}{4}$.

4.2.2 Definição experimental de probabilidade

Muitas vezes, a alocação de probabilidades baseia-se em observações do passado. Seja um experimento aleatório com espaço amostral Ω e um evento A de interesse. Suponha que esse experimento seja repetido n vezes e o evento A ocorreu $n(A)$ vezes. A frequência relativa do evento A é dada por:

$$f(A) = \frac{n(A)}{n} \quad (4.2)$$

À medida que o experimento é repetido mais e mais vezes, sob as mesmas condições, a frequência relativa do evento A tenderá a ficar cada vez mais próxima da probabilidade de ocorrência do evento A . Mais especificamente:

$$P(A) = \lim_{n \rightarrow \infty} f(A) = \lim_{n \rightarrow \infty} \frac{n(A)}{n} \quad (4.3)$$

Exemplo 4.5 Um fabricante de lâmpadas fluorescentes precisa especificar o tempo de garantia de um de seus modelos. Embora os projetistas estimem que o tempo médio de vida do modelo seja de 5.000 horas, não se sabe exatamente como as lâmpadas irão comportar-se. E sem esse conhecimento seria temerário especificar o tempo de garantia.

Ao definir o experimento aleatório como ligar a lâmpada e registrar o tempo (em horas) que ela funciona, o espaço amostral é formado pelo conjunto de todos os valores maiores ou iguais a zero, ou seja:

$$\Omega = \{t, \text{ tal que } t \geq 0\}$$

Seja o evento:

$$A_t = \text{a lâmpada funcionar até o tempo } t$$

Podemos repetir o experimento com um número n suficientemente grande de lâmpadas.⁴ Com os resultados do experimento, podemos calcular as frequências relativas:

$$f(A_t) = \frac{n(A_t)}{n} \quad (4.4)$$

para diversos valores de t . Essas frequências relativas podem ser usadas como valores aproximados das probabilidades $P(A_t)$ e, assim, definir adequadamente o tempo de garantia, de tal forma que ele não seja demasiadamente longo, pois aí seria necessário repor muitas lâmpadas (custo financeiro alto), mas também não seja muito curto, o que pode gerar a suspeita de um produto com baixa qualidade, acarretando perda de mercado.

Em muitas situações, é impossível realizar o experimento diversas vezes. Veja o exemplo seguinte.

Exemplo 4.6 Quando estudamos o regime de vazões de um rio, com o objetivo de avaliar a viabilidade da construção de uma usina hidrelétrica, não é possível replicarmos os diversos meses e anos, fenômenos climáticos e eventual in-

⁴ Há métodos estatísticos para calcular n (tamanho da amostra), conforme será visto no Capítulo 7.

tervenção humana. Nesse caso, é bastante comum a utilização de dados históricos.⁵ Supondo que as condições atuais e futuras sejam razoavelmente semelhantes àquelas nas quais os dados foram obtidos, podemos ter uma ideia sobre as probabilidades dos eventos de interesse através das frequências relativas dos dados históricos.

4.2.3 Axiomas e propriedades da probabilidade

Independentemente de como são obtidas, usando a definição clássica ou a experimental, as probabilidades atendem a alguns *axiomas*. Formalmente, seja um experimento aleatório e um espaço amostral Ω associado a ele. A cada evento E_i ($i = 1, 2, \dots$) associaremos um número real denominado *probabilidade de ocorrência de E_i* , $P(E_i)$, que deve satisfazer aos seguintes axiomas:

- a) $0 \leq P(E_i) \leq 1$
- b) $P(\Omega) = 1$ e
- c) Se E_1, E_2, \dots, E_n são eventos mutuamente exclusivos, então $P(E_1 \cup E_2 \cup \dots \cup E_n) = P(E_1) + P(E_2) + \dots + P(E_n)$.

O axioma (a) afirma que uma probabilidade é sempre um número entre 0 e 1. O axioma (b) afirma que, ao realizar o experimento, sempre vai ocorrer algum dos resultados possíveis, razão pela qual o espaço amostral é chamado de *evento certo*. Já o axioma (c) é menos intuitivo. Ele afirma que, ao unir eventos formados por resultados diferentes, a probabilidade de ocorrer essa união é dada pela soma das probabilidades de cada evento.

Para ilustrar os axiomas, retomemos o experimento de lançar um dado e observar o lado voltado para cima. Temos $\Omega = \{1, 2, 3, 4, 5, 6\}$. Ao realizar o experimento, certamente vai ocorrer algum elemento de Ω ; logo, $P(\Omega) = 1$. Vamos considerar, por exemplo, os eventos associados a cada resultado, isto é, $E_i = i$ ($i = 1, 2, \dots, 6$). Se for suposto que o dado é perfeitamente equilibrado e os lançamentos imparciais (dado honesto), podemos atribuir, pela definição clássica, as seguintes probabilidades: $P(E_i) = 1/6$ ($i = 1, 2, \dots, 6$). Note que esses eventos são mutuamente exclusivos ($E_i \cap E_j = \Phi, \forall i \neq j$) e a probabilidade da união de quaisquer desses eventos é dada pela soma das probabilidades de cada um. Por exemplo, pela definição clássica, Expressão 4.1, temos:

$$P(E_1 \cup E_2 \cup E_3) = P(\{1, 2, 3\}) = \frac{3}{6} = \frac{1}{2}$$

O mesmo valor pode ser obtido pelo axioma (c):

5 Alguns rios brasileiros têm dados de vazões coletados desde a década de 1930.

$$P(E_1 \cup E_2 \cup E_3) = P(E_1) + P(E_2) + P(E_3) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$$

Seguem algumas propriedades básicas da probabilidade:

1. $P(\emptyset) = 0$

Se o experimento é realizado, algum resultado certamente vai ocorrer ($P(\Omega) = 1$). Portanto, \emptyset nunca ocorre ($P(\emptyset) = 0$). \emptyset é conhecido como *evento impossível*.

2. Para o caso discreto, isto é, quando os resultados possíveis podem ser listados, então, pelo axioma (c), a probabilidade de qualquer evento pode ser obtida pela soma das probabilidades dos resultados individuais, ou seja, se $A \subseteq \Omega = \{\omega_1, \omega_2, \omega_3, \dots\}$, então:

$$P(A) = \sum_{i: \omega_i \in A} P(\omega_i) \tag{4.5}$$

No experimento do dado, por exemplo:

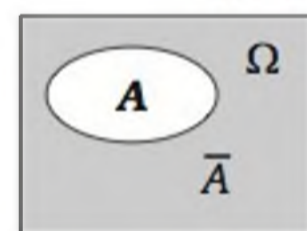
$$P(\text{número par}) = P(\{2, 4, 6\}) = P(2) + P(4) + P(6) = \frac{3}{6} = \frac{1}{2}$$

Observe que esse processo de calcular probabilidades pode ser usado mesmo quando o espaço amostral não for equiprovável.

3. Sejam $A \subseteq \Omega$ e \bar{A} o evento complementar de A , então:

$$P(\bar{A}) = 1 - P(A) \tag{4.6}$$

Note que, ao unir A e \bar{A} temos o espaço amostral Ω , que tem probabilidade igual a 1. Pelo axioma (c), temos a expressão do evento complementar.

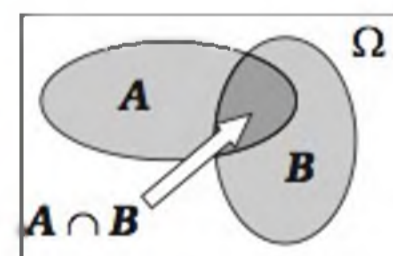


No experimento do dado, temos, por exemplo, $P(\text{ocorrer seis}) = P(\{6\}) = \frac{1}{6}$. Pela propriedade do evento complementar: $P(\text{não ocorrer seis}) = 1 - \frac{1}{6} = \frac{5}{6}$.

4. (*Regra da soma das probabilidades*). Sejam A e B eventos quaisquer, então:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{4.7}$$

Note, pelo esquema ao lado, que, ao somar $P(A)$ e $P(B)$, estamos contando duas vezes os pontos do conjunto $A \cap B$. Logo, ao calcular $P(A \cup B)$, é necessário excluir uma vez $P(A \cap B)$.



No experimento do dado, sejam: $A = \{2, 4, 6\}$ e $B = \{3, 4, 5, 6\}$. Portanto, $P(A) = 1/2$, $P(B) = 2/3$ e $P(A \cap B) = P(\{4, 6\}) = 1/3$. A probabilidade de ocorrer um número maior do que 1 pode ser calculada por $P(A \cup B) = 1/2 + 2/3 - 1/3 = 5/6$. Note que é o mesmo valor que obteríamos se calculássemos diretamente por (4.1).

EXERCÍCIOS

3. Retira-se, ao acaso, uma carta de um baralho de 52 cartas. Calcule a probabilidade de:
 - a) a carta não ser de ouros;
 - b) ser uma carta de ouros ou uma figura.
4. Depois de um longo período de testes, verificou-se que o procedimento A de recuperação de informação corre um risco de 2% de não oferecer resposta satisfatória. No procedimento B , o risco cai para 1%. O risco de ambos os procedimentos apresentarem resposta insatisfatória é de 0,5%. Qual é a probabilidade de pelo menos um dos procedimentos apresentar resposta insatisfatória?
5. De um conjunto de cinco empresas, deseja-se selecionar, aleatoriamente, uma empresa, mas com probabilidade proporcional ao número de funcionários. O número de funcionários da Empresa A é 20; de B é 15; de C é 7; de D é 5 e de E é 3.
 - a) Qual é a probabilidade de cada uma das empresas ser a selecionada?
 - b) Qual é a probabilidade de a Empresa A não ser selecionada?
6. Considere que a probabilidade de ocorrer k defeitos ortográficos em uma página de jornal é dada por:

$$p(k) = \frac{1}{e \cdot k!} \quad (e \approx 2,7183)$$

Tomando-se uma página qualquer, calcule a probabilidade de:

- a) não ocorrer erro;
 - b) ocorrer mais do que dois erros.
7. Mostre que:

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

4.3 PROBABILIDADE CONDICIONAL E INDEPENDÊNCIA

Muitas vezes, há interesse em calcular a probabilidade de ocorrência de um evento A , dada a ocorrência de um evento B . Exemplos:

- Qual é a probabilidade de chover amanhã em Florianópolis, sabendo que choveu hoje?
- Qual é a probabilidade de um dispositivo eletrônico funcionar sem problemas por 200 horas consecutivas, sabendo que ele já funcionou por 100 horas?
- Qual é a probabilidade de que um dos três servidores de correio eletrônico fique congestionado, sabendo que um deles está inoperante?

Em outras palavras, queremos calcular a probabilidade de ocorrência de A condicionada à ocorrência prévia de B . Essa probabilidade é representada por $P(A|B)$ (lê-se *probabilidade de A dado B*).

Exemplo 4.7 Os dados, a seguir, representam o sumário de um dia de observação em um posto de qualidade, em que se avalia o peso dos pacotes de leite produzidos num laticínio.

Condição do peso	Tipo do leite			
	B (B)	C (C)	UHT (U)	Total
Dentro das especificações (D)	500	4.500	1.500	6.500
Fora das especificações (F)	30	270	50	350
Total	530	4.770	1.550	6.850

Retira-se, ao acaso, um pacote de leite da população de 6.850 unidades. Sejam D e F os eventos que representam se o pacote retirado está dentro ou fora das especificações, respectivamente. Da mesma forma, B , C e U são eventos que representam o tipo do leite. Pergunta-se:

- a) Qual é a probabilidade de o pacote de leite estar fora das especificações?

Resp.: Como o espaço amostral é composto de 6.850 unidades, sendo que 350 satisfazem ao evento, então:

$$P(F) = \frac{350}{6.850} = 0,051$$

- b) Qual a probabilidade de o pacote de leite retirado estar fora das especificações, sabendo-se que é do tipo UHT?

Resp.: Nesse caso, o espaço amostral ficou restrito às 1.550 unidades de leite UHT. Destas, 50 satisfazem ao evento. Então:

$$P(F|U) = \frac{50}{1.550} = 0,032$$

Note que, se o numerador e o denominador de $P(F|U)$ forem divididos pelo número total de unidades, temos:

$$P(F|U) = \frac{50}{1.550} = \frac{50/6.850}{1.550/6.850} = \frac{P(F \cap U)}{P(U)}$$

que é a relação usada na definição formal de probabilidade condicional.

Sejam A e B eventos quaisquer, sendo $P(B) > 0$. Definimos a *probabilidade condicional de A dado B* por:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (4.8)$$

Note que no denominador temos a probabilidade do evento que supostamente aconteceu, mas calculada nas condições originais do experimento.

Se houver interesse no oposto, isto é, na probabilidade de ocorrência de B condicionada à ocorrência prévia de A , sendo $P(A) > 0$, temos:

$$P(B|A) = \frac{P(B \cap A)}{P(A)} \quad (4.9)$$

É importante ressaltar que a operação de intersecção é *comutativa*, implicando $P(A \cap B) = P(B \cap A)$.

Exemplo 4.8 Seja o lançamento de 2 dados não viciados e a observação das faces voltadas para cima. Suponha que haja interesse nas probabilidades dos seguintes eventos:

- Faces iguais, sabendo que a soma é menor ou igual a 5.
- Soma das faces menor ou igual a 5, sabendo que as faces são iguais.

Inicialmente, vamos explicitar o espaço amostral desse experimento, que é formado por todas as $6 \times 6 = 36$ possíveis combinações de resultados dos dois dados, ou seja:

$$\Omega = \left\{ \begin{array}{cccccc} (1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6) \\ (2,1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6) \\ (3,1) & (3,2) & (3,3) & (3,4) & (3,5) & (3,6) \\ (4,1) & (4,2) & (4,3) & (4,4) & (4,5) & (4,6) \\ (5,1) & (5,2) & (5,3) & (5,4) & (5,5) & (5,6) \\ (6,1) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6) \end{array} \right\}$$

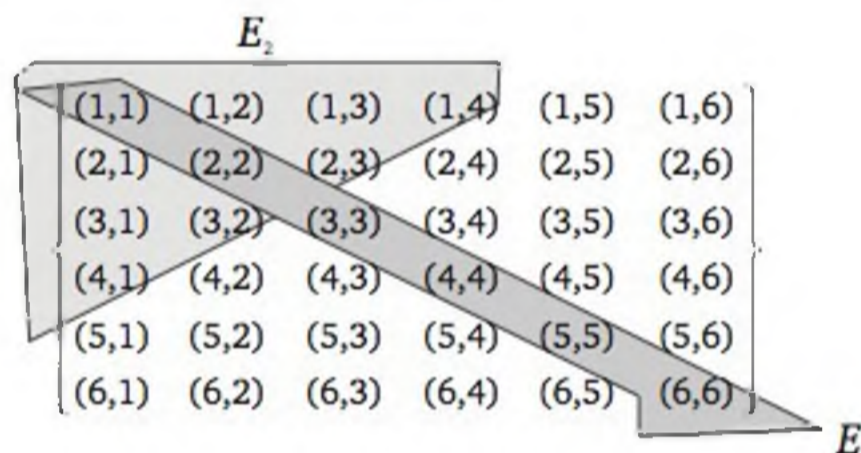
Considere os eventos:

$E_1 = \text{faces iguais} = \{(1,1), (2,2), (3,3), (4,4), (5,5), (6,6)\}$ e

$E_2 = \text{soma das faces é menor ou igual a 5} =$

$= \{(1,1), (1,2), (1,3), (1,4), (2,1), (2,2), (2,3), (3,1), (3,2), (4,1)\}$.

Portanto, $E_1 \cap E_2 = \{(1,1), (2,2)\}$. Esquematicamente:



Calculando

- a) A probabilidade de as faces serem iguais, sabendo que a soma é menor ou igual a 5. Ou seja:

$$P(E_1|E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)} = \frac{2/36}{10/36} = \frac{2}{10} = 0,2$$

Note que, se o espaço amostral for restringido ao evento conhecido, E_2 , temos 10 resultados possíveis, sendo que 2 satisfazem também ao evento de interesse, E_1 , o que torna natural a probabilidade condicional ser $2/10$.

- b) A probabilidade de a soma das faces ser menor ou igual a 5, sabendo que as faces são iguais. Ou seja:

$$P(E_2|E_1) = \frac{P(E_2 \cap E_1)}{P(E_1)} = \frac{2/36}{6/36} = \frac{2}{6} = 0,333\dots$$

4.3.1 A regra do produto

Uma das consequências da expressão da probabilidade condicional (4.8) é a *regra do produto*, obtida ao isolar a probabilidade da interseção. Ou seja:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \Rightarrow \boxed{P(A \cap B) = P(B) \cdot P(A|B)} \quad (4.10)$$

que fornece uma fórmula de calcular a probabilidade de ambos os eventos (A e B) ocorrerem. Em (4.10), o evento condicionado é B , mas o inverso também é possível, pois

$$P(B|A) = \frac{P(B \cap A)}{P(A)} \Rightarrow \boxed{P(A \cap B) = P(A) \cdot P(B|A)} \quad (4.11)$$

Para três eventos, A , B e C , a *regra do produto* pode ser escrita como

$$\boxed{P(A \cap B \cap C) = P(A) \cdot P(B|A) \cdot P(C|A \cap B)} \quad (4.12)$$

É importante que seja observada a sequência lógica dos eventos para montar as expressões precedentes.

Exemplo 4.9. Uma caixa contém 4 cartões amarelos e 8 vermelhos. Retiramos, ao acaso, 2 cartões, um após o outro, sem reposição, e observamos as cores dos dois cartões.

- a) Qual é a probabilidade de que ambos sejam amarelos?

Chamando de A_i o evento que representa cartão amarelo na i -ésima extração e V_i o evento que representa cartão vermelho na i -ésima extração ($i = 1, 2$), temos o seguinte espaço amostral:

$$\Omega = \{(A_1, A_2), (A_1, V_2), (V_1, A_2), (V_1, V_2)\}$$

A probabilidade de interesse é $P\{(A_1, A_2)\}$, que também pode ser colocada em termos de interseção: $P(A_1 \cap A_2)$, isto é, a probabilidade de ocorrer amarelo na primeira extração e amarelo na segunda extração. Para a aplicação da *regra do produto*, $P(A_1 \cap A_2) = P(A_1) \cdot P(A_2|A_1)$, calculamos:

$$P(A_1) = \frac{4}{12} = \frac{1}{3} \text{ (pois existem 4 amarelos dentre os 12 cartões) e}$$

$P(A_2|A_1) = \frac{3}{11}$ (pois, supondo que tenha sido extraído cartão amarelo na primeira extração, restaram 3 amarelos dentre 11 cartões). Logo, $P(A_1 \cap A_2) = P(A_1) \cdot P(A_2|A_1) = \frac{1}{3} \cdot \frac{3}{11} = \frac{1}{11}$.

b) Como alocar probabilidades a todos os elementos do espaço amostral?

Nesse caso, podemos construir uma árvore, indicando todas as situações possíveis (*árvore de probabilidades*). Veja a Figura 4.3.

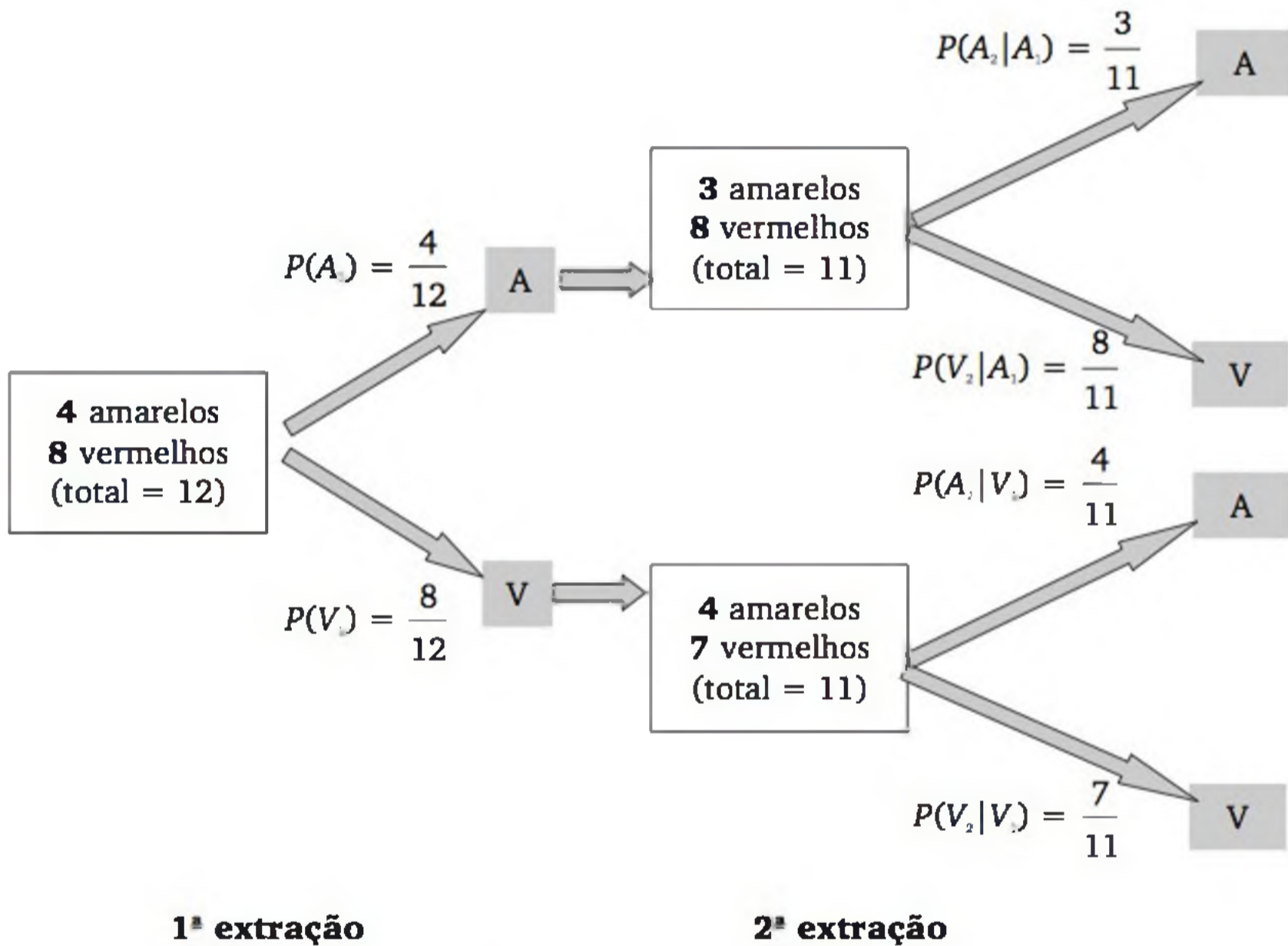


Figura 4.3 *Árvore de probabilidades – retiradas sem reposição (A = amarelo; V = vermelho).*

Com base na Figura 4.1, podemos calcular as probabilidades de todos os resultados do espaço amostral, como segue:

$$P\{(A_1, A_2)\} = P(A_1 \cap A_2) = P(A_1) \cdot P(A_2|A_1) = \frac{4}{12} \cdot \frac{3}{11} = \frac{1}{11}$$

$$P\{(A_1, V_2)\} = P(A_1 \cap V_2) = P(A_1) \cdot P(V_2|A_1) = \frac{4}{12} \cdot \frac{8}{11} = \frac{8}{33}$$

$$P\{(V_1, A_2)\} = P(V_1 \cap A_2) = P(V_1) \cdot P(A_2|V_1) = \frac{8}{12} \cdot \frac{4}{11} = \frac{8}{33}$$

$$P\{(V_1, V_2)\} = P(V_1 \cap V_2) = P(V_1) \cdot P(V_2|V_1) = \frac{8}{12} \cdot \frac{7}{11} = \frac{14}{33}$$

Observe que a soma dos quatro resultados possíveis é igual a 1 (axioma (b) da probabilidade).

- c) Qual é a probabilidade de ocorrer exatamente 1 cartão amarelo?

Queremos a probabilidade de ocorrer (A_1, V_2) ou (V_1, A_2) . Em termos da linguagem de conjuntos, queremos a união dos dois resultados (eventos). Como esses eventos são mutuamente exclusivos (não podem ocorrer simultaneamente), então a probabilidade é dada pela soma, ou seja:

$$P\{(A_1, V_2) \cup (V_1, A_2)\} = P(A_1, V_2) + P(V_1, A_2) = \frac{8}{33} + \frac{8}{33} = \frac{16}{33}$$

- d) Considere a retirada de 3 cartões. Qual é a probabilidade de que os dois primeiros sejam vermelhos e o terceiro seja amarelo?

Desejamos calcular

$$P(V_1 \cap V_2 \cap A_3) = P(V_1) \cdot P(V_2|V_1) \cdot P(A_3|V_1 \cap V_2)$$

Os dois primeiros fatores já foram calculados anteriormente. Para calcular $P(A_3|V_1 \cap V_2)$ basta considerar a caixa com 2 cartões vermelhos a menos, ou seja, com 4 amarelos e 6 vermelhos. Assim:

$$P(A_3|V_1 \cap V_2) = \frac{4}{10}$$

Logo:

$$P(V_1 \cap V_2 \cap A_3) = \frac{8}{12} \cdot \frac{7}{11} \cdot \frac{4}{10} = \frac{28}{165}$$

4.3.2 Eventos independentes

O Exemplo 4.10 é parecido com o exemplo anterior, mas com a amostragem feita *com reposição*. Verifique que os cálculos tornam-se mais simples, pois a configuração da urna não se altera na segunda extração.

Exemplo 4.10 Uma caixa contém 4 cartões amarelos e 8 vermelhos. Retiram-se, ao acaso, 2 cartões da caixa, um após o outro, sendo que o primeiro

cartão é reposto antes da retirada do segundo (*amostragem com reposição*), e observa-se a cor dos dois cartões.

A Figura 4.4 apresenta a árvore de probabilidades desse experimento.

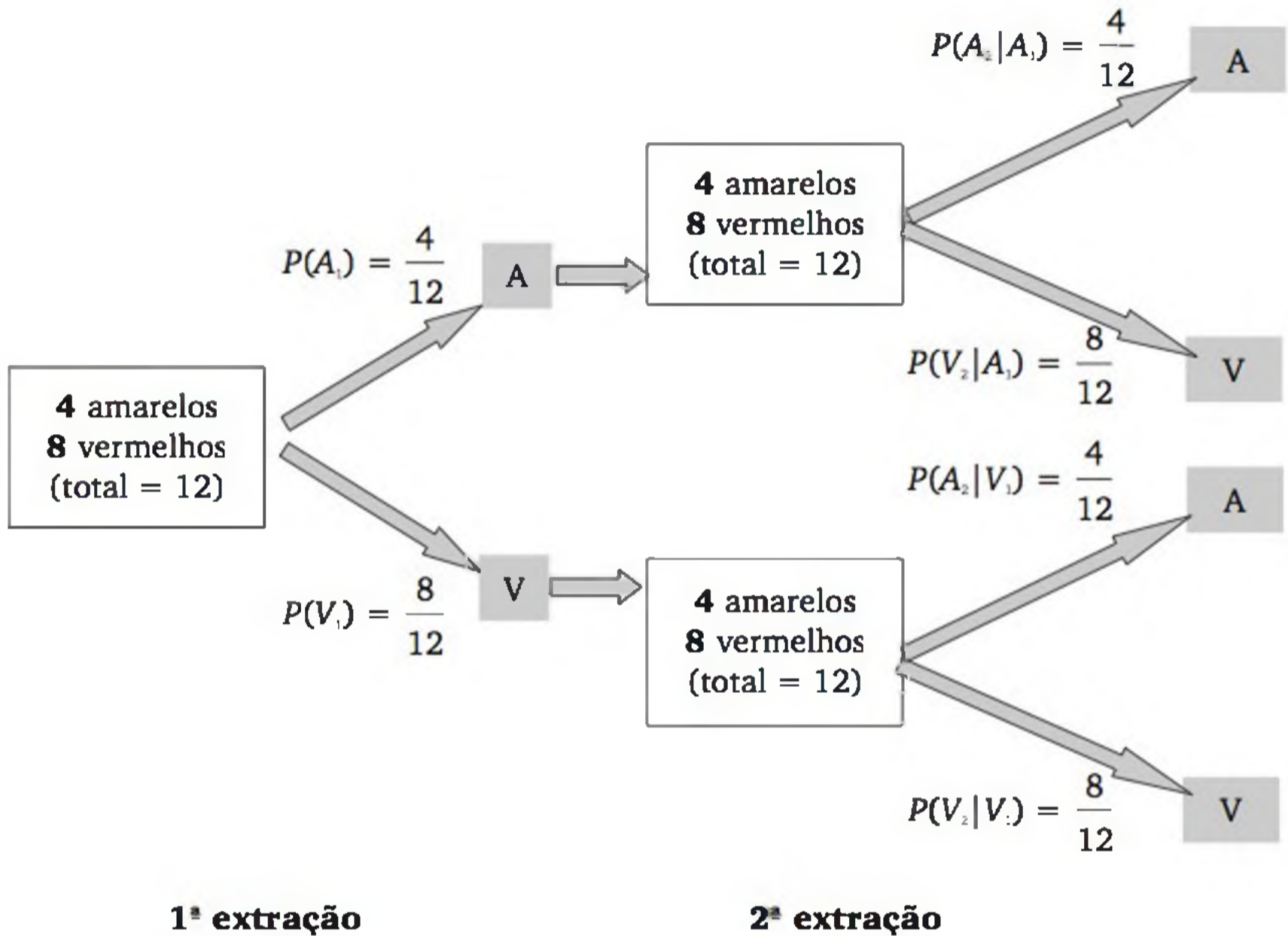


Figura 4.4 *Árvore de probabilidades – retiradas com reposição (A = amarelo; V = vermelho).*

Note que nessa situação, $P(A_2|A_1) = P(A_2|V_1) = 4/12$, ou seja, não importa se saiu cartão amarelo ou vermelho na primeira extração, a probabilidade de sair amarelo na segunda extração é de $4/12$ – há *independência* entre os eventos. Assim, basta escrever $P(A_2)$, sem condicionante.

Dois ou mais eventos são **independentes** quando a ocorrência de um dos eventos não influencia a probabilidade da ocorrência dos outros.

Se dois eventos A e B são independentes, então:

$$P(A|B) = P(A) \text{ e} \tag{4.13}$$

$$P(B|A) = P(B) \tag{4.14}$$

Como consequência de (4.14), a *regra do produto* pode ser simplificada da seguinte forma:

$$P(A \cap B) = P(A) \cdot P(B|A) = P(A) \cdot P(B) \quad (4.15)$$

Essa relação é usada para definir formalmente eventos independentes, ou seja:

A e B são independentes $\Leftrightarrow P(A \cap B) = P(A) \cdot P(B)$

A definição de independência ainda pode ser ampliada para mais eventos, como segue:

$E_1, E_2 \dots E_n$ são independentes $\Leftrightarrow P(E_1 \cap E_2 \cap \dots \cap E_n) = P(E_1) \cdot P(E_2) \dots P(E_n)$

Embora a implicação seja dos dois lados, normalmente as condições do experimento permitem verificar se é razoável supor independência entre os eventos e, em caso afirmativo, o cálculo da probabilidade da interseção pode ser fatorado nas probabilidades dos eventos independentes.

Quando a população for bastante grande em relação ao tamanho da amostra, mesmo que a amostragem seja feita *sem* reposição, podemos supor independência. Imagine que no experimento do Exemplo 4.9 haja 4.000 cartões amarelos e 8.000 vermelhos. Ao extrair dois cartões, a probabilidade de sair amarelo na segunda extração é de aproximadamente $4/12$, independentemente de ter saído amarelo ou vermelho na primeira extração.

Exemplo 4.11. Considere um sistema composto de n componentes ligados em série, de tal forma que, se um componente falhar, o sistema todo falha. Esquematicamente:



Se os componentes operam *independentemente* e cada um tem probabilidade p de falhar, qual é a probabilidade de o sistema funcionar?

Resp.: $(1 - p)^n$ (verifique o porquê; use a regra do produto para eventos independentes).

EXERCÍCIOS

8. Para testar se um sistema especialista responde satisfatoriamente a um usuário, foram feitas cinco perguntas, cada uma com quatro alternativas de resposta. Se o sistema escolhe as alternativas aleatoriamente, qual é a probabilidade de ele responder corretamente a todas as cinco perguntas?
9. Com respeito ao Exemplo 4.7, calcule:
 $P(D)$; $P(B)$; $P(D \cap B)$; $P(D|B)$; $P(B|D)$.

4.4 TEOREMA DA PROBABILIDADE TOTAL

Exemplo 4.12. Imagine que você utiliza peças de quatro fornecedores, que têm diferentes desempenhos quanto a sua qualidade. As peças são classificadas como *conformes* ou *não conformes* e você conhece a proporção de peças *não conformes* de cada fornecedor (p_1, p_2, p_3 e p_4). Considere a formação de um lote com peças dos quatro fornecedores, conforme ilustra a Figura 4.5. Se você selecionar, ao acaso, uma das peças do lote, qual é a probabilidade de ela ser *não conforme*?

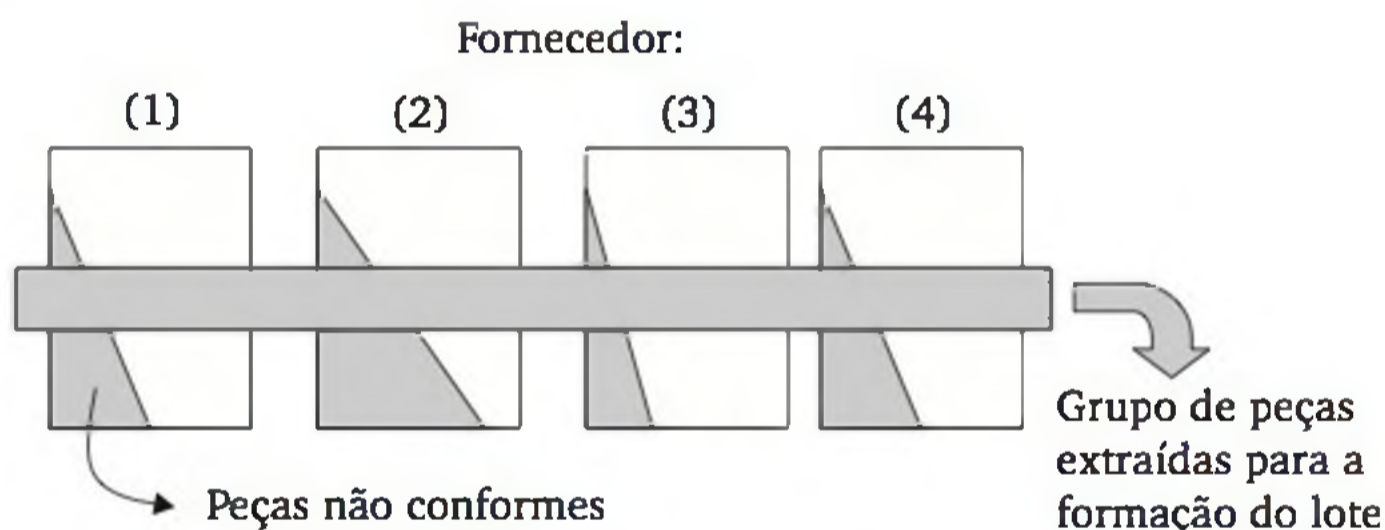


Figura 4.5 Ilustração da formação de um lote de peças provindas de quatro fornecedores.

A resposta seria simples se você soubesse de qual fornecedor é a peça selecionada, mas você não sabe. O chamado *teorema da probabilidade total* permite solucionar esse problema.

Considere o espaço amostral particionado em k eventos, E_1, E_2, \dots, E_k , satisfazendo às seguintes condições:

- $E_i \cap E_j = \emptyset$ para todo $i \neq j$ (eventos *mutuamente exclusivos*);
- $E_1 \cup E_2 \cup \dots \cup E_k = \Omega$ (eventos *exaustivos*) e
- $P(E_i) > 0$ para $i = 1, 2, \dots, k$. Veja a Figura 4.6.

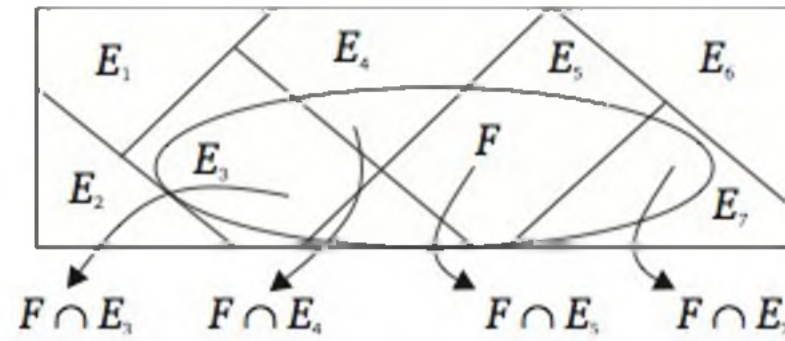


Figura 4.6 Partição do espaço amostral em eventos mutuamente exclusivos.

Seja um evento F qualquer, referente ao espaço amostral Ω . Então:

$$F = (F \cap E_1) \cup (F \cap E_2) \cup \dots \cup (F \cap E_k)$$

onde os eventos $(F \cap E_i)$ ($i = 1, 2, \dots, n$) são mutuamente exclusivos entre si. Logo:

$$\begin{aligned} P(F) &= P[(F \cap E_1) \cup (F \cap E_2) \cup \dots \cup (F \cap E_k)] = \\ &= P(F \cap E_1) + P(F \cap E_2) + \dots + P(F \cap E_k) \end{aligned}$$

Usando a *regra do produto*, temos a seguinte equação, conhecida como o *teorema da probabilidade total*:

$$P(F) = P(E_1) \cdot P(F|E_1) + P(E_2) \cdot P(F|E_2) + \dots + P(E_k) \cdot P(F|E_k)$$

ou

$$P(F) = \sum_{i=1}^k P(E_i) \cdot P(F|E_i) \tag{4.16}$$

Naturalmente, algumas $P(F|E_i)$ poderão assumir valor zero por não haver interseção entre F e E_i . O teorema da probabilidade total pode ser interpretado fisicamente como uma medida do peso de cada um dos eventos E_i , na contribuição para formar o evento F .

Exemplo 4.12 (continuação) Os eventos E_i representam as procedências das peças (fornecedores 1, 2, 3 e 4), e o evento F representa *peça não conforme*. Repare que os eventos E_i (fornecedores) são mutuamente exclusivos, pois a peça somente pode ser originária de um dos fornecedores; e que o evento F tem interseção com cada um deles (uma vez que todos os fornecedores produzem peças não conformes).

Suponha a mesma probabilidade para todos os fornecedores, isto é,

$$P(E_1) = P(E_2) = P(E_3) = P(E_4) = 0,25$$

e as probabilidades de não conformidade para cada fornecedor sejam:

$$p_1 = P(F | E_1) = 0,1; \quad p_2 = P(F | E_2) = 0,1; \quad p_3 = P(F | E_3) = 0,2; \\ p_4 = P(F | E_4) = 0,4$$

Então, usando (4.16), a probabilidade de *não conforme* é dada por:

$$P(F) = (0,25)(0,1) + (0,25)(0,1) + (0,25)(0,2) + (0,25)(0,4) = 0,20.$$

4.5 TEOREMA DE BAYES

O teorema de Bayes está intimamente relacionado ao teorema da probabilidade total. Supõem-se as mesmas condições (eventos E_i mutuamente exclusivos e exaustivos e um evento F qualquer). Basicamente, o teorema de Bayes permite obter a probabilidade de que um dos eventos E_i ocorra, sabendo-se que o evento F ocorreu. Para o caso das peças dos quatro fornecedores, o Teorema de Bayes permite responder a questões do tipo: “sabendo-se que a peça é *não conforme*, qual é a probabilidade de que tenha vindo do fornecedor 4?”

De maneira geral, usando a expressão (4.8), temos:

$$P(E_i|F) = \frac{P(E_i \cap F)}{P(F)}$$

Usando a regra do produto (4.10), podemos escrever o chamado *Teorema de Bayes*:

$$P(E_i|F) = \frac{P(E_i) \cdot P(F|E_i)}{P(F)} \quad (4.17)$$

onde $P(F)$ é calculado por (4.16).

Exemplo 4.12 (continuação) Sabendo-se que a peça é não conforme, qual é a probabilidade de que ela tenha vindo do fornecedor 4?

Lembrando que já calculamos $P(F) = 0,20$, então, aplicando (4.17), temos:

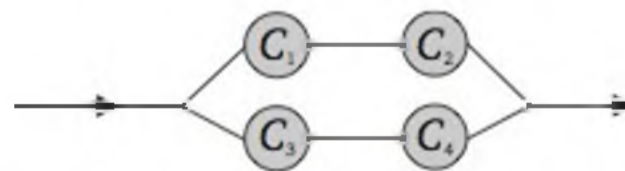
$$P(E_4|F) = \frac{P(E_4) \cdot P(F|E_4)}{P(F)} = \frac{(0,25)(0,40)}{0,20} = 0,50$$

EXERCÍCIOS

- 10.** Uma caixa contém três cartões verdes, quatro amarelos, cinco azuis e três vermelhos. Dois cartões são retirados da caixa, ao acaso, um após o outro, sem reposição. Anotam-se suas cores. Calcular a probabilidade de que:
- os dois cartões sejam da mesma cor;
 - os dois cartões sejam verdes, sabendo-se que são da mesma cor.
- 11.** Uma rede local de computadores é composta por um servidor e cinco clientes (A, B, C, D e E). Registros anteriores indicam que dos pedidos de determinado tipo de processamento, realizados através de uma consulta, cerca de 10% vêm do cliente A, 15% do B, 15% do C, 40% do D e 20% do E. Se o pedido não for feito de forma adequada, o processamento apresentará erro. Usualmente, ocorrem os seguintes percentuais de pedidos inadequados: 1% do cliente A, 2% do cliente B, 0,5% do cliente C, 2% do cliente D e 8% do cliente E.
- Qual é a probabilidade de o sistema apresentar erro?
 - Qual é a probabilidade de que o processo tenha sido pedido pelo cliente E, sabendo-se que apresentou erro?

EXERCÍCIOS COMPLEMENTARES

- 12.** A probabilidade de que Joãozinho resolva este problema é 0,5. A probabilidade de que Mariazinha resolva este problema é 0,7. Qual é a probabilidade de o problema ser resolvido se ambos tentarem independentemente?
- 13.** Um sistema tem dois componentes que operam independentemente. Suponha que as probabilidades de falha dos componentes 1 e 2 sejam 0,1 e 0,2, respectivamente. Determinar a probabilidade de o sistema funcionar nos dois casos seguintes:
- os componentes são ligados em série (isto é, ambos devem funcionar para que o sistema funcione);
 - os componentes são ligados em paralelo (isto é, basta um funcionar para que o sistema funcione).
- 14.** Um sistema tem quatro componentes que operam independentemente, sendo que cada componente tem probabilidade 0,1 de não funcionar. O sistema é ligado da seguinte forma:



Determinar a probabilidade de o sistema funcionar.

- 15.** De acordo com certa tábua de mortalidade, a probabilidade de José estar vivo daqui a 20 anos é de 0,6, e a mesma probabilidade para Manuel é de 0,9. Determinar:
- P (ambos estarem vivos daqui a 20 anos);
 - P (nenhum estar vivo daqui a 20 anos);
 - P (um estar vivo e outro estar morto daqui a 20 anos).
- 16.** Após um longo processo de seleção para preenchimento de duas vagas de emprego para engenheiro, uma empresa chegou a um conjunto de 9 engenheiros e 6 engenheiras, todos com capacitação bastante semelhante. Indeciso, o setor de recursos humanos decidiu realizar um sorteio para preencher as duas vagas oferecidas.
- construa o modelo probabilístico, considerando que se esteja observando o sexo (masculino ou feminino) dos sorteados;
 - qual é a probabilidade de que ambos os selecionados sejam do mesmo sexo?
 - sabendo-se que ambos os selecionados são do mesmo sexo, qual é a probabilidade de serem homens?
- 17.** Está sendo avaliada a qualidade de um lote de peças numa indústria cerâmica, onde estão misturados 30 pisos e 40 azulejos.
- retira-se uma peça ao acaso do lote e observa-se o tipo de cerâmica. Construa o modelo probabilístico para esta situação;
 - retiram-se duas peças ao acaso do lote, uma após a outra, com reposição, e observa-se o tipo de cerâmica. Construa o modelo probabilístico para esta situação;
 - repita o item (b), supondo que não haja reposição;
 - registros anteriores da qualidade indicaram que 1,5% dos azulejos e cerca de 0,7% dos pisos apresentaram defeitos. Retira-se, ao acaso, uma peça do lote. Qual é a probabilidade de a peça apresentar defeito?
 - para as condições do item (d), qual é a probabilidade de a peça ser piso, uma vez que apresentou defeito?
- 18.** A caixa I tem 8 peças boas e 2 defeituosas; a caixa II tem 6 peças boas e 4 defeituosas; a caixa III tem 15 peças boas e 5 defeituosas.
- tira-se, aleatoriamente, uma peça de cada caixa. Determinar a probabilidade de serem todas boas;
 - escolhe-se uma caixa ao acaso e tira-se uma peça. Determinar a probabilidade de ser defeituosa;

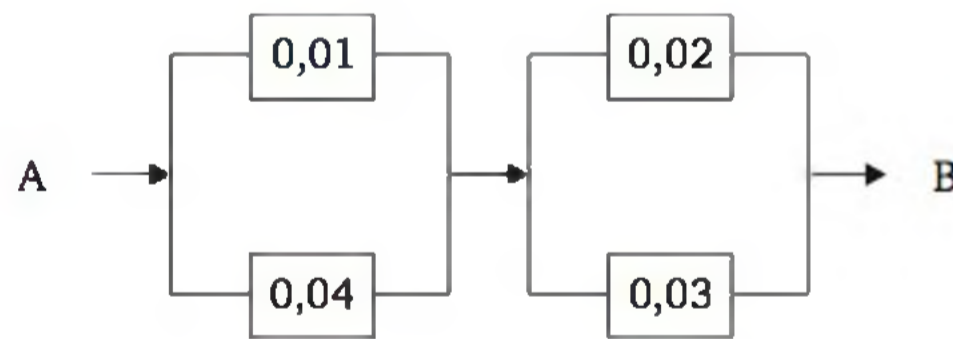
- c) escolhe-se uma caixa ao acaso e tira-se uma peça. Calcular a probabilidade de ter sido escolhida a caixa I, sabendo-se que a peça é defeituosa.

19. A qualidade de CDs foi avaliada em termos da resistência a arranhão e adequação das trilhas. Os resultados de 1.000 CDs foram:

Resistência a arranhão	Adequação das trilhas	
	Aprovado	Reprovado
Alta	700	140
Baixa	100	60

Se um CD for selecionado ao acaso desse lote de 1.000 CDs, qual é a probabilidade de ele:

- a) ter resistência a arranhão alta e ser aprovado na avaliação das trilhas?
 - b) ter resistência a arranhão alta ou ser aprovado na avaliação das trilhas?
 - c) ser aprovado na avaliação das trilhas, dado que tem resistência a arranhão alta?
 - d) ter resistência a arranhão alta, dado que foi aprovado na avaliação das trilhas?
20. Certo sistema funciona somente se houver um caminho fechado de A a B, com componentes que funcionam. Os componentes funcionam independentemente um dos outros. O sistema é esquematizado abaixo, assim como as probabilidades de falha de cada componente:



Calcule a probabilidade de o sistema funcionar.

21. Dois números inteiros são extraídos, aleatoriamente e sem reposição, do intervalo $[-20, 29]$. Esses dois números são multiplicados. Qual é a probabilidade de o produto ser positivo?

Variáveis Aleatórias Discretas

5.1 VARIÁVEL ALEATÓRIA

Um conceito de fundamental importância para a estatística indutiva é o de *variáveis aleatórias*. Para entender esse conceito, imagine que um dado comum vai ser lançado. Tente dizer qual será o número resultante. É claro que, *antes* do lançamento, não podemos dizer qual é o número que ocorrerá, pois o resultado depende do fator sorte e, por isso, é uma *variável aleatória*.

Uma **variável aleatória** pode ser entendida como variável quantitativa, cujo resultado (valor) depende de fatores aleatórios.

Outros exemplos de variáveis aleatórias são:

- a) número de coroas obtido no lançamento de duas moedas;
- b) número de itens defeituosos em uma amostra retirada, aleatoriamente, de um lote;
- c) número de defeitos em um azulejo que sai da linha de produção;
- d) número de pessoas que visitam um determinado *site*, num certo período de tempo;
- e) volume de água perdido por dia, num sistema de abastecimento;
- f) resistência ao desgaste de um tipo de aço, num teste padrão;

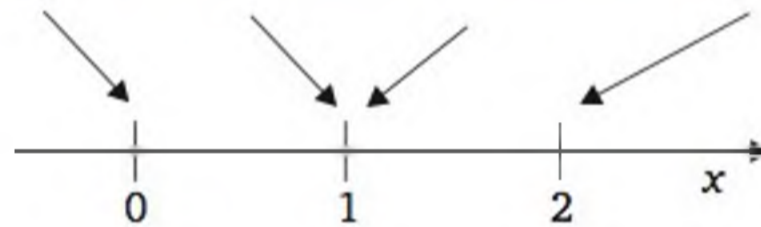
- g) tempo de resposta de um sistema computacional;
- h) grau de empeno em um azulejo que sai da linha de produção.

Todos os exemplos acima têm uma característica comum: além do resultado ser quantitativo (valor real), não podemos prevê-lo com exatidão, pois ele depende de experimento aleatório.

Embora no exemplo do dado os valores que a variável aleatória pode assumir coincidam com o espaço amostral do experimento, este não é um caso geral. No exemplo (a), *lançamento de 2 moedas*, o espaço amostral mais completo é $\Omega = \{(cara, cara), (cara, coroa), (coroa, cara), (coroa, coroa)\}$, enquanto que a variável aleatória *número de coroas* assume valores no conjunto $\{0, 1, 2\}$. Mas existe uma relação (função) entre os dois conjuntos, conforme mostra o esquema a seguir:

$$\Omega = \{(cara, cara), (cara, coroa), (coroa, cara), (coroa, coroa)\}$$

X:



Formalmente, uma **variável aleatória** é uma função que associa elementos do espaço amostral ao conjunto de números reais.

As variáveis aleatórias podem ser discretas ou contínuas, conforme mostra a Figura 5.1.

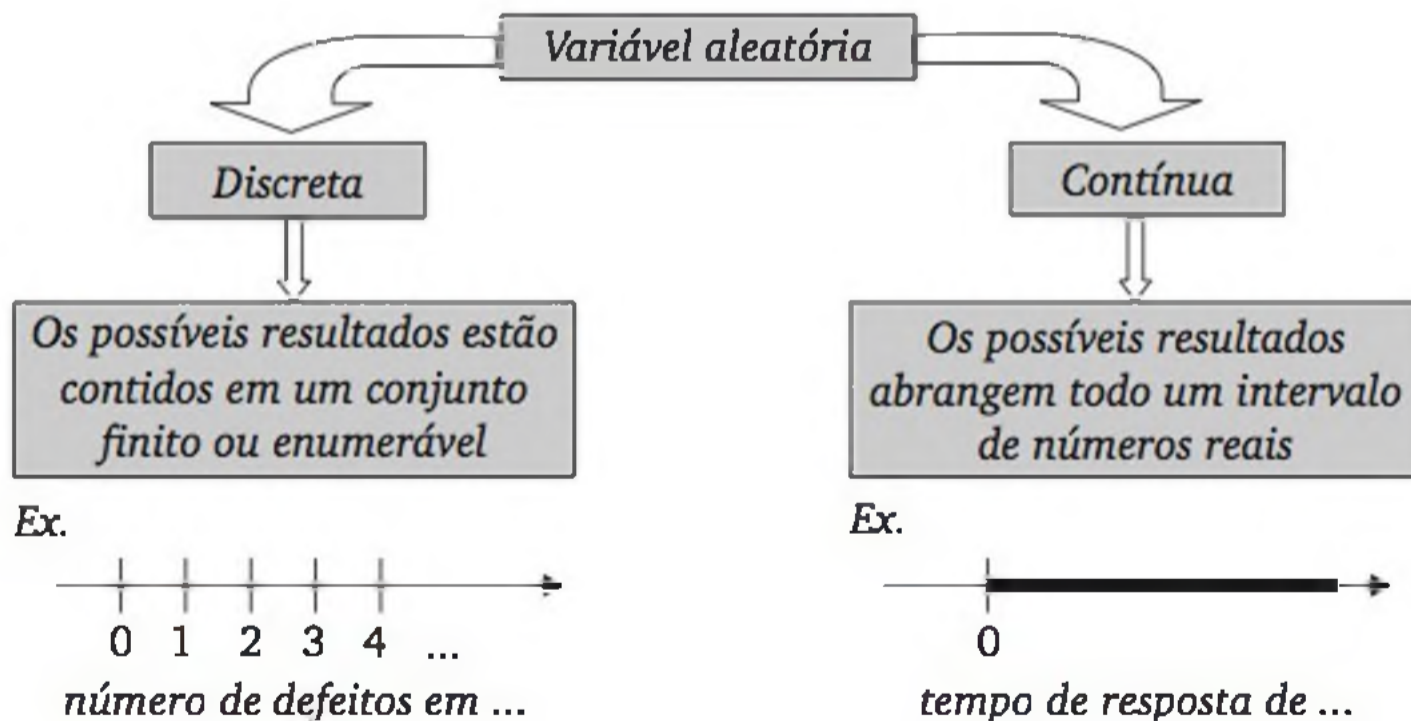


Figura 5.1 Variáveis aleatórias discretas e contínuas.

Os casos (a) – (d) são exemplos de variáveis aleatórias *discretas* e os casos (g) – (h) são exemplos de variáveis aleatórias *contínuas*. No restante deste capítulo, trataremos apenas do primeiro caso.

Cabe observar que as variáveis qualitativas também podem ser caracterizadas como variáveis aleatórias discretas, desde que as representemos como *variáveis indicadoras* 0 ou 1. Por exemplo, ao avaliar azulejos que saem de uma linha de produção, cada azulejo pode ser classificado como *bom* ($X = 0$) ou *defeituoso* ($X = 1$). Nesse caso, a variável aleatória discreta X está definida como variável indicadora de item defeituoso.¹

5.1.1 Distribuição de probabilidades

Definida uma variável aleatória discreta, temos a descrição do que *pode* ocorrer no experimento aleatório. Em alguns casos e sob certas suposições, temos duas informações:

- quais resultados podem ocorrer;
- qual é a probabilidade de cada resultado acontecer.

Exemplo 5.1 Seja a variável aleatória $X = \text{número obtido no lançamento de um dado comum}$. Se assumirmos o dado perfeitamente equilibrado e o lançamento imparcial, podemos alocar as seguintes probabilidades aos valores possíveis de X :

Valores possíveis x	Probabilidades $p(x)$
1	$1/6$
2	$1/6$
3	$1/6$
4	$1/6$
5	$1/6$
6	$1/6$
Total	1

Ou, mais resumidamente, $p(j) = 1/6$ ($j = 1, 2, 3, 4, 5, 6$).

¹ Se houver mais de duas categorias (por exemplo, A, B e C), podemos usar mais de uma variável indicadora. No caso de três categorias, podem ser empregadas as variáveis aleatórias X e Y , onde $X = 1$ se ocorrer B, e $X = 0$ caso contrário; $Y = 1$ se ocorrer C, e $Y = 0$ caso contrário. A ocorrência de A estaria representada por $X = 0$ e $Y = 0$.

A **distribuição de probabilidades** de uma variável aleatória X é a descrição do conjunto de probabilidades associadas aos possíveis valores de X , conforme foi ilustrado no exemplo precedente. Observe que a soma das probabilidades dos valores possíveis de X é igual a 1 (um).

Se X for discreta, com possíveis valores $\{x_1, x_2, \dots\}$, então a *distribuição de probabilidades* de X pode ser apresentada pela chamada **função de probabilidade**, que associa a cada valor possível x_i a sua probabilidade de ocorrência $p(x_i)$, ou seja:

$$p(x_i) = P(X = x_i) \quad (i = 1, 2, \dots)$$

Uma *função de probabilidade* deve satisfazer:

- a) $p(x_i) \geq 0$;
- b) $\sum_i p(x_i) = 1$

Existe certa similaridade entre as distribuições de probabilidades e as distribuições de frequências vistas no Capítulo 3. Contudo, na distribuição de probabilidades são mostrados os *possíveis valores* e não os valores efetivamente observados. Além disso, as probabilidades são, geralmente, alocadas a partir de suposições a respeito do experimento aleatório em questão, enquanto as frequências são obtidas com efetivas realizações do experimento.

A Figura 5.2 apresenta gráficos que podem ser usados para representar a distribuição de probabilidade de uma variável aleatória discreta. O gráfico em hastes (do lado esquerdo) é típico para variáveis aleatórias discretas. Já o gráfico em forma de histograma (do lado direito) é construído com o cuidado de a área total ser igual à unidade.

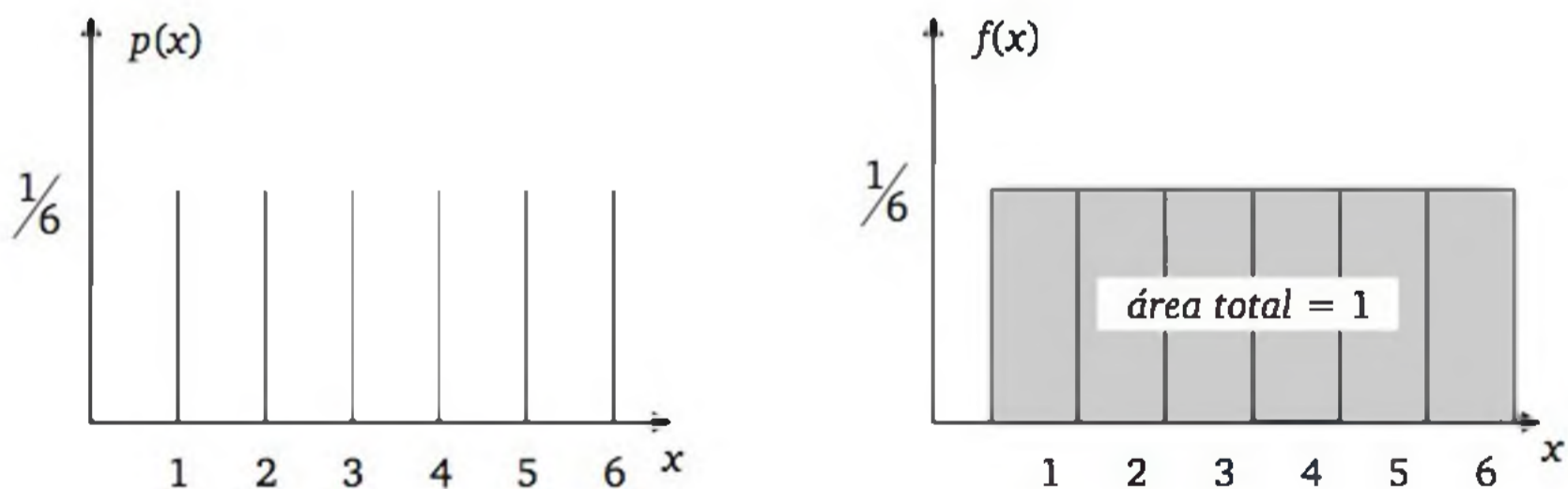


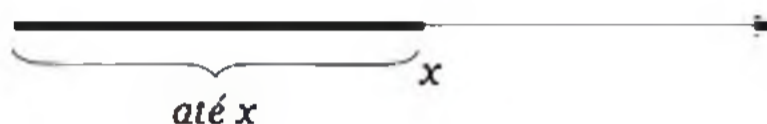
Figura 5.2 Representações gráficas da distribuição de probabilidades da variável aleatória $X =$ número obtido no lançamento de um dado comum.

5.1.2 Função de distribuição acumulada

Outra forma de representar uma distribuição de probabilidades de uma variável aleatória é através de sua função de distribuição acumulada, que é definida por:

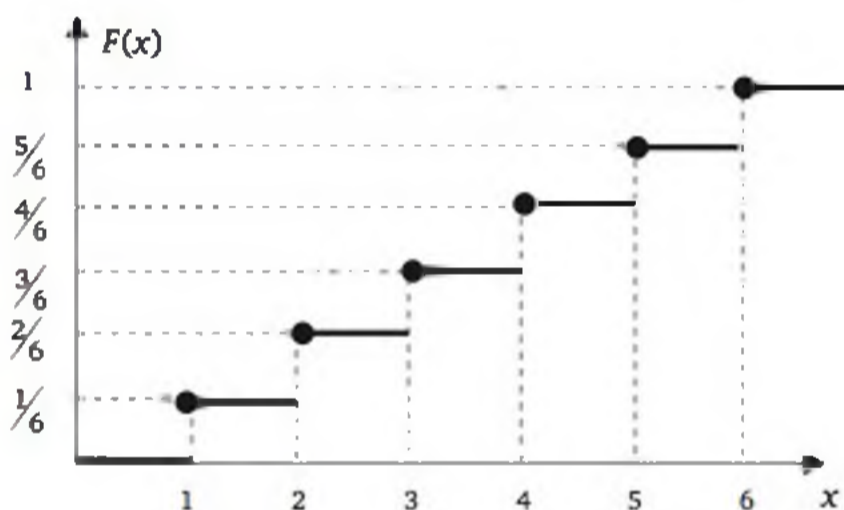
$$F(x) = P(X \leq x), \quad \forall x \in \mathfrak{R} \tag{5.1}$$

Assim, para todo $x \in \mathfrak{R}$, a função de distribuição acumulada descreve a probabilidade de ocorrer um valor até x , conforme é ilustrado a seguir:



A variável aleatória $X = \text{número obtido no lançamento de um dado comum}$ tem a seguinte função de distribuição acumulada:

$$F(x) = \begin{cases} 0 & \text{se } x < 1 \\ 1/6 & \text{se } 1 \leq x < 2 \\ 2/6 & \text{se } 2 \leq x < 3 \\ 3/6 & \text{se } 3 \leq x < 4 \\ 4/6 & \text{se } 4 \leq x < 5 \\ 5/6 & \text{se } 5 \leq x < 6 \\ 1 & \text{se } x \geq 6 \end{cases}$$



Observe que os pontos em que a função de probabilidade descreve probabilidades não nulas correspondem a saltos na função de distribuição acumulada e, também, que a altura de cada salto equivale ao valor da probabilidade naquele ponto. Assim, para todo $x \in \mathfrak{R}$, há uma relação direta entre $p(x)$ e $F(x)$.

5.1.3 Valor esperado e variância

Na análise exploratória de dados, discutimos algumas medidas (particularmente, a média, a variância e o desvio padrão – Capítulo 3) para sintetizar as informações sobre distribuições de frequências de variáveis quantitativas. De forma análoga, essas medidas também podem ser definidas para as variáveis aleatórias, com o objetivo de sintetizar características relevantes de uma distribuição de probabilidades. Considere uma variável aleatória X e sua distribuição de probabilidades:

Valores possíveis	Probabilidades
x_1	p_1
x_2	p_2
x_3	p_3
...	...
x_k	p_k
Total	1

A *média* ou *valor esperado* de uma variável aleatória X é dado por:

$$\mu = E(X) = \sum_{j=1}^k x_j p_j \tag{5.2}$$

E a *variância* por:

$$\sigma^2 = V(X) = \sum_{j=1}^k (x_j - \mu)^2 p_j \tag{5.3}$$

Alternativamente, a *variância* pode ser calculada por:²

$$V(X) = E(X^2) - \mu^2 \tag{5.4}$$

onde: $E(X^2) = \sum_{j=1}^k x_j^2 p_j$

E o *desvio padrão* é dado por:

$$\sigma = DP(X) = \sqrt{Var(X)} \tag{5.5}$$

Retomemos o exemplo da variável aleatória $X =$ *número obtido no lançamento de um dado comum*, em que a função de probabilidade é dada por:

$$p(j) = 1/6 \quad (j = 1, 2, 3, 4, 5, 6)$$

2 A demonstração da equivalência de (5.3) e (5.4) é apresentada a seguir:

$$\begin{aligned} V(X) &= \sum_{j=1}^k (x_j - \mu)^2 p_j = \sum_{j=1}^k (x_j^2 - 2x_j \mu + \mu^2) p_j = \sum_{j=1}^k x_j^2 p_j - 2\mu \sum_{j=1}^k x_j p_j + \sum_{j=1}^k p_j \mu^2 = \\ &= \sum_{j=1}^k x_j^2 p_j - 2\mu \mu + \mu^2 \sum_{j=1}^k p_j = \sum_{j=1}^k x_j^2 p_j - 2\mu^2 + \mu^2 = \sum_{j=1}^k x_j^2 p_j - \mu^2 = E(X^2) - \mu^2. \end{aligned}$$

e calculemos o valor esperado e a variância.

$$E(X) = 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) + 6\left(\frac{1}{6}\right) = 3,5$$

$$E(X^2) = 1^2\left(\frac{1}{6}\right) + 2^2\left(\frac{1}{6}\right) + 3^2\left(\frac{1}{6}\right) + 4^2\left(\frac{1}{6}\right) + 5^2\left(\frac{1}{6}\right) + 6^2\left(\frac{1}{6}\right) = 15,167$$

$$\text{Assim, } V(X) = E(X^2) - \mu^2 = 15,167 - (3,5)^2 = 2,92.$$

Considerando que as probabilidades podem ser interpretadas como limite da frequência relativa quando o experimento é executado muitas e muitas vezes, podemos interpretar o *valor esperado* como a média aritmética dos resultados da variável aleatória se o experimento pudesse ser repetido infinitas vezes. Assim, se pudéssemos lançar o dado infinitas vezes, obteríamos, em média, 3,5 pontos por lançamento. Já a variância informa sobre a dispersão dos possíveis valores.

Observe que, no presente exemplo, o valor esperado $\mu = 3,5$ é um número que a variável aleatória *não* pode assumir. Fisicamente, o *valor esperado* corresponde ao *centro de gravidade* da distribuição de probabilidades.

Algumas propriedades

Sendo c uma constante e X e Y variáveis aleatórias, as seguintes relações podem ser comprovadas:

a) $E(c) = c$	f) $V(c) = 0$
b) $E(X + c) = E(X) + c$	g) $V(X + c) = V(X)$
c) $E(cX) = cE(X)$	h) $V(cX) = c^2V(X)$
d) $E(X + Y) = E(X) + E(Y)$	i) $DP(cX) = c DP(X)$
e) $E(X - Y) = E(X) - E(Y)$	

As relações (b) e (g) mostram que ao somar uma constante a uma variável aleatória, a distribuição de probabilidades é deslocada por esta constante, mas a variabilidade é preservada. Todavia, ao multiplicar a variável aleatória por uma constante – relações (c) e (h) –, o centro da distribuição é deslocado na mesma proporção e a variabilidade também é alterada (ver Figura 5.3).

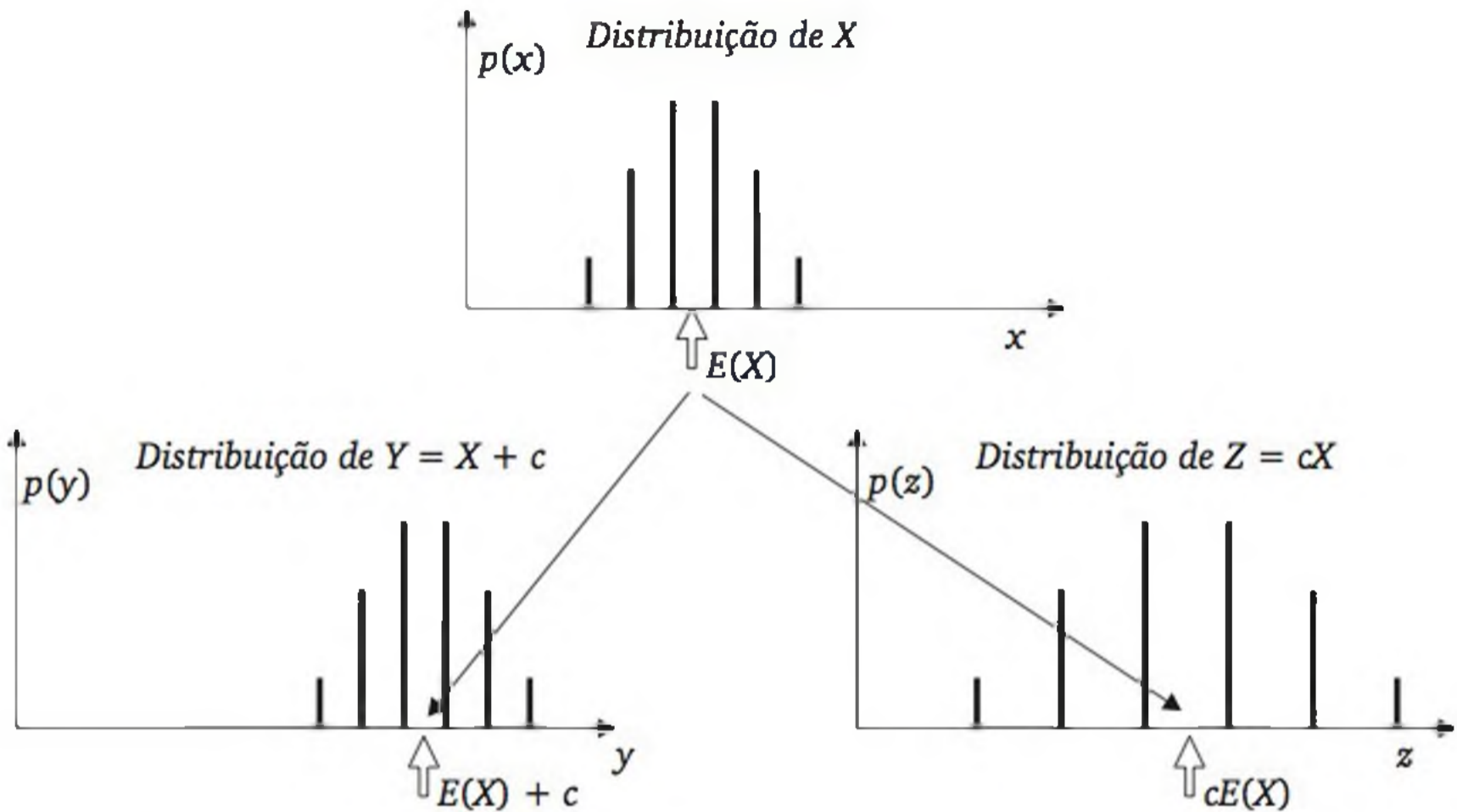


Figura 5.3 Efeito da soma e da multiplicação de uma constante sobre uma variável aleatória.

5.1.4 Variáveis aleatórias independentes

Considere que um dado seja lançado duas vezes, sob as mesmas condições. Seja X_i a variável aleatória que representa o número de pontos obtido no i -ésimo lançamento ($i = 1, 2$). Supondo que o dado seja perfeitamente equilibrado e o lançamento imparcial, então X_1 e X_2 têm a mesma função de probabilidade, que é dada por:

$$p(j) = 1/6 \quad (j = 1, 2, 3, 4, 5, 6)$$

Considere, também, a variável aleatória S como o número total de pontos obtidos nos dois lançamentos, isto é,

$$S = X_1 + X_2$$

A função de probabilidade de S pode ser obtida com base em funções de probabilidades de X_1 e X_2 , resultando em:

s	2	3	4	5	6	7	8	9	10	11	12
$p(s)$	$1/36$	$2/36$	$3/36$	$4/36$	$5/36$	$6/36$	$5/36$	$4/36$	$3/36$	$2/36$	$1/36$

Note que, se o primeiro lançamento for realizado e ocorrer o resultado x_1 , sendo $x_1 \in \{1, 2, 3, 4, 5, 6\}$, a função de probabilidade de X_2 permanece a mesma, pois o segundo lançamento independe do primeiro. Assim, podemos dizer que X_1 e X_2 são *variáveis aleatórias independentes*.

No entanto, a função de probabilidade de S é alterada. Por exemplo, se ocorrer o valor 1 ($X_1 = 1$), então a distribuição de S passa a ser:

Possíveis valores de S dado que $X_1 = 1$	2	3	4	5	6	7
Probabilidades	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Já se ocorrer $X_1 = 6$, passamos a ter:

Possíveis valores de S dado que $X_1 = 6$	7	8	9	10	11	12
Probabilidades	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Ou seja, X_1 e S *não são variáveis aleatórias independentes*.

Em geral, X_1, X_2, \dots, X_n podem ser consideradas **variáveis aleatórias independentes** se o conhecimento de uma não altera as distribuições de probabilidades das demais.

Muitas vezes as características do experimento permitem-nos avaliar se as variáveis aleatórias envolvidas são independentes. Essa condição é importante, pois a maioria dos métodos estatísticos é desenvolvida na suposição de que as observações provêm de variáveis aleatórias independentes.

Considerando a regra do produto, podemos calcular a probabilidade de *sair número par nos dois lançamentos do dado* simplesmente multiplicando a probabilidade de *sair par em cada lançamento*, ou seja, $(\frac{1}{2}) \cdot (\frac{1}{2}) = \frac{1}{4}$. Podemos representar o evento *sair número par nos dois lançamentos do dado* por:

$$\{X_1 \in (2, 4, 6), X_2 \in (2, 4, 6)\}$$

E a probabilidade deste evento por:

$$\begin{aligned} P\{X_1 \in (2, 4, 6), X_2 \in (2, 4, 6)\} &= P\{X_1 \in (2, 4, 6)\} \cdot P\{X_2 \in (2, 4, 6)\} = \\ &= \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}. \end{aligned}$$

Para um conjunto de variáveis aleatórias X_1, X_2, \dots, X_n , definimos:

X_1, X_2, \dots, X_n são variáveis aleatórias independentes se e só se:

$$P\{X_1 \in E_1, X_2 \in E_2, \dots, X_n \in E_n\} = P(X_1 \in E_1) \cdot P(X_2 \in E_2) \cdot P(X_n \in E_n)$$

para quaisquer conjuntos E_1, E_2, \dots, E_n .

Se X e Y são variáveis aleatórias independentes, então:

$$V(X + Y) = V(X) + V(Y) \quad (5.6)$$

$$V(X - Y) = V(X) + V(Y) \quad (5.7)$$

Note, por (5.7), que a variância da diferença de duas variáveis aleatórias independentes é a soma das variâncias de cada variável aleatória. Isso decorre da propriedade (h) (Seção 5.1.3).

EXERCÍCIOS

1. Apresente a função de probabilidade para as seguintes variáveis aleatórias:
 - a) Número de caras obtido com o lançamento de uma moeda honesta;
 - b) Número de caras obtido no lançamento de duas moedas honestas;
 - c) Número de peças com defeito em uma amostra de duas peças, sorteadas aleatoriamente de um grande lote, em que 40% das peças são defeituosas;
 - d) Número de peças com defeito em uma amostra de três peças, sorteadas aleatoriamente de um grande lote, em que 40% das peças são defeituosas.
2. Apresente, sob forma gráfica, a distribuição de probabilidades do Exercício 1(d).
3. Apresente a função de probabilidade acumulada do Exercício 1(d).
4. Calcule os valores esperados e as variâncias das distribuições de probabilidade do Exercício 1.
5. Considere que um produto pode estar perfeito (B), com defeito leve (DL) ou com defeito grave (DG). Seja a seguinte distribuição do lucro (em R\$), por unidade vendida desse produto:

Produto	x	$p(x)$
B	6	0,7
DL	0	0,2
DG	-2	0,1

- a) Calcule o valor esperado e a variância do lucro.
 - b) Se, com a redução de desperdícios, foi possível aumentar uma unidade no lucro de cada unidade do produto, qual é o novo valor esperado e a variância do lucro por unidade?
 - c) E se o lucro duplicou, qual é o novo valor esperado e variância do lucro por unidade?
6. Certo tipo de conserva tem peso líquido médio de 900 g, com desvio padrão de 10 g. A embalagem tem peso médio de 100 g, com desvio padrão de 4 g. Suponha que o processo de enchimento das embalagens controla o peso líquido, de tal forma que se possa supor independência entre o peso líquido e o peso da embalagem. Qual é a média e o desvio padrão do peso bruto?

5.2 PRINCIPAIS DISTRIBUIÇÕES DISCRETAS

Na seção anterior, construímos as distribuições de probabilidades de algumas variáveis aleatórias, empregando nosso conhecimento para o cálculo das probabilidades envolvidas. Nesta seção, estudaremos alguns *modelos probabilísticos padrões*, que podem ser usados em diversas situações práticas. O problema passa a ser, então, determinar *qual modelo* é o mais adequado para a situação em estudo e como aplicá-lo adequadamente.

Lembremos que, para identificarmos uma variável aleatória discreta, temos de conhecer *quais resultados* podem ocorrer e *quais são as probabilidades* associadas aos resultados. A seguir, vamos ver os principais modelos discretos.

5.2.1 Distribuição de Bernoulli

Talvez os experimentos mais simples são aqueles em que observamos a presença ou não de alguma característica, que são conhecidos como *ensaios de Bernoulli*. Alguns exemplos:

- a) lançar uma moeda e observar se ocorre *cara* ou não;
- b) lançar um dado e observar se ocorre *seis* ou não;
- c) numa linha de produção, observar se um item, tomado ao acaso, é ou não *defeituoso*;
- d) verificar se um servidor de uma *intranet* está ou não ativo.

Denominamos *sucesso* e *fracasso* os dois eventos possíveis em cada caso.³ O ensaio de Bernoulli é caracterizado por uma variável aleatória X , definida por $X = 1$, se sucesso; $X = 0$, se fracasso. A função de probabilidade de X (*Distribuição de Bernoulli*) é dada por

x	$p(x)$
0	$1 - p$
1	p
Total	1

onde $p = P\{\text{sucesso}\}$. A distribuição fica completamente especificada ao atribuímos um valor para p . No exemplo (a), se o lançamento for imparcial e a moeda perfeitamente equilibrada, $p = 1/2$. Em (b), com suposição análoga, $p = 1/6$. Outras características da distribuição de Bernoulli:

$$E(X) = p \quad (5.8)$$

$$V(X) = p \cdot (1 - p) \quad (5.9)$$

$$F(x) = \begin{cases} 0 & \text{se } x < 0 \\ 1 - p & \text{se } 0 \leq x < 1 \\ 1 & \text{se } x \geq 1 \end{cases} \quad (5.10)$$

5.2.2 Distribuição Binomial

Na maior parte das vezes, são realizados n ensaios de Bernoulli. O interesse está no número X de ocorrências de *sucesso*, como nos exemplos a seguir:

- lançar uma moeda cinco vezes e observar o número de *caras*;
- numa linha de produção, observar dez itens, tomados ao acaso, e verificar quantos estão *defeituosos*;
- verificar, num dado instante, o número de processadores ativos, num sistema com multiprocessadores;
- verificar o número de *bits* que não estão afetados por ruídos, em um pacote com n *bits*.

Nos exemplos precedentes, se for possível supor:

- ensaios *independentes*;
- $P\{\text{sucesso}\} = p$, constante para todo ensaio ($0 < p < 1$).

³ No presente contexto, o termo *sucesso* não significa algo **bom**, mas simplesmente um resultado ou evento no qual temos interesse; e *fracasso*, o outro resultado ou evento possível.

Temos, então, exemplos de *experimentos binomiais*.

Uma variável aleatória com distribuição binomial de parâmetros n e p pode ser apresentada por:

$$X = X_1 + X_2 + \dots + X_n \tag{5.11}$$

onde X_1, X_2, \dots, X_n são variáveis aleatórias independentes, sendo cada uma delas com distribuição de Bernoulli de parâmetro p . Como X_i será 0 ou 1, dependendo da ocorrência ou não de sucesso no i -ésimo ensaio ($i = 1, 2, \dots, n$), então a soma X corresponderá ao *número de sucessos*. Para especificarmos a função de probabilidade de X , considere o seguinte exemplo:

Exemplo 5.1 Uma indústria processadora de suco classifica os carregamentos de laranja que chegam a suas instalações em A, B ou C . Suponha independência entre as chegadas dos carregamentos, isto é, a classificação de um não altera a classificação dos demais. Suponha também que a probabilidade p de classificação na classe A é a mesma para todos os carregamentos. Para os próximos 4 carregamentos, seja X a variável aleatória que representa o *número de carregamentos classificados na classe A*. Vamos calcular a probabilidade de que X assumo o valor x , isto é, a probabilidade de que x carregamentos sejam classificados na classe A ($x = 0, 1, 2, 3, 4$).

Para cada carregamento, seja S (sucesso) quando este for classificado na classe A ; e seja F (fracasso) quando este for classificado em outra classe. A Figura 5.4 mostra todas as possíveis sequências de resultados, os possíveis valores de X e as correspondentes probabilidades.

Resultados possíveis de 4 carregamentos:

				SSFF		
				SFSF		
		SFFF		SFFS	SSSF	
		FSFF		FSSF	SSFS	
		FFSF		FSFS	SFSS	
	FFFF	FFFS		FFSS	FSSS	SSSS
Valores de X :	0	1	2	3	4	
	↓	↓	↓	↓	↓	
Probabilidades:	$(1 - p)^4$	$4p(1 - p)^3$	$6p^2(1 - p)^2$	$4p^3(1 - p)$	p^4	

Figura 5.4 Construção de uma distribuição binomial com $n = 4$ e p genérico.

Explicando as probabilidades da Figura 5.4. O evento $X = 0$ ocorre quando nenhum carregamento é classificado na classe A ($FFFF$), cuja probabilidade é

$(1 - p)(1 - p)(1 - p)(1 - p) = (1 - p)^4$. O evento $X = 1$ ocorre quando um carregamento for classificado na classe A (*SFFF* ou *FSFF* ou *FFSF* ou *FFFS*). Como cada um desses resultados tem probabilidade $p(1 - p)^3$, a probabilidade do evento $X = 1$ é igual a $4p(1 - p)^3$. As outras probabilidades podem ser obtidas de forma análoga.

Coefficientes binomiais

Na Figura 5.4, podemos observar que no cálculo da probabilidade do evento $X = 1$, contamos de quantas maneiras poderia aparecer um sucesso entre as quatro possibilidades, assim encontramos a quantidade quatro, correspondente às seguintes sequências de respostas: *SFFF*, *FSFF*, *FFSF* e *FFFS*.

Em geral, na distribuição binomial, para calcular a probabilidade do evento $X = x$, onde x é um valor possível da variável aleatória X , precisamos conhecer o número de maneiras em que podemos combinar os x sucessos entre os n ensaios. Esse valor, conhecido como coeficiente binomial, entra no cálculo da probabilidade como um coeficiente das potências de p e $1 - p$, como verificamos na Figura 5.4.

Vamos representar o número de combinações que podemos fazer com x elementos, numa sequência de n elementos (sendo $x \leq n$), por $\binom{n}{x}$. Esse número de combinações pode ser calculado pela seguinte expressão:

$$\boxed{\binom{n}{x} = \frac{n!}{(n - x)! x!}} \tag{5.12}$$

onde $n! = n(n - 1)(n - 2)...1$ (lê-se *n fatorial*) e, por convenção, $0! = 1$. Por exemplo, para $n = 4$ temos os seguintes coeficientes binomiais:

$$\begin{aligned} x = 0: \binom{4}{0} &= \frac{4!}{4! \cdot 0!} = \frac{4!}{4!} = 1 & x = 3: \binom{4}{3} &= \frac{4!}{1! \cdot 3!} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{1 \cdot 3 \cdot 2 \cdot 1} = 4 \\ x = 1: \binom{4}{1} &= \frac{4!}{3! \cdot 1!} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1 \cdot 1} = 4 & x = 4: \binom{4}{4} &= \frac{4!}{0! \cdot 4!} = \frac{4!}{4!} = 1 \\ x = 2: \binom{4}{2} &= \frac{4!}{2! \cdot 2!} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 2 \cdot 1} = 6 \end{aligned}$$

Expressão geral da distribuição binomial

Seja X uma variável aleatória com distribuição binomial de parâmetros n e p (sendo $0 < p < 1$). A probabilidade de X assumir um certo valor x , pertencente ao conjunto $\{0, 1, 2, \dots, n\}$, é dada pela expressão:

$$p(x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x} \tag{5.13}$$

Exemplo 5.2 (continuação). Historicamente, 30% dos carregamentos são classificados na classe A, em que podemos supor que a probabilidade p de um carregamento ser classificado na classe A é 0,3. Entre os quatro próximos carregamentos, calculemos a probabilidade de exatamente dois serem classificados na classe A.

Temos $n = 4$ e $p = 0,3$. Assim, a probabilidade de x carregamentos serem classificados na classe A é dada por:

$$p(x) = \binom{4}{x} \cdot (0,3)^x \cdot (0,7)^{4-x} \quad (x = 0, 1, 2, 3, 4)$$

Em particular, $p(2) = \binom{4}{2} \cdot (0,3)^2 \cdot (0,7)^{4-2} = 6 \cdot (0,09) \cdot (0,49) = 0,2646$

A Tabela 1 do apêndice apresenta probabilidades da binomial para $n \leq 15$ e p múltiplo de 0,05. O Exemplo 5.3 ilustra o uso da tabela.

Exemplo 5.2 Dados históricos mostram que 70% das pessoas que acessam a página p23 da *internet* também acessam a página p24. Obteremos, através da tabela da distribuição binomial, a probabilidade de que, nos dez próximos acessos à p23, a maioria também acesse a p24.

Note que temos um experimento binomial, com $n = 10$ e $p = 0,7$ (supondo independência entre os acessos). Usando a *Tabela da distribuição binomial*, podemos especificar a distribuição de $X = \text{número de pessoas que também acessam a p24}$. A probabilidade de ocorrer o evento *a maioria também acessar a p24* corresponde, em termos da variável aleatória X , ao evento $\{X > 5\}$, como ilustramos ao lado. A probabilidade deste evento será a soma dos resultados individuais, ou seja:

$$\begin{aligned} P(X > 5) &= \\ &= p(6) + p(7) + p(8) + p(9) + p(10) = \\ &= 0,2001 + 0,2668 + 0,2335 + 0,1211 + 0,0282 = \\ &= 0,8497. \end{aligned}$$

Parte da Tabela 1

		p
n	x	0,70
10	0	0,0000
	1	0,0001
	2	0,0014
	3	0,0090
	4	0,0368
	5	0,1029
	6	0,2001
	7	0,2668
	8	0,2335
	9	0,1211
	10	0,0282

$x > 5$

Se X tem distribuição binomial de parâmetros n e p , então seu *valor esperado* e sua *variância* podem ser calculados por:

$$E(X) = n.p \quad (5.14)$$

$$V(X) = n.p.(1 - p) \quad (5.15)$$

A Figura 5.5 mostra duas distribuições binomiais com a indicação da posição dos respectivos valores esperados.

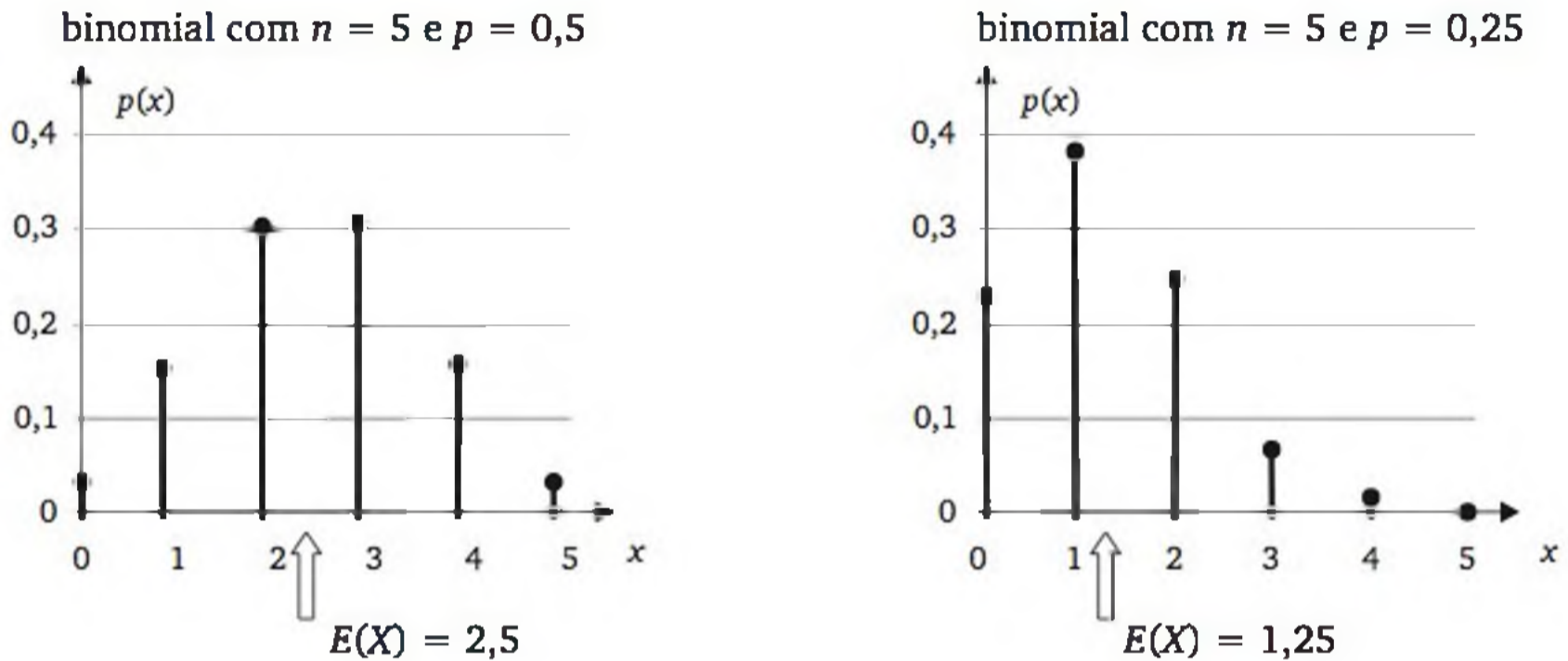


Figura 5.5 Representações gráficas de distribuições binomiais.

Distribuições binomiais com $p = 0,5$ são simétricas, mas são assimétricas quando $p \neq 0,5$. A assimetria aumenta à medida que p aproxima-se de zero (assimetria positiva) ou de um (assimetria negativa).

EXERCÍCIOS

7. Dados históricos mostram que 5% dos itens provindos de um fornecedor apresentam algum tipo de defeito. Considerando um lote com 20 itens, calcular a probabilidade de:
 - a) haver algum item com defeito;
 - b) haver exatamente dois itens defeituosos;
 - c) haver mais de dois itens defeituosos;
 - d) qual é o número esperado de itens defeituosos no lote?
 - e) e de itens bons?
8. Apresente o gráfico da variável aleatória do Exemplo 5.2 sob a forma de um histograma, indique a posição do valor esperado e represente $P(X > 5)$ como uma área na figura.

5.2.3 Distribuição hipergeométrica

Considere o problema básico de inspeção por amostragem, em que observamos uma amostra de n itens de um lote com N itens, sendo r defeituosos. Avaliamos o número X de itens defeituosos na amostra. A variável aleatória X aparenta ser binomial, mas só é realmente binomial se:

- a seleção da amostra for *aleatória* (para garantir a mesma probabilidade p de sair item defeituoso em todos os ensaios);
- *com reposição* (para garantir *independência* entre os ensaios).

A segunda condição não costuma ser satisfeita na prática. Se a amostragem for *aleatória*, mas *sem reposição*, a distribuição de X é conhecida como *hipergeométrica* de parâmetros N , n e r (ver Figura 5.6).

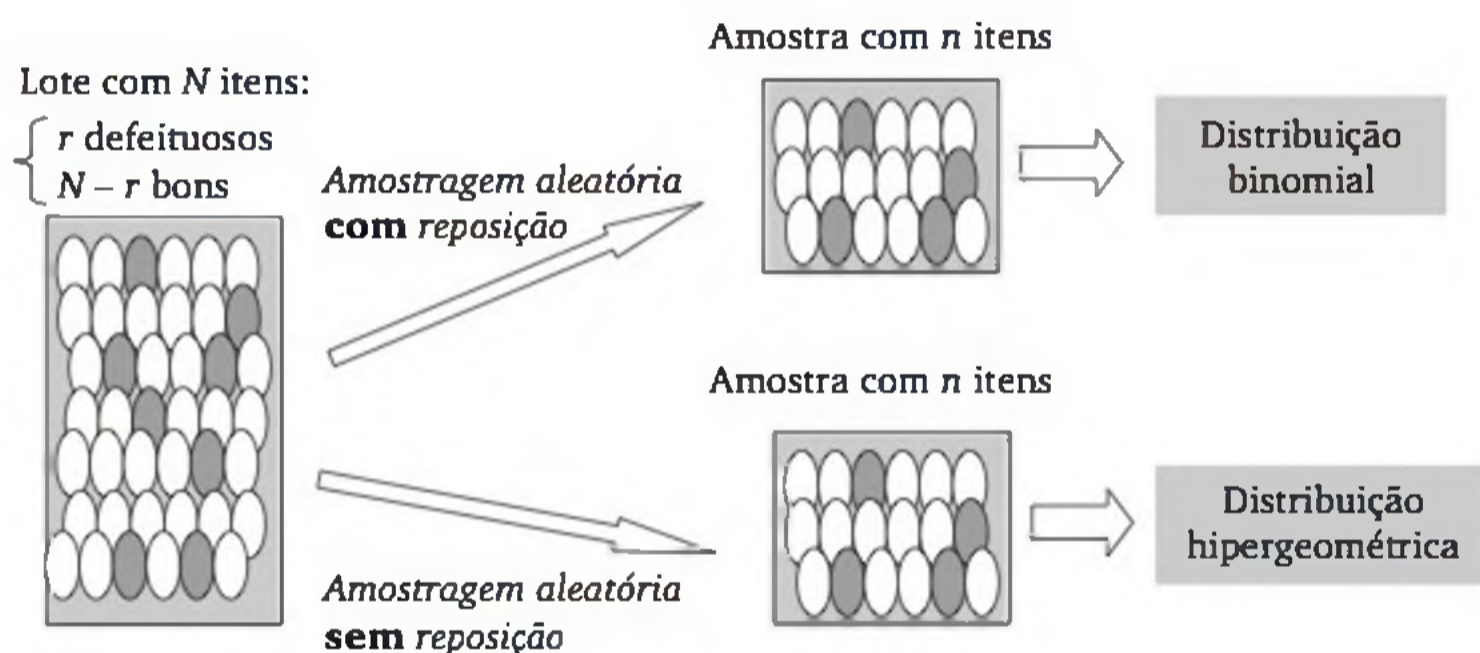


Figura 5.6 Caso típico de inspeção por amostragem. Variável aleatória em estudo: $X =$ número de defeituosos na amostra.

A função de probabilidade de X é expressa por:

$$p(x) = \frac{\binom{r}{x} \cdot \binom{N-r}{n-x}}{\binom{N}{n}} \quad [x = 0, 1, \dots, \min(r, n)] \quad (5.16)$$

com valor esperado e variância dados por:

$$E(X) = n.p \quad (5.17)$$

$$V(X) = n.p.(1 - p) \cdot \frac{N - n}{N - 1} \quad (5.18)$$

onde $p = \frac{r}{N}$

Exemplo 5.3 Placas de vídeo são expedidas em lotes de 30 unidades. Antes que a remessa seja aprovada, um inspetor escolhe aleatoriamente cinco placas do lote e as inspeciona. Se nenhuma das placas inspecionadas for defeituosa, o lote é aprovado. Se uma ou mais forem defeituosas, todo o lote é inspecionado. Supondo que haja três placas defeituosas no lote, qual é a probabilidade de que o controle da qualidade aponte para a inspeção total?

Seja X o número de placas defeituosas na amostra. Desejamos calcular:

$$P(X \geq 1) = 1 - P(X = 0)$$

Usando o modelo hipergeométrico:

$$p(0) = \frac{\binom{3}{0} \cdot \binom{30-3}{5-0}}{\binom{30}{5}} = \frac{\binom{27}{5}}{\binom{30}{5}} = \frac{80.730}{142.506} = 0,5665$$

Logo, $P(X \geq 1) = 1 - 0,5665 = 0,4335$.

É importante ressaltar que quando N é muito maior do que n , a distribuição hipergeométrica pode ser aproximada pela binomial. Muitos autores prescrevem uma relação $n/N \leq 0,05$ para que seja possível fazer a aproximação.⁴ Nesse caso, a binomial tem parâmetros $n =$ tamanho da amostra e $p = r/N$.

EXERCÍCIOS

9. Qual é a probabilidade do Exemplo 5.3, se a inspeção completa for feita somente quando forem encontradas mais do que uma placa defeituosa na amostra?
10. Calcule o valor esperado e a variância da variável aleatória definida no Exemplo 5.3.

5.2.4 Distribuição de Poisson

Considere as situações em que se avalia o número de ocorrências de um tipo de evento por unidade de tempo, de comprimento, de área, ou de volume. Por exemplo:

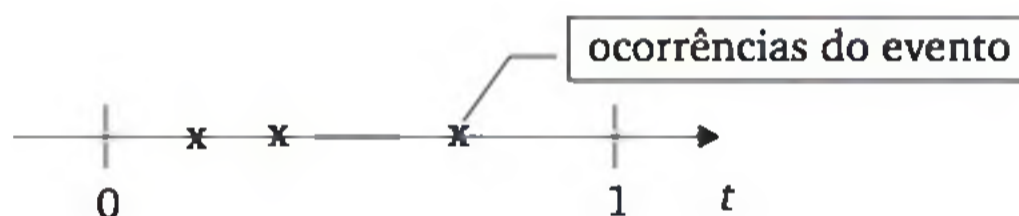
⁴ Observe que se N for muito maior que n , as retiradas, mesmo feitas *sem* reposição, não irão modificar em demasia as probabilidades condicionais de ocorrências de *sucessos* (e de *fracassos*), na sequência de ensaios.

- número de consultas a uma base de dados em um minuto;
- número de pedidos a um servidor num intervalo de tempo;
- número de erros de tipografia em um formulário;
- número de defeitos em um m^2 de piso cerâmico;
- número de pulsações radioativas em um intervalo de tempo, na desintegração dos núcleos de uma substância radioativa.

Suposições básicas:

- *independência* entre as ocorrências do evento considerado;
- os eventos ocorrem de forma *aleatória*, de tal forma que não haja tendência de aumentar ou reduzir as ocorrências do evento, no intervalo considerado.

Para desenvolvermos a distribuição de Poisson, consideremos a variável aleatória $X = \text{número de consultas a uma base de dados em um minuto}$. Ou seja, X é a contagem de ocorrências de consultas no intervalo de tempo $[0, 1)$, como representado a seguir:



Considere o intervalo $[0, 1)$ particionado em n subintervalos de amplitude $\Delta t = 1/n$:



Seja n suficientemente grande para que a probabilidade de ocorrer duas ou mais consultas, em cada subintervalo de amplitude Δt , seja desprezível. Assim, considere que em cada subintervalo só possa ocorrer 0 ou 1 consulta. Sendo p a probabilidade de ocorrer uma consulta em Δt , as probabilidades associadas a X podem ser calculadas, aproximadamente, pela binomial (Expressão 5.13). Então,

$$p(x) = P(X = x) \approx \binom{n}{x} \cdot p^x \cdot (1 - p)^{n-x}, \quad x = 0, 1, \dots, n \quad (5.19)$$

Mas quando $n \rightarrow \infty$ e $p \rightarrow 0$, de tal sorte que o valor esperado $E(X) = n \cdot p \rightarrow \lambda$, sendo $\lambda > 0$, é possível mostrar que:

$$\binom{n}{x} \cdot p^x \cdot (1 - p)^{n-x} \xrightarrow[\substack{p \rightarrow 0 \\ np \rightarrow \lambda}]{n \rightarrow \infty} \frac{e^{-\lambda} \lambda^x}{x!}, x = 0, 1, 2, \dots \quad (5.20)$$

Então, sendo λ a taxa média de consultas por unidade de tempo, as probabilidades de X podem ser calculadas pela chamada *distribuição de Poisson*, cuja função de probabilidade é dada por:

$$p(x) = \frac{e^{-\lambda} \lambda^x}{x!}, x = 0, 1, 2, \dots \quad (5.21)$$

sendo:

$$E(X) = V(X) = \lambda. \quad (5.22)$$

Exemplo 5.4 Supondo que as consultas num banco de dados ocorrem de forma independente e aleatória, com uma taxa média de três consultas por minuto, calculemos a probabilidade de que no próximo minuto ocorram menos do que três consultas.

Seja X o número de consultas por minuto. Então:

$$P(X < 3) = p(0) + p(1) + p(2) = \frac{e^{-3} 3^0}{0!} + \frac{e^{-3} 3^1}{1!} + \frac{e^{-3} 3^2}{2!} = 0,4232$$

A Tabela 2 do apêndice apresenta as probabilidades acumuladas da Poisson, isto é:

$$F(x) = P(X \leq x) = \sum_{j=0}^x \frac{\lambda^j e^{-\lambda}}{j!} \quad (5.23)$$

No Exemplo 5.4, usando a Tabela 2, temos:

$$P(X < 3) = P(X \leq 2) = F(2) = 0,4232$$

Exemplo 5.4 (continuação) Calculemos a probabilidade de que nos próximos dois minutos ocorram mais do que 5 consultas.

Observe que a unidade de tempo alterou de um para dois minutos. Mas se a taxa média é de três ocorrências por minuto, então, em dois minutos, a taxa média é de seis ocorrências. Logo, no presente problema, $\lambda = 6$ e

$$P(X > 5) = 1 - P(X \leq 5) = 1 - F(5) = 1 - 0,4457 = 0,5543$$

Aproximação da binomial pela Poisson

Justificamos a distribuição de Poisson a partir da binomial, fazendo $n \rightarrow \infty$ e $p \rightarrow 0$ (Expressão 5.20). Logo, em experimentos binomiais, quando n for muito grande e p for muito pequeno, podemos usar a distribuição de Poisson com:

$$\lambda = n.p \quad (5.24)$$

Observe que se n for grande (acima de 100, por exemplo) as combinações da binomial ficam difíceis de serem calculadas. Nesse caso, o uso da aproximação Poisson torna-se imprescindível.

Exemplo 5.5 Seja uma linha de produção em que a taxa de itens defeituosos é de 0,5%. Calculemos a probabilidade de ocorrer mais do que quatro defeituosos, em uma amostra de 500 itens.

$$\lambda = n.p = (500) \cdot (0,005) = 2,5$$

$$P(X > 4) = 1 - P(X \leq 4) = 1 - F(4) = 1 - 0,8912 = 0,1088$$

EXERCÍCIOS

11. Mensagens chegam a um servidor de acordo com uma distribuição de Poisson, com taxa média de cinco chegadas por minuto.
 - a) Qual é a probabilidade de que duas chegadas ocorram em um minuto?
 - b) Qual é a probabilidade de que uma chegada ocorra em 30 segundos?
12. Em um canal de comunicação digital, a probabilidade de se receber um *bit* com erro é de 0,0002. Se 10.000 *bits* forem transmitidos por esse canal, qual é a probabilidade de que mais de quatro *bits* sejam recebidos com erro?

EXERCÍCIOS COMPLEMENTARES

13. Um armazém é abastecido mensalmente, sendo que a taxa média de abastecimento é 30 unidades/dia, com desvio padrão de 3 unidades/dia. A demanda média é de 25 unidades/dia, com desvio padrão de 4 unidades/dia. Suponha que o abastecimento e a demanda sejam independentes e, além disso, a demanda e o abastecimento num dia não alteram o abastecimento e a demanda nos dias seguintes. Qual é o valor esperado e o desvio padrão do excedente de produtos, no período de um mês?

- 14.** Suponha que 10% dos clientes que compram a crédito em uma loja deixam de pagar regularmente suas contas (prestações). Se num particular dia, a loja vende a crédito para 10 pessoas, qual é a probabilidade de que mais de 20% delas deixem de pagar regularmente as contas? Suponha que as 10 pessoas que fizeram crediário nesse dia correspondam a uma amostra aleatória de clientes potenciais desta loja.
- 15.** Em um sistema de transmissão de dados, existe uma probabilidade igual a 0,05 de um lote de dados ser transmitido erroneamente. Foram transmitidos 20 lotes de dados para a realização de um teste de análise da confiabilidade do sistema.
- Qual é o modelo teórico mais adequado para esse caso? Por quê?
 - Calcule a probabilidade de haver erro na transmissão.
 - Calcule a probabilidade de que haja erro na transmissão em exatamente 2 dos 20 lotes de dados.
 - Qual é o número esperado de erros no teste realizado?
- 16.** Numa fábrica, 3% dos artigos produzidos são defeituosos. O fabricante pretende vender 4000 peças e recebeu 2 propostas:
- Proposta 1:* o comprador A propõe examinar uma amostra de 80 peças. Se houver 3 ou menos defeituosas, ele paga 60 unidades monetárias (u.m.) por peça; caso contrário, ele paga 30 u.m. por peça.
- Proposta 2:* o comprador B propõe examinar 40 peças. Se todas forem perfeitas, ele está disposto a pagar 65 u.m. por peça; caso contrário, ele paga 20 u.m. por peça.
- Qual é a melhor proposta? (Calcule o valor esperado da venda em cada proposta.)
- 17.** O departamento de qualidade de uma empresa seleciona, aleatoriamente, alguns itens que chegam à empresa e submete-os a testes. Para avaliar um lote de transformadores de pequeno porte, o departamento de qualidade selecionou, aleatoriamente, 10 transformadores. Ele vai recomendar a aceitação do lote se não existir item defeituoso na amostra. Supondo que o processo produtivo desses transformadores gera um percentual de 3% de defeituosos, responda:
- Qual é a probabilidade de que o lote venha a ser aceito?
 - Ao analisar 8 lotes de transformadores, com amostras aleatórias de 10 itens em cada lote, qual é a probabilidade de que, no máximo, um lote seja rejeitado?
- 18.** Na comunicação entre servidores, uma mensagem é dividida em n pacotes, os quais são enviados em forma de códigos. Pelo histórico da rede, sabe-se

que cada pacote tem uma pequena probabilidade, igual a 0,01, de não chegar corretamente a seu destino e, além disso, o fato de um pacote não chegar ao destino não altera a probabilidade dos demais chegarem corretamente. Um programa corretivo garante o envio correto da mensagem quando o número de pacotes enviados erroneamente não passar de 10% do total de pacotes da mensagem. Qual é a probabilidade de uma mensagem composta de 20 pacotes ser enviada corretamente? Responder usando:

- a) a distribuição binomial;
- b) a distribuição de Poisson.

19. Uma central telefônica recebe, em média, 300 chamadas na hora de maior movimento, e pode processar, no máximo, 10 ligações por minuto. Utilizando a distribuição de Poisson, calcular a probabilidade de que a capacidade da mesa seja ultrapassada em dado minuto do horário de pico.

20. Um piso cerâmico tem, em média, 0,01 defeito por m^2 . Em uma área de 10 m x 10 m desse piso, calcule a probabilidade de ocorrer algum defeito.

21. Placas de circuito integrado são avaliadas após serem preenchidas com *chips* semicondutores. Considere que foi produzido um lote de 20 placas e selecionadas 5 para avaliação. Calcule a probabilidade de encontrar pelo menos uma placa defeituosa, supondo que o lote tenha 4 defeituosas e que tenha sido realizada:

- a) uma amostragem aleatória com reposição;
- b) uma amostragem aleatória sem reposição.

22. Suponha que o número de falhas em certo tipo de placa plástica tenha distribuição de Poisson, com taxa média de 0,05 defeito por m^2 . Na construção de um barco, é necessário cobrir uma superfície de 3 m x 2 m com essa placa.

- a) Qual é a probabilidade de que não haja falhas nessa superfície?
- b) Qual é a probabilidade de que haja mais que uma falha nessa superfície?
- c) Na construção de 5 barcos, qual é a probabilidade de que pelo menos 4 não apresentem defeito na superfície plástica?

23. Um item é vendido em lotes de 200 unidades. Normalmente o processo de fabricação gera 5% de itens defeituosos. Um comprador compra cada lote por R\$ 100,00 (alternativa 1). Um outro comprador faz a seguinte proposta: de cada lote, ele escolhe uma amostra de 15 peças; se a amostra tem 0 defeituoso, ele paga R\$ 200,00; 1 defeituoso, ele paga R\$ 50,00; mais que 1 defeituoso, ele paga R\$ 5,00 (alternativa 2). Em média, qual alternativa

é mais vantajosa para o fabricante? (Calcule os valores esperados das duas alternativas.)

- 24.** Na produção de rolhas de cortiça, não é possível garantir qualidade homogênea, devido às variações internas nas placas de cortiça. Em função disso, um equipamento separa as rolhas que saem da linha de produção em duas categorias: *A* e *B*. Os dados históricos mostram que 40% são classificadas como *A* e 60% como *B*. O fabricante vende por R\$ 100,00 o milhar de rolhas da categoria *A*; e por R\$ 60,00 o milhar da categoria *B*.

Um comprador propõe comprar a produção diária da fábrica. Ele fará um plano de amostragem, extraindo 8 rolhas aleatoriamente. Se encontrar mais que 5 rolhas da categoria *A*, ele paga R\$ 200,00; caso contrário, ele paga R\$ 50,00. Pede-se:

- a)** Qual é a probabilidade do comprador encontrar mais que 5 rolhas da classe *A*?
 - b)** Qual é o valor esperado da venda do fabricante, por milhar de rolhas vendidas, se ele aceitar a proposta do comprador? Em termos do valor esperado da venda, a proposta do comprador é mais vantajosa do que a venda separada por categoria?
 - c)** Qual é a variância da venda do fabricante, por milhar de rolhas vendidas, se ele aceitar a proposta do comprador?
- 25.** Suponha que as requisições a um sistema ocorram de forma independente e que a taxa média de ocorrências é três requisições por minuto, constante no período em estudo. Calcule a probabilidade de:
- a)** ocorrer mais que uma requisição no próximo minuto;
 - b)** ocorrer mais que uma requisição no próximo minuto, sabendo-se que é certa a ocorrência de pelo menos uma (pois, você mesmo fará uma requisição no próximo minuto).

Variáveis Aleatórias Contínuas

6.1 CARACTERIZAÇÃO DE UMA VARIÁVEL ALEATÓRIA CONTÍNUA

Muitas variáveis aleatórias que surgem na vida de um engenheiro ou de um profissional da informática têm natureza eminentemente contínua, tais como:

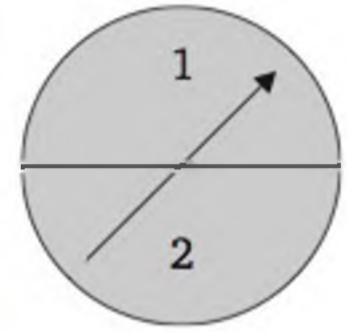
- tempo de resposta de um sistema computacional;
- rendimento de um processo químico;
- tempo de vida de um componente eletrônico;
- resistência de um material etc.

Outras vezes, há variáveis aleatórias discretas, com grande número de possíveis resultados, em que é preferível usar um modelo aproximado contínuo no lugar do modelo exato discreto. É o caso de:

- número de transações por segundo de uma CPU;
- número de defeitos numa amostra de 5.000 itens etc.

Para entender as peculiaridades das variáveis aleatórias contínuas, imagine o seguinte experimento.

Exemplo 6.1a Um círculo é dividido em dois setores de mesmo tamanho (180° cada um), aos quais são atribuídos os números 1 e 2. Um ponteiro é preso ao centro do círculo e girado, conforme mostra a figura ao lado. Seja a variável aleatória discreta $X = \text{número do setor apontado quando o ponteiro pára de girar}$. A distribuição de probabilidades de X , considerando que todos os pontos sejam equiprováveis, pode ser especificada pela função de probabilidade da Figura 6.1.



x	$p(x)$
1	$\frac{1}{2}$
2	$\frac{1}{2}$
Total	1

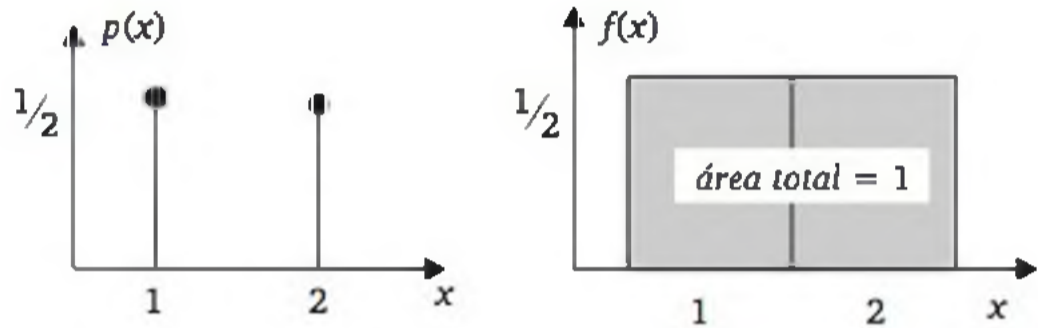


Figura 6.1 Três formas de apresentação da função de probabilidade do experimento aleatório do Exemplo 6.1a.

A representação com gráfico de hastes é típica para variáveis discretas. Apresentamos, também, um gráfico em forma de histograma, em que as probabilidades podem ser representadas por área. No caso do exemplo em questão, as bases dos retângulos são iguais à unidade, o que faz com que a área seja igual à altura do retângulo.¹

Exemplo 6.1b Considere, agora, o círculo dividido em quatro setores de mesmo tamanho (90° cada um). A distribuição de probabilidades da variável aleatória discreta $X = \text{número do setor apontado quando o ponteiro pára de girar}$ é apresentada na Figura 6.2.

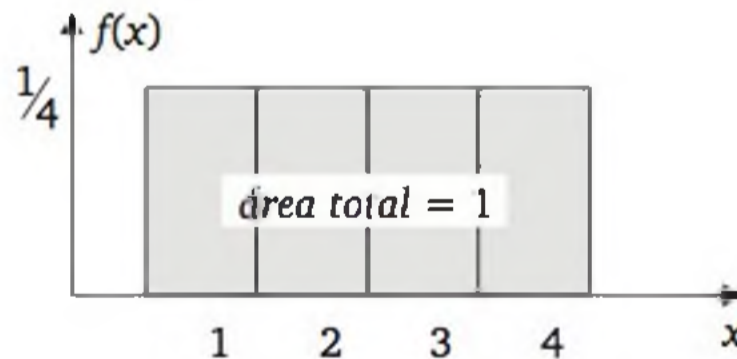
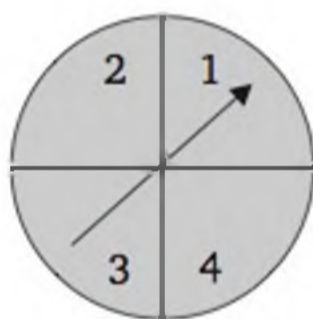


Figura 6.2 Representação gráfica da função de probabilidade do experimento aleatório do Exemplo 6.1b.

¹ Ao representarmos probabilidades por áreas, devemos tomar o cuidado para que a área total seja igual à unidade.

Exemplo 6.1c Imagine que o círculo seja dividido em 8, 16 e 32 setores. A Figura 6.3 mostra a função de probabilidade de X em cada caso.

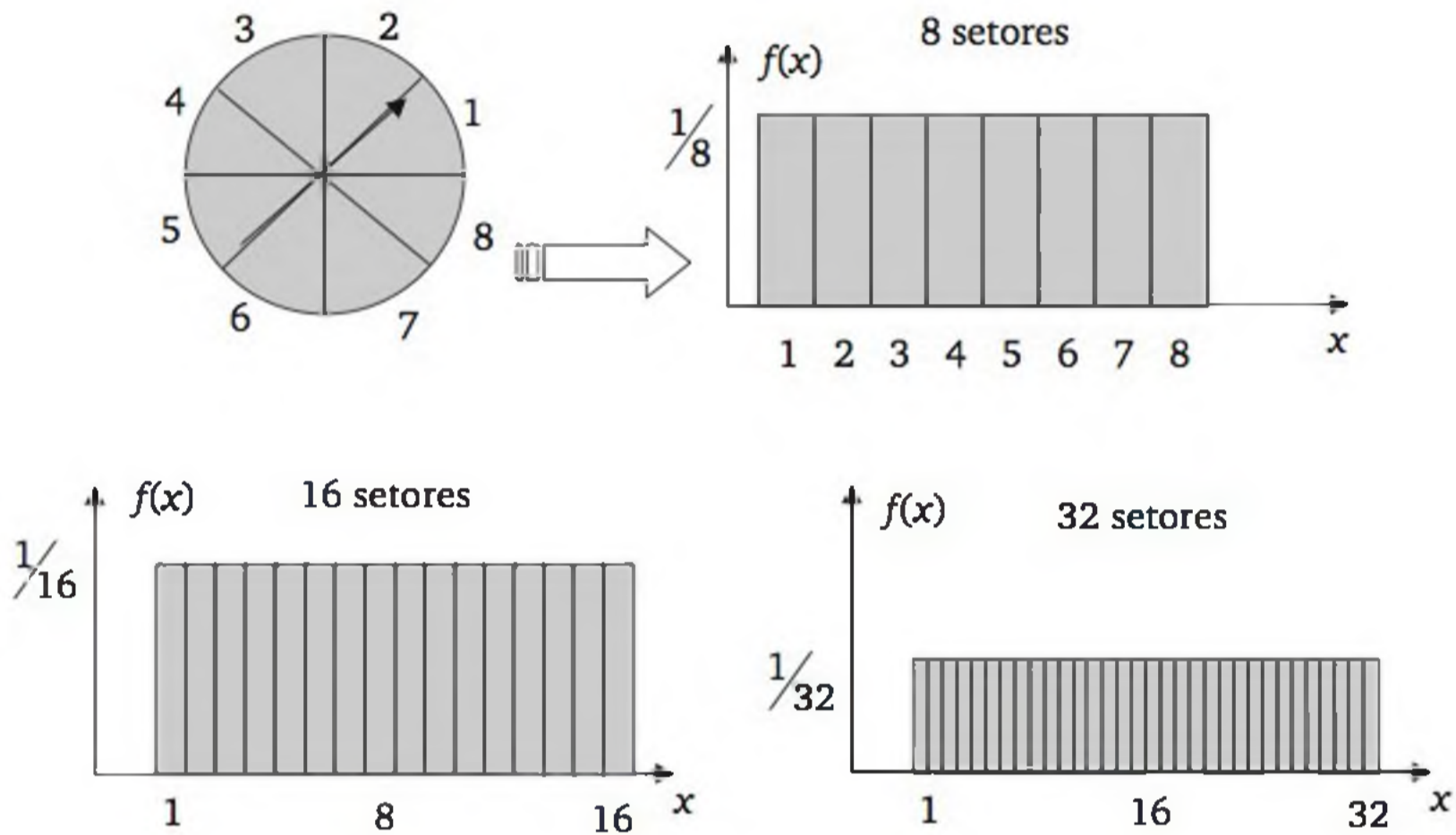


Figura 6.3 Representações gráficas das funções de probabilidade dos experimentos aleatórios do Exemplo 6.1c.

É fácil verificar que à medida que aumentamos o número de divisões no círculo, o número de possíveis setores (resultados de uma variável aleatória discreta) vai aumentando, e a probabilidade de cada resultado ocorrer (representada pela área de um retângulo) vai sendo reduzida. Teoricamente, o círculo pode ser dividido em infinitos setores, o que torna inviável a representação tabular ou gráfica da distribuição de probabilidades, da forma como fizemos no Exemplo 6.1. Em termos matemáticos, teríamos:

$$P(x) = P(X = x) = \lim_{n \rightarrow \infty} \frac{1}{n} = 0, \quad \forall x = 1, 2, \dots \quad (6.1)$$

Uma alternativa melhor é definir uma variável aleatória contínua, como veremos na seção seguinte.

6.1.1 Função densidade de probabilidade

Exemplo 6.2 Considere um círculo, com medidas de ângulos, em graus, a partir de determinada origem, como mostra a figura ao lado. Nesse círculo, há um ponteiro que é colocado a girar.

Seja a variável aleatória contínua $X = \text{ângulo formado entre a posição que o ponteiro pára e a linha horizontal do lado direito}$. Considerando que não existe região de preferência para o ponteiro parar, a distribuição de probabilidade de X pode ser representada por uma função que assume um valor constante e positivo em todo o intervalo $[0^\circ, 360^\circ)$, de tal forma que as probabilidades possam ser representadas por áreas sob a curva dessa função. Como certamente vai ocorrer um resultado em $[0^\circ, 360^\circ)$, então a área sob a função neste intervalo deve ser igual a 1, e nula fora deste intervalo. A Figura 6.4 ilustra a distribuição de probabilidades de X , através da chamada *função densidade de probabilidade*, e mostra a relação entre uma área e um evento.

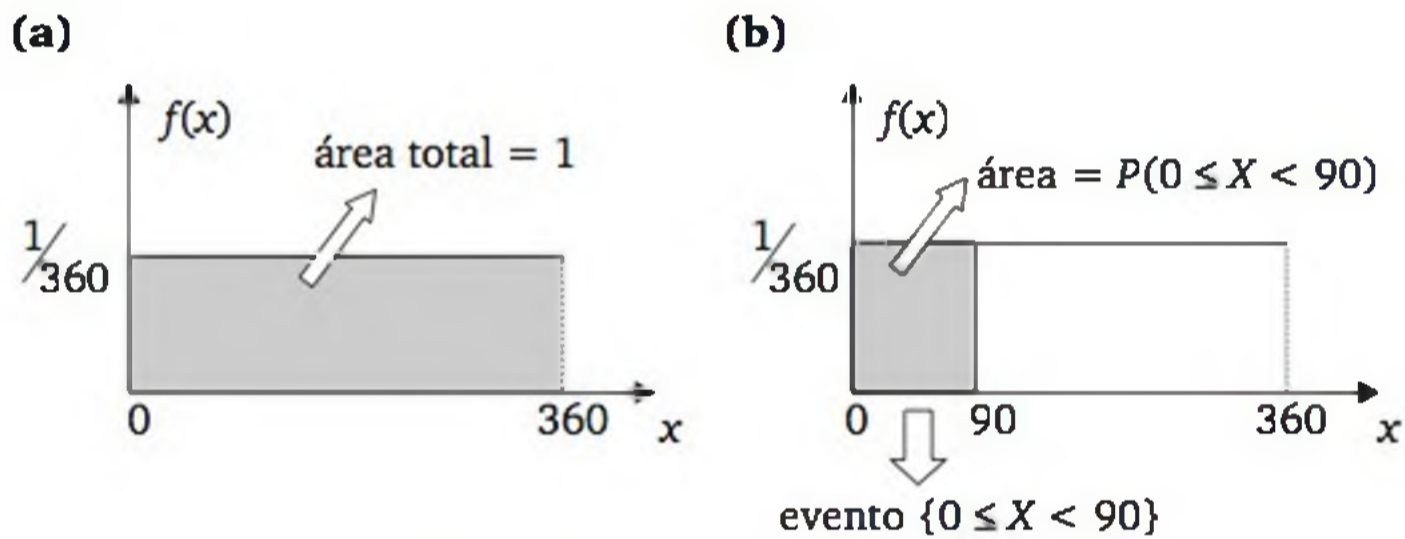
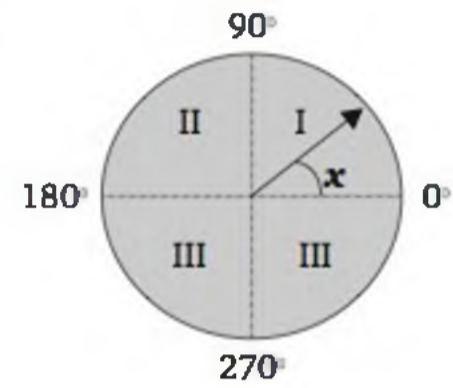
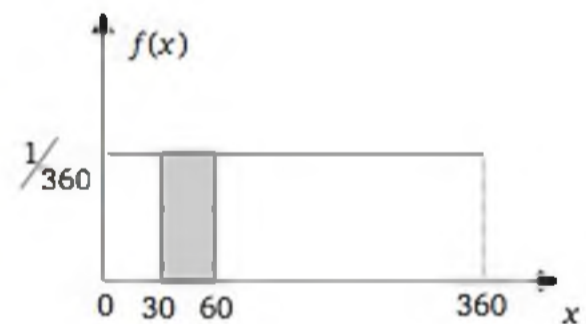


Figura 6.4 (a) Função densidade de probabilidade da variável aleatória do Exemplo 6.2; e (b) Probabilidade do evento $\{0 \leq X < 90\}$, representada por uma área.

Note que os eventos associados a uma variável aleatória contínua são intervalos (ou coleção de intervalos) dos números reais. Com base na função densidade de probabilidade, podemos calcular probabilidades de eventos desse tipo. Por exemplo, qual é a probabilidade do ponteiro parar no intervalo $[30^\circ, 60^\circ]$? Tomando a área do retângulo indicado na figura ao lado, temos:



$$P(30 \leq X \leq 60) = \frac{1}{360} \cdot (60 - 30) = \frac{30}{360} = \frac{1}{12}$$

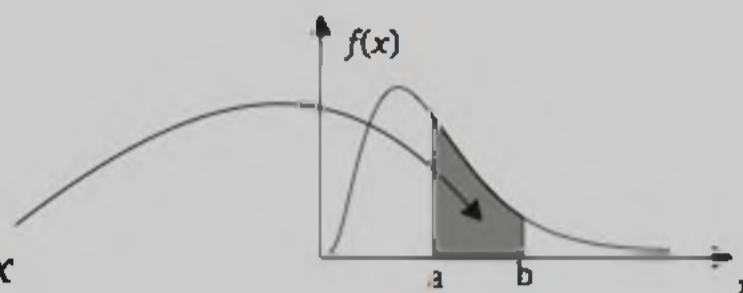
Observe que a inclusão ou exclusão dos extremos não altera a probabilidade, pois uma linha tem área nula. Ou seja, para uma variável aleatória contínua, a probabilidade de ocorrer um particular valor é igual a zero.

As probabilidades de eventos associados a uma variável aleatória contínua X podem ser calculadas através de uma **função densidade de probabilidade** f , que deve satisfazer:

a) $f(x) \geq 0, \forall x \in \mathfrak{R}$ e

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

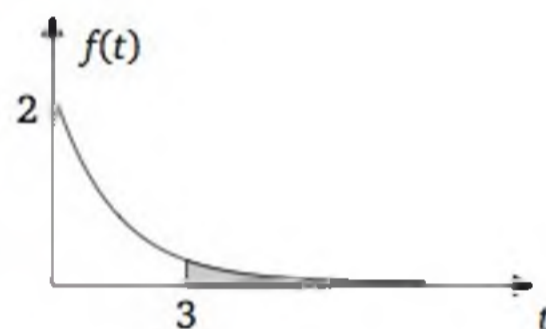
Se $A = [a, b]$, então $P(A) = \int_a^b f(x) dx$



Exemplo 6.3 Seja a variável aleatória T definida como o tempo de resposta na consulta a um banco de dados, em minutos. Suponha que essa variável aleatória tenha a seguinte função densidade de probabilidade:

$$f(t) = \begin{cases} 2e^{-2t}, & \text{para } t \geq 0 \\ 0, & \text{para } t < 0 \end{cases}$$

Vamos calcular a probabilidade de a resposta demorar mais do que 3 minutos, isto é, $P(T > 3)$.



$$P(T > 3) = \int_3^{+\infty} f(t) dt = \int_3^{+\infty} 2e^{-2t} dt = 2 \left[-\frac{1}{2} e^{-2t} \right]_3^{+\infty} = 0 + e^{-2(3)} = e^{-6}$$

6.1.2 Função de distribuição acumulada

Como X é uma variável aleatória contínua com função de densidade de probabilidade f , definimos sua *função de distribuição acumulada* por:

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(s) ds, \quad \forall x \in \mathfrak{R} \quad (6.2)$$

Considere a função densidade de probabilidade do Exemplo 6.3:

$$f(t) = \begin{cases} 2e^{-2t}, & \text{para } t \geq 0 \\ 0, & \text{para } t < 0 \end{cases}$$

Vamos obter a função de distribuição acumulada. Como a expressão matemática se altera no ponto zero, devemos considerar os dois seguintes casos:

para $t < 0$,

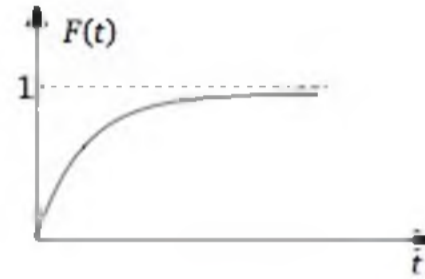
$$F(t) = \int_{-\infty}^t f(s) ds = \int_{-\infty}^t 0 ds = 0$$

e para $t \geq 0$,

$$F(t) = \int_{-\infty}^t f(s)ds = \int_{-\infty}^0 0ds + \int_0^t 2e^{-2s} ds = 0 + \left[-e^{-2s}\right]_0^t = 1 - e^{-2t}$$

Resumindo, a função de distribuição acumulada da variável aleatória T é dada por:

$$F(t) = \begin{cases} 1 - e^{-2t}, & \text{para } t \geq 0 \\ 0, & \text{para } t < 0 \end{cases}$$



Cabe observar que é possível obter qualquer probabilidade através da função de distribuição acumulada. Para $a < b$, temos:

$$P(X < a) = P(X \leq a) = F(a) \tag{6.3}$$

$$P(X > b) = 1 - F(b) \tag{6.4}$$

$$P(a < X < b) = F(b) - F(a) \tag{6.5}$$

Retomando o Exemplo 6.3, o cálculo de $P(T > 3)$ pode ser feito aplicando (6.4):

$$P(T > 3) = 1 - P(T \leq 3) = 1 - F(3) = 1 - [1 - e^{-2(3)}] = e^{-6}$$

Dada a função de distribuição acumulada F , podemos obter a função densidade de probabilidade f por:

$$f(x) = \frac{d}{dx} F(x) \tag{6.6}$$

para todo ponto x em que F é derivável.² Assim, a função F também caracteriza a distribuição de probabilidades de uma variável aleatória.

6.1.3 Valor esperado e variância

Uma variável aleatória contínua X , com função densidade de probabilidade f , tem *valor esperado* e *variância* definidos por:

$$\mu = E(X) = \int_{-\infty}^{+\infty} xf(x)dx \tag{6.7}$$

$$\sigma^2 = V(X) = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x)dx \tag{6.8}$$

² No conjunto finito de pontos em que F não é derivável, podemos arbitrar valores para f .

As interpretações dessas medidas podem ser feitas de forma análoga ao caso discreto. Além disso, todas as propriedades enunciadas para o caso discreto continuam válidas para o caso contínuo, em especial

$$\boxed{V(X) = E(X^2) - \mu^2} \quad (6.9)$$

onde: $E(X^2) = \int_{-\infty}^{+\infty} x^2 f(x) dx$

Retomando o Exemplo 6.3, em que a variável aleatória T era caracterizada por

$$f(t) = \begin{cases} 2e^{-2t}, & \text{para } t \geq 0 \\ 0, & \text{para } t < 0 \end{cases}$$

temos: $\mu = E(T) = \int_{-\infty}^{+\infty} tf(t) dt = \int_{-\infty}^0 t \cdot 0 dt + \int_0^{+\infty} t \cdot 2e^{-2t} dt = 0 + 2 \int_0^{+\infty} te^{-2t} dt$

Integrando $\int_0^{+\infty} t \cdot 2e^{-2t} dt$ por partes, obtemos $\mu = 1/2$.

Temos, também,

$$E(T^2) = \int_{-\infty}^{+\infty} t^2 f(t) dt = \int_{-\infty}^0 t^2 \cdot 0 dt + \int_0^{+\infty} t^2 \cdot 2e^{-2t} dt = 0 + 2 \int_0^{+\infty} t^2 e^{-2t} dt$$

Com certo esforço matemático, obtemos $E(T^2) = 1/2$. Então:

$$\sigma^2 = V(T) = E(T^2) - \mu^2 = \frac{1}{2} - \left(\frac{1}{2}\right)^2 = \frac{1}{4}$$

EXERCÍCIOS

1. Seja um ponto escolhido aleatoriamente no intervalo $[0, 1]$.
 - a) Apresente uma função densidade de probabilidade para este experimento.
 - b) Obtenha a função de distribuição acumulada.
 - c) Calcule o valor esperado e a variância.
2. Um profissional de Computação observou que seu sistema gasta entre 20 e 24 segundos para realizar determinada tarefa. Considere a probabilidade uniforme em $[20, 24]$, isto é, todo subintervalo de mesma amplitude em

[20, 24] tem a mesma probabilidade. Como pode ser descrita, gráfica e algebricamente, a função densidade de probabilidade? Sob essa densidade, calcule:

- a) $P(X > 23)$;
- b) $E(X)$;
- c) $V(X)$.

3. Com respeito ao exercício anterior, mas supondo probabilidades maiores em torno de 22 segundos e a densidade decrescendo, simétrica e linearmente, até os extremos 20 e 24 segundos. Como pode ser descrita, gráfica e algebricamente, a função densidade de probabilidade? Sob essa densidade, calcule:

- a) $P(X > 23)$;
- b) $E(X)$;
- c) $V(X)$.

Comparando os gráficos das funções de densidade de probabilidade dos dois exercícios, você acha razoáveis as diferenças encontradas nos três itens?

4. Seja X uma variável aleatória com função de distribuição acumulada

$$F(x) = \begin{cases} 1 - e^{-x}, & \text{para } x \geq 0 \\ 0, & \text{para } x < 0 \end{cases}$$

Obtenha a função densidade de probabilidade de X .

5. Seja X com função densidade de probabilidade dada por

$$f(x) = \begin{cases} x, & \text{para } 0 \leq x < 1 \\ 2 - x, & \text{para } 1 \leq x < 2 \\ 0, & \text{para } x \notin [0, 2) \end{cases}$$

Calcule:

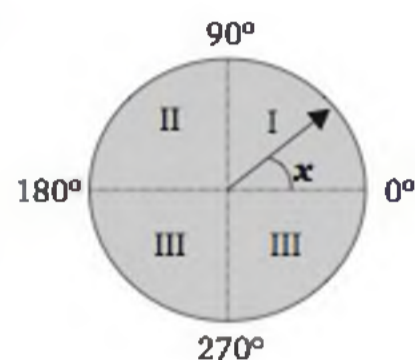
- a) $P(0 < X < 5)$
- b) $P(0 < X < 1)$
- c) $P(1/3 < X < 3/2)$
- d) $E(X)$
- e) $V(X)$

6.2 PRINCIPAIS MODELOS CONTÍNUOS

Nesta seção serão descritos três modelos contínuos bastante conhecidos.

6.2.1 Distribuição uniforme

Relembremos o Exemplo 6.2, onde tínhamos um círculo e um ponteiro que era colocado a girar. A variável aleatória de interesse era $X = \text{ângulo formado entre a posição que o ponteiro pára e a linha horizontal do lado direito}$. Supôs-se, também, não existir região de preferência para o ponteiro parar. Nessas condições, podemos considerar que todo intervalo de mesma amplitude, contido em $[0^\circ, 360^\circ)$, tem a mesma probabilidade de ocorrência. É um experi-



mento típico em que a chamada *distribuição uniforme* é apropriada.

Uma variável aleatória X tem *distribuição uniforme* de parâmetros α e β , sendo $\beta > \alpha$, se sua densidade é especificada por:

$$f(x) = \begin{cases} \frac{1}{\beta - \alpha}, & \text{para } x \in [\alpha, \beta] \\ 0, & \text{para } x \notin [\alpha, \beta] \end{cases} \tag{6.10}$$

Em consequência, sua distribuição acumulada é dada por (ver Figura 6.5):

$$F(x) = \begin{cases} 0, & \text{para } x < \alpha \\ \frac{x - \alpha}{\beta - \alpha}, & \text{para } \alpha \leq x < \beta \\ 1, & \text{para } x \geq \beta \end{cases} \tag{6.11}$$

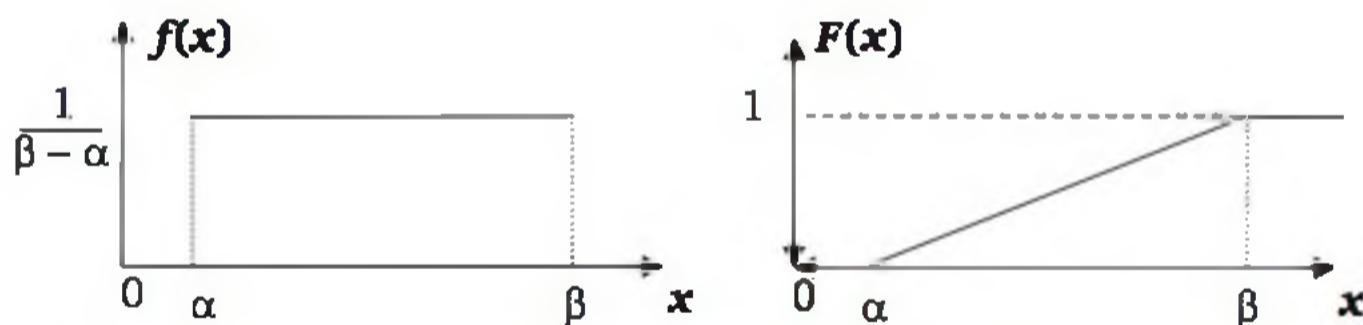


Figura 6.5 Representação gráfica da função densidade de probabilidade e da função de distribuição acumulada de uma variável aleatória com distribuição uniforme em $[\alpha, \beta]$.

O valor esperado e a variância de uma distribuição uniforme são:

$$E(X) = \frac{\alpha + \beta}{2} \tag{6.12}$$

$$V(X) = \frac{(\beta - \alpha)^2}{12} \tag{6.13}$$

Note que o valor esperado da distribuição uniforme é exatamente o ponto médio do intervalo $[\alpha, \beta]$, ou seja, nessa distribuição fica evidente que μ representa o centro de gravidade da massa descrita pela função densidade de probabilidade.

6.2.2 Distribuição exponencial

O modelo exponencial tem forte relação com o modelo discreto de Poisson. Enquanto a distribuição de Poisson pode ser usada para modelar o *número de ocorrências* em um período contínuo (de tempo ou de comprimento), a distribuição exponencial pode modelar a variável aleatória contínua que representa o *intervalo* (de tempo ou de comprimento) entre as ocorrências. Exemplos:

- a) tempo (em minutos) até a próxima consulta a uma base de dados;
- b) tempo (em segundos) entre pedidos a um servidor;
- c) distância (em metros) entre defeitos de uma fita.

A distribuição exponencial pode ser usada quando as suposições de Poisson (independência entre as ocorrências e taxa média de ocorrência constante no intervalo considerado) estiverem satisfeitas. A Figura 6.6 ilustra a relação entre as duas distribuições.

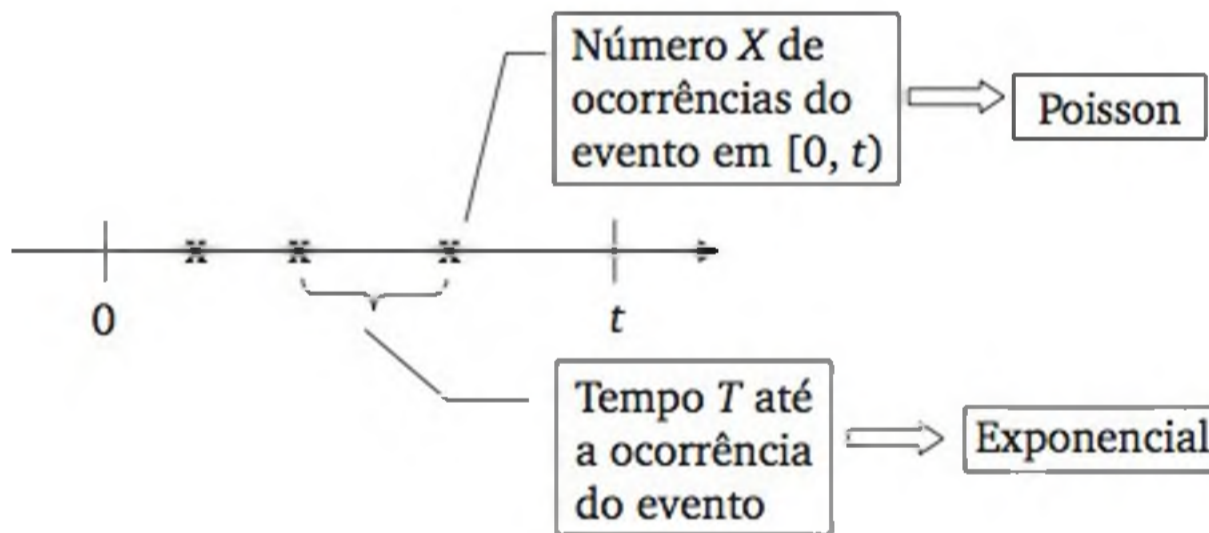
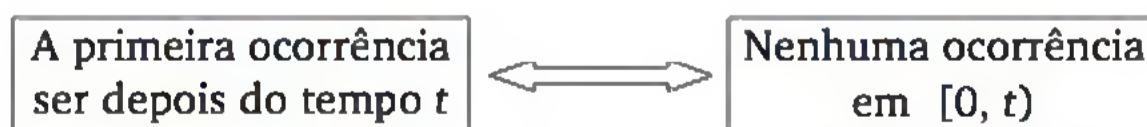


Figura 6.6 Relação entre a distribuição de Poisson e a exponencial.

Para chegarmos à formulação matemática da distribuição exponencial, vamos considerar a equivalência entre os dois seguintes eventos:



Sejam as variáveis aleatórias:

X_t = número de ocorrências no intervalo de tempo $[0, t)$; e

T = tempo entre as ocorrências.

Sendo λ a taxa média de ocorrências por unidade de tempo, então, considerando independência entre as ocorrências, X_t tem distribuição de Poisson com parâmetro λt . E a equivalência entre os dois eventos pode ser expressa por:

$$\boxed{T > t} \iff \boxed{X_t = 0}$$

Logo,

$$P(T > t) = P(X_t = 0) = \frac{(\lambda t)^0 e^{-\lambda t}}{0!} = e^{-\lambda t}$$

aplicação da expressão de Poisson

Usando o evento complementar, podemos definir para todo $t > 0$ a função de distribuição acumulada de uma variável aleatória T com distribuição exponencial:

$$\boxed{F(t) = P(T \leq t) = 1 - e^{-\lambda t}} \quad (6.14)$$

Em consequência, para $t > 0$ temos a função densidade de probabilidade dada por:

$$\boxed{f(t) = \frac{d}{dt} F(t) = \lambda e^{-\lambda t}} \quad (6.15)$$

Para $t \leq 0$, definimos $F(t) = f(t) = 0$ (ver a Figura 6.7).

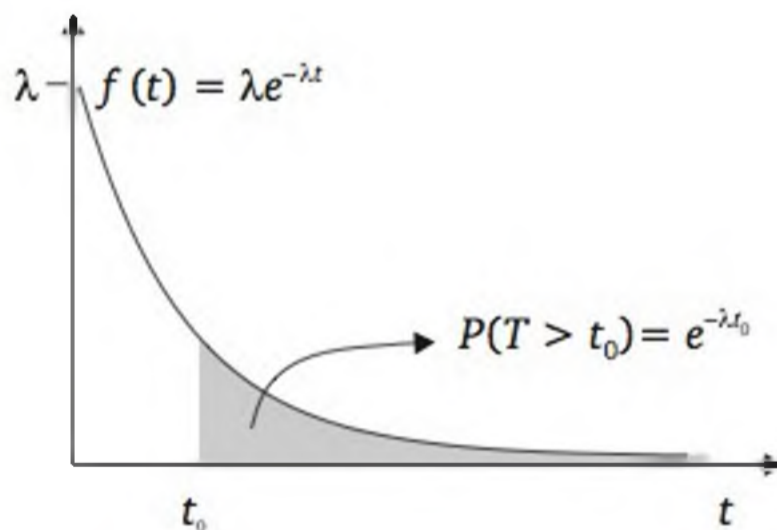


Figura 6.7 Representação gráfica da função densidade de probabilidade de uma variável aleatória com distribuição exponencial.

Em geral, é mais fácil partir do complemento de $F(t)$ para calcular as probabilidades, ou seja, para $t > 0$,

$$P(T > t) = e^{-\lambda t} \quad (6.16)$$

Exemplo 6.3 (continuação) Dada a variável aleatória $T =$ tempo de resposta na consulta a um banco de dados (em minutos) com função densidade de probabilidade

$$f(t) = \begin{cases} 2e^{-2t} & \text{para } t \geq 0 \\ 0, & \text{para } t < 0 \end{cases}$$

ou seja, uma exponencial com $\lambda = 2$, calcular a probabilidade da consulta demorar mais que 3 minutos, isto é, $P(T > 3)$. Podemos partir da função de densidade, fazendo:

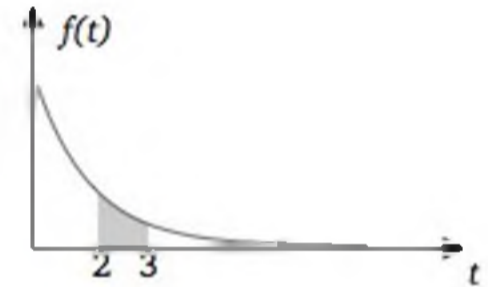
$$P(T > 3) = \int_3^{+\infty} f(t)dt = \int_3^{+\infty} 2e^{-2t} dt = e^{-6}$$

Ou podemos usar (6.16), obtendo:

$$P(T > 3) = e^{-2(3)} = e^{-6}$$

Considere, agora, o cálculo da probabilidade $P(2 \leq T \leq 3)$, isto é, a próxima consulta ocorrer no intervalo de 2 a 3 minutos. Podemos fazer

$$P(2 \leq T \leq 3) = \int_2^3 2e^{-2t} dt$$



ou usar (6.5):

$$P(2 \leq T \leq 3) = P(T \geq 2) - P(T > 3) = e^{-2(2)} - e^{-2(3)} = e^{-4} - e^{-6} = 0,0158$$

Para uma variável aleatória T , com distribuição exponencial de parâmetro λ , temos:

$$E(T) = \frac{1}{\lambda} \quad (6.17)$$

$$V(T) = \frac{1}{\lambda^2} \quad (6.18)$$

Um exemplo do cálculo do valor esperado e da variância de uma exponencial foi feito na Seção 6.1.3. Observe que podemos obter os mesmos resultados com (6.17) e (6.18).

EXERCÍCIOS

6. O tempo de vida (em horas) de um transistor é uma variável aleatória T com distribuição exponencial. O tempo médio de vida do transistor é de 500 horas.
- Calcule a probabilidade de o transistor durar mais do que 500 horas.
 - Calcule a probabilidade de o transistor durar entre 300 e 1000 horas.
 - Sabendo-se que o transistor já durou 500 horas, calcule a probabilidade de ele durar mais 500 horas.
7. Usando a expressão de probabilidade condicional (Capítulo 4), mostrar que para $s, t > 0$, vale a seguinte relação para uma variável aleatória T exponencial:

$$P(T > s + t | T > s) = P(T > t)$$

Essa propriedade é conhecida como “falta de memória”, pois não importa o que aconteceu no passado ($T \leq s$), mas apenas a partir do momento em que se inicia a observação, que pode ser considerado como o instante zero. Nesse contexto, a distribuição exponencial é inadequada para representar “tempo de vida” de itens que sofrem efeito de fadiga.

6.2.3 Distribuição normal

A normal é considerada a distribuição de probabilidades mais importante, pois permite modelar uma infinidade de fenômenos naturais e, além disso, possibilita realizar aproximações para calcular probabilidades de muitas variáveis aleatórias que têm outras distribuições. É muito importante também na inferência estatística, como será observado nos capítulos seguintes.

A distribuição normal é caracterizada por uma função de probabilidade, cujo gráfico descreve uma curva em forma de sino, como mostra a Figura 6.8. Essa forma de distribuição evidencia que há maior probabilidade de a variável aleatória assumir valores próximos do centro.

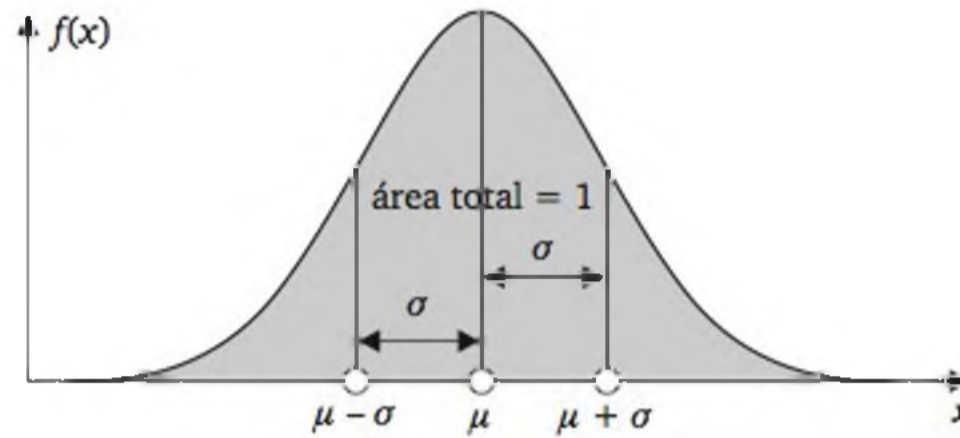


Figura 6.8 Representação gráfica da função densidade de probabilidade normal e a indicação de seus dois parâmetros: μ e σ .

Dados os parâmetros $\mu \in \mathfrak{R}$ e $\sigma > 0$, a função densidade de probabilidade da normal é dada por:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < +\infty \tag{6.19}$$

Com certo esforço matemático, é possível mostrar que o *valor esperado* e a *variância* da distribuição normal são dados por:

$$E(X) = \mu \tag{6.20}$$

$$V(X) = \sigma^2 \tag{6.21}$$

A Figura 6.9 mostra diferentes curvas normais, em função dos valores de μ e σ . As distribuições da Figura 6.8 podem representar, por exemplo, medidas da dureza de aço produzido sob diferentes condições. A distribuição (1) representa a dureza do aço em uma situação padrão; e a distribuição (2), as medidas de dureza do aço após um processo de melhoria da qualidade, em que aumentou a dureza média. A distribuição (3) representa as medidas de dureza do aço quando o processo está sob rígido controle; enquanto a distribuição (4) quando fora de controle, o que acarreta aumento na variabilidade.

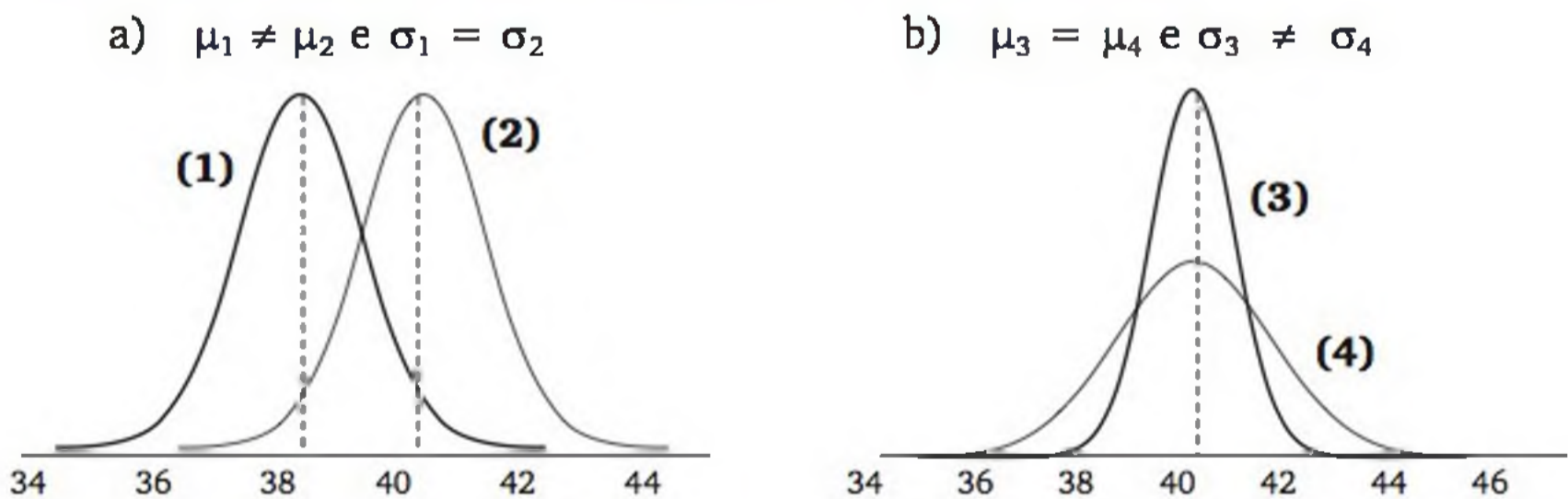


Figura 6.9 Diferentes distribuições normais em função dos parâmetros μ e σ .

Na sequência, representaremos uma variável aleatória X com distribuição normal de média μ e variância σ^2 por $X : N(\mu, \sigma^2)$. Seguem outras características do modelo normal:

- a curva é simétrica em torno de μ , em consequência, os valores da média (μ) e da mediana (m_d) são iguais, e também $P(X < \mu - \alpha) = P(X > \mu + \alpha), \forall \alpha \in \mathfrak{R}$;
- teoricamente, a curva prolonga-se de $-\infty$ a $+\infty$, sendo $\lim_{x \rightarrow \pm\infty} f(x) = 0$;
- a área total sob a curva é igual a 1, ou seja, $\int_{-\infty}^{+\infty} f(x)dx = 1$;
- qualquer combinação linear de variáveis aleatórias normais é também uma variável aleatória normal; em especial, se X_1 e X_2 são variáveis aleatórias independentes e $X_1 : N(\mu_1, \sigma_1^2)$ e $X_2 : N(\mu_2, \sigma_2^2)$, então $\forall \alpha, b \in \mathfrak{R}, Y = aX_1 + bX_2$ tem distribuição normal com

$$E(Y) = a\mu_1 + b\mu_2 \tag{6.22}$$

$$V(Y) = a^2 \sigma_1^2 + b^2 \sigma_2^2 \tag{6.23}$$

- afastamentos da média, em unidades de desvio padrão, preservam a mesma área sob a curva, independentemente dos valores de μ e σ (ver a Figura 6.10).

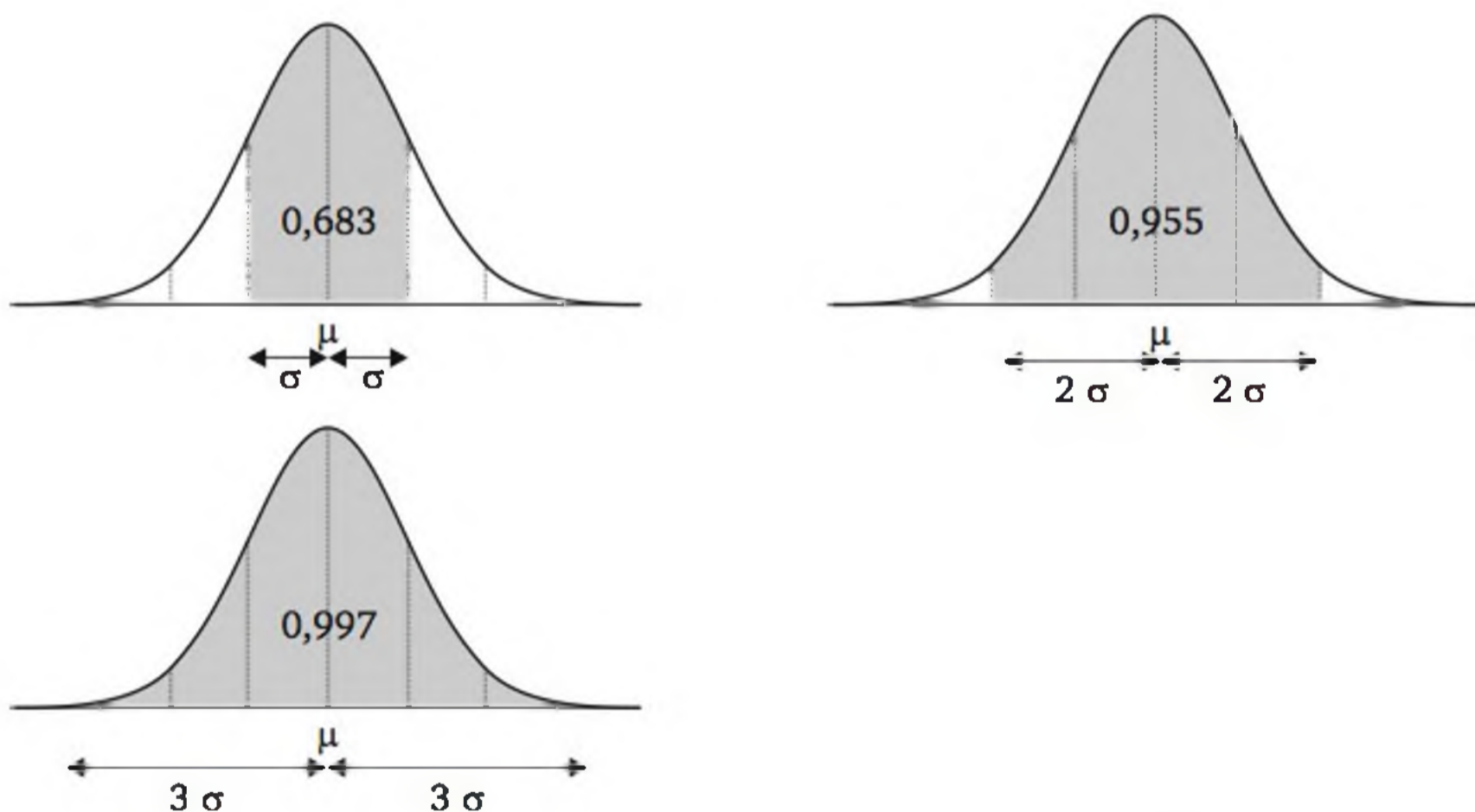


Figura 6.10 Afastamentos da média, em unidades de desvio padrão, preservam a mesma área sob a curva normal.

Seja $X : N(\mu, \sigma^2)$, então a variável aleatória

$$Z = \frac{X - \mu}{\sigma} \tag{6.24}$$

tem distribuição normal com média zero e desvio padrão unitário, ou seja, $Z : N(0,1)$, que também é conhecida como **distribuição normal padrão**. Qualquer área (probabilidade) sob a densidade de X pode ser avaliada sob a densidade de Z , conforme ilustra a Figura 6.11. Dessa forma, qualquer problema relativo a uma distribuição normal pode ser pensado em termos da distribuição normal padrão.

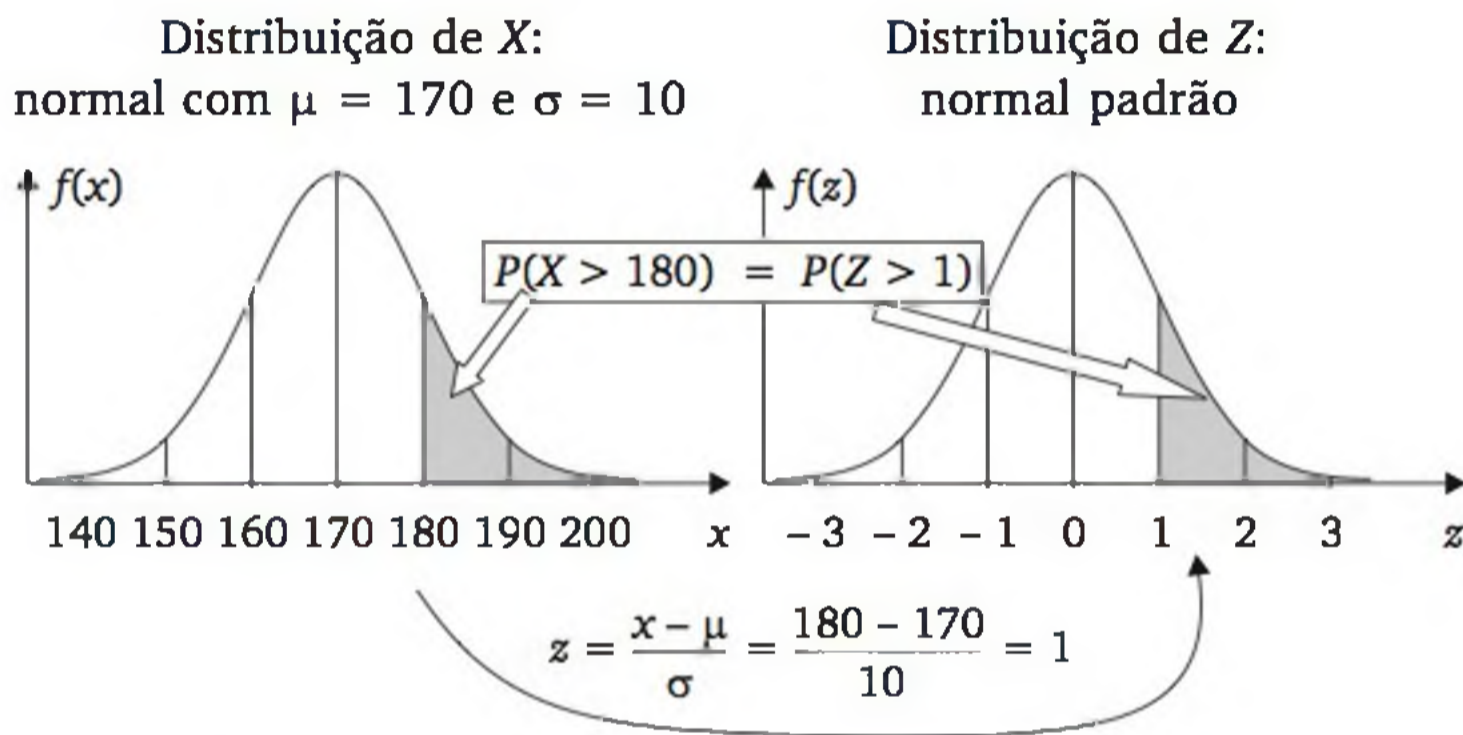


Figura 6.11 Transformação do evento $\{X > 180\}$, da distribuição normal de parâmetros $\mu = 170$ e $\sigma = 10$, em um evento da distribuição normal padrão: $\{Z > 1\}$.

Tabela da distribuição normal padrão

Como vimos, as probabilidades de uma variável com distribuição normal podem ser representadas por áreas sob a curva da distribuição normal padrão. No apêndice, apresentamos a Tabela 3, que relaciona valores positivos de z , com áreas sob a cauda superior da curva. Os valores de z são apresentados com duas decimais. A primeira decimal fica na coluna da esquerda e a segunda decimal na linha do topo da tabela. A Figura 6.12 mostra como podemos usar a Tabela 3 para encontrar uma área sob a cauda superior da curva.

z	segunda decimal de z				
	0,00	0,01	0,02	...	0,09
0,0					
0,1					
0,2		0,4168			
...					

(área na cauda superior)

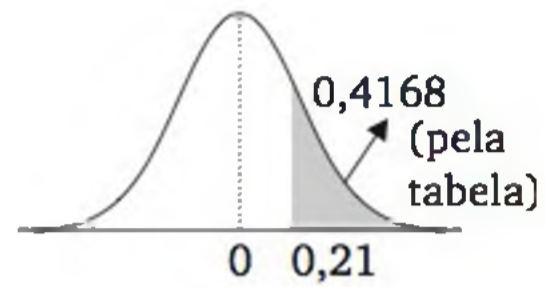
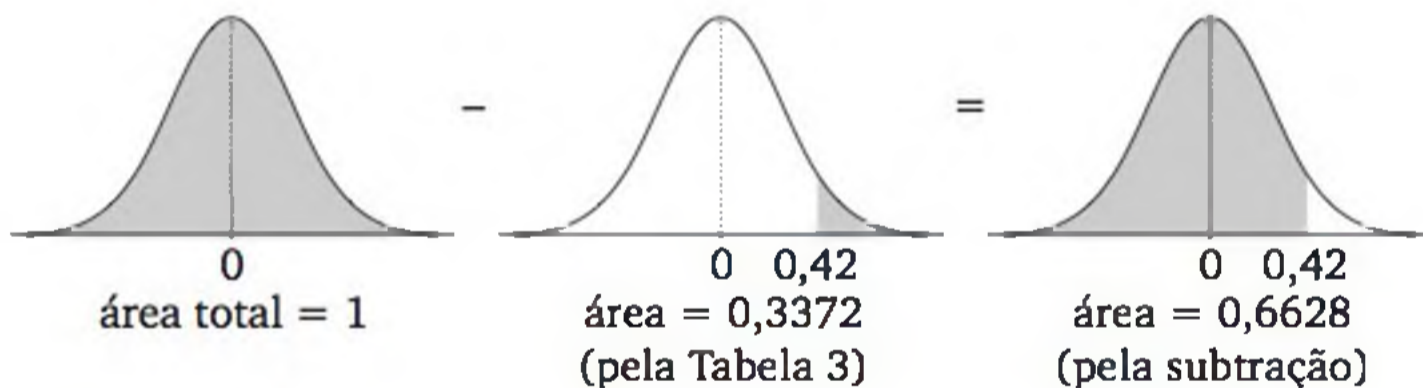
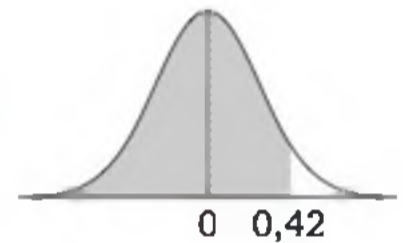


Figura 6.12 Ilustração do uso da tabela da distribuição normal padrão (Tabela 3 do apêndice) para encontrar $P(Z > 0,21)$.

A área 0,4168 corresponde à probabilidade $P(Z > 0,21) = 1 - \Phi(0,21)$, onde Φ representa a função de distribuição acumulada da normal padrão. Ou seja, a Tabela 3 fornece os valores $1 - \Phi(z)$, para $z = 0,01, 0,02, \dots, 3,00$.

Exemplo 6.4 Seja Z uma variável aleatória com distribuição normal padrão. Vamos usar a Tabela 3 para encontrar as seguintes probabilidades:

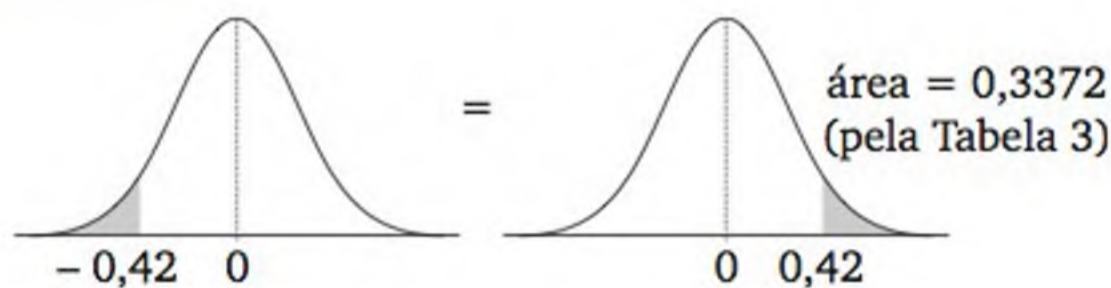
a) $P(Z < 0,42)$. Esta probabilidade corresponde à área da distribuição normal padrão indicada ao lado. Podemos obter esta área, fazendo a seguinte operação:



Mais formalmente,

$$\Phi(0,42) = P(Z < 0,42) = 1 - P(Z \geq 0,42) = 1 - 0,3372 = 0,6628$$

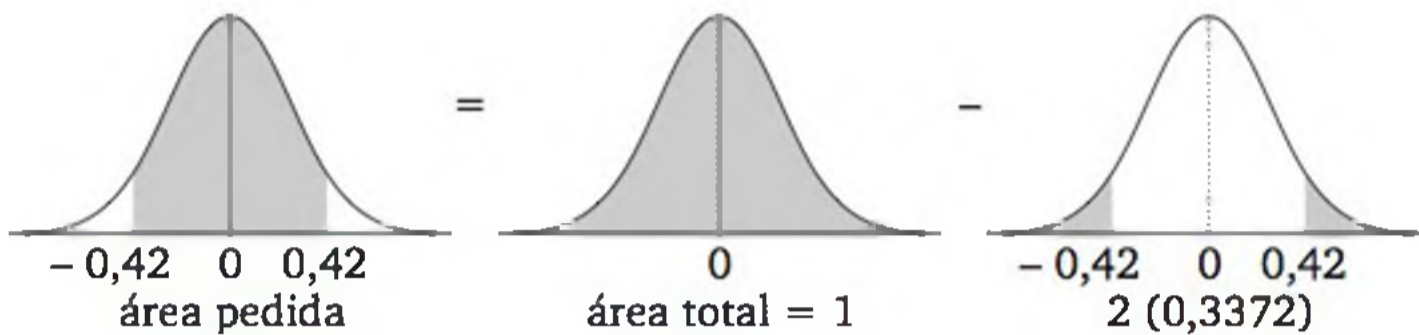
b) $P(Z < 0,42)$. O esquema seguinte mostra esta probabilidade em termos de área e como podemos usar a simetria da curva para obtê-la na Tabela 3.



Ou seja,

$$P(Z < -0,42) = P(Z > 0,41) = 0,3372$$

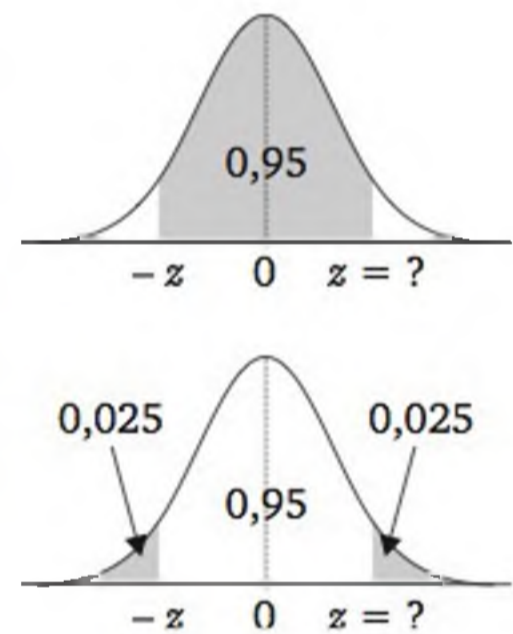
c) $P(-0,42 < Z < 0,42)$.



Então, $P(-0,42 < Z < 0,42) = 1 - 2 (0,3372) = 0,3256$.

Exemplo 6.5 Na distribuição normal padrão, qual é o valor de z , tal que $P(-z < Z < z) = 0,95$? (Veja figura ao lado.)

Considerando a simetria da curva normal e o fato de a área total sob a curva ser igual a 1 (um), podemos transformar esta pergunta em: *Qual é o valor de z tal que $P(Z > z) = 0,025$?* A figura ao lado ilustra a equivalência entre as duas perguntas.



Entrando com o valor de área 0,025 na Tabela 3 do apêndice, encontramos o valor de z igual a 1,96. Esse processo é ilustrado a seguir.

z	0,00	0,01	...	0,06	...	0,09
...				↑		
1,9	←			0,025		
...						

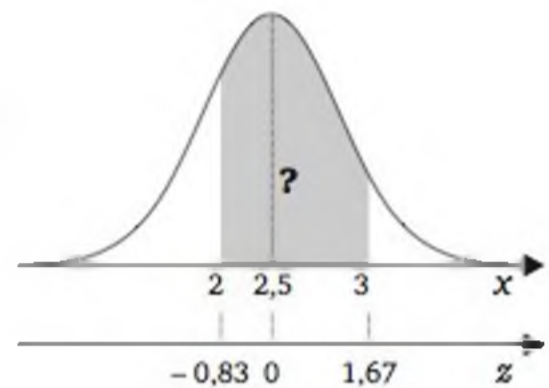
Exemplo 6.6 Suponha que a absorção de água (%) em certo tipo de piso cerâmico tenha distribuição normal com média 2,5 e desvio padrão 0,6. Selecionando, aleatoriamente, uma unidade desse piso, qual é a probabilidade de ele acusar absorção de água entre 2% e 3,5%?

Solução: Primeiramente, precisamos transformar os valores de absorção de água (x) em valores padronizados (z), por (6.24), isto é,

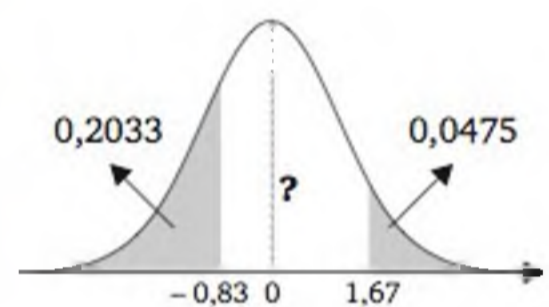
$$z = \frac{x - \mu}{\sigma} = \frac{x - 2,5}{0,6}$$

Para $x = 2$, temos: $z = \frac{2 - 2,5}{0,6} = -0,83$

e para $x = 3,5$, temos: $z = \frac{3,5 - 2,5}{0,6} = 1,67$.



Usando a Tabela 3 do apêndice, encontramos para $z = 0,83$ e $z = 1,67$ as respectivas áreas nas extremidades da curva: 0,2033 e 0,0475 (lembrando que para valores negativos de z , como $-0,83$, procuramos na Tabela 3 seu valor absoluto, 0,83). É fácil observar, pela figura ao lado, que a probabilidade desejada corresponde ao complemento da soma dessas áreas, ou seja:



$$P(2 < X < 3,5) = P(-0,83 < Z < 1,67) = 1 - (0,2033 + 0,0475) = 0,7492.$$

EXERCÍCIOS

8. Seja Z uma variável aleatória com distribuição normal padrão. Calcule:

- $P(Z > 1,65)$;
- $P(Z < 1,65)$;
- $P(-1 < Z < 1)$;
- $P(-2 < Z < 2)$;
- $P(-3 < Z < 3)$;
- $P(Z > 6)$;
- o valor de z , tal que $P(-z < Z < z) = 0,90$;
- o valor de z , tal que $P(-z < Z < z) = 0,99$.

9. Suponha que o tempo de resposta na execução de um algoritmo é uma variável aleatória com distribuição normal de média 23 segundos e desvio padrão de 4 segundos. Calcule:
- a) a probabilidade de o tempo de resposta ser menor do que 25 segundos;
 - b) a probabilidade de o tempo de resposta ficar entre 20 e 30 segundos.
10. Certo tipo de conserva tem peso líquido (X_1) com média de 900 g e desvio padrão de 10 g. A embalagem tem peso (X_2) com média de 100 g e desvio padrão de 4 g. Suponha X_1 e X_2 independentes e com distribuições normais.
- a) Qual é a probabilidade de o peso bruto ser superior a 1.020 g?
 - b) Qual é a probabilidade do peso bruto estar entre 980 e 1.020 g?

6.3 A NORMAL COMO LIMITE DE OUTRAS DISTRIBUIÇÕES

Muitas distribuições de probabilidade aproximam-se da distribuição normal. É o caso da binomial quando n é grande e da Poisson quando λ é grande.

6.3.1 Aproximação normal à binomial

Nos experimentos binomiais, quando n é muito grande, o uso da função de probabilidade binomial é impraticável, pois os coeficientes binomiais tornam-se exageradamente grandes. Já vimos que nos casos em que n é grande e p é muito pequeno, podemos usar a distribuição de Poisson para calcular, aproximadamente, as probabilidades de uma binomial. Quando n é grande e p não é próximo de 0 ou de 1, a distribuição normal pode ser usada para calcular, aproximadamente, as probabilidades de uma binomial.

A Figura 6.13 apresenta gráficos das distribuições de probabilidades binomiais com $n = 1, 10$ e 50 e $p = 0,5$ e $0,2$.

Observando a Figura 6.13, verificamos que quando $n = 50$, a forma da distribuição binomial é parecida com a curva de uma distribuição normal. Observe, ainda, que se $p = 0,5$, a aproximação já parece razoável para $n = 10$.

De maneira geral, as condições para fazer uma aproximação da distribuição binomial para a normal são:

- 1) n **grande** e
- 2) p **não** muito próximo de 0 (zero) ou de 1 (um).

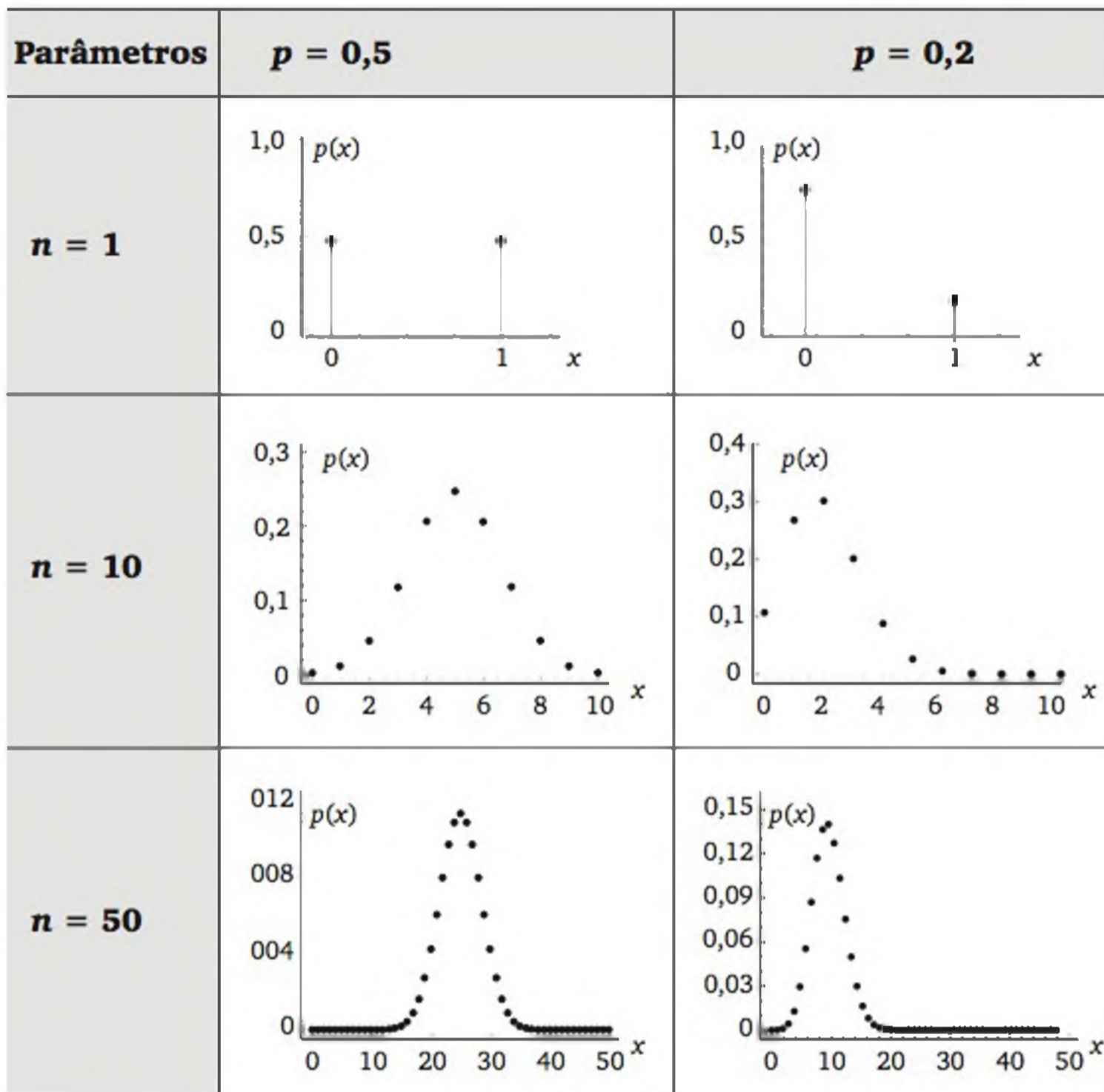


Figura 6.13 Distribuições binomiais para diferentes valores de n e p .

Uma regra prática, sugerida por vários autores, considera a aproximação razoável se as duas seguintes inequações estiverem satisfeitas:

$$np \geq 5 \text{ e} \tag{6.25}$$

$$n(1 - p) \geq 5 \tag{6.26}$$

Os parâmetros μ e σ da distribuição normal devem-se identificar ao valor esperado e ao desvio padrão do modelo binomial, ou seja:

$\mu = np$	(6.27)
$\sigma = \sqrt{np(1 - p)}$	(6.28)

Exemplo 6.7 Historicamente, 10% dos pisos cerâmicos, que saem de uma linha de produção, têm algum defeito leve. Se a produção diária é de 1000 unidades, qual é a probabilidade de ocorrer mais de 120 itens defeituosos?

Pelas características do experimento, a variável aleatória $Y = \text{número de defeituosos na amostra}$ tem distribuição binomial com parâmetros $n = 1000$ e $p = 0,1$. Verificamos, também, que as condições 6.25 e 6.26 estão satisfeitas, pois

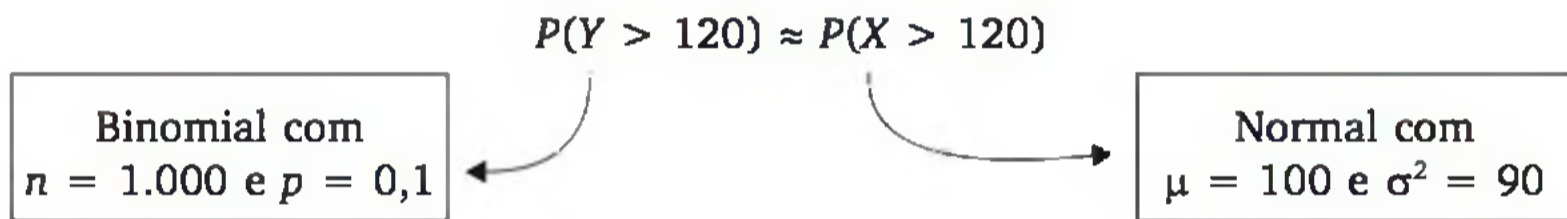
- a) $np = (1000)(0,1) = 100 \geq 5$ e
 b) $n(1 - p) = (1000)(0,9) = 900 \geq 5$

Usando (6.27) e (6.28), temos:

$$\mu = np = 1000 \cdot (0,1) = 100 \text{ e}$$

$$\sigma = \sqrt{np(1 - p)} = \sqrt{1000 \cdot (0,1) \cdot (0,9)} = \sqrt{90}$$

Considere X uma variável aleatória normal com média $\mu = 100$ e variância $\sigma^2 = 90$. Então:



Valor padronizado: $z = \frac{x - \mu}{\sigma} = \frac{120 - 100}{\sqrt{90}} = 2,11$

Assim, $P(X > 120) = P(Z > 2,11) = 0,0174$

Correção de continuidade

Ao calcularmos probabilidades de eventos oriundos de experimentos binomiais como áreas sob uma curva normal, estamos fazendo uma aproximação de uma variável aleatória discreta, que só assume valores inteiros, para uma variável contínua, cujos eventos constituem intervalos de números reais. Nesse contexto, devemos fazer alguns ajustes, como ilustra o exemplo seguinte.

Exemplo 6.8 Seja Y o número de caras obtido em dez lançamentos de uma moeda honesta. Vamos calcular a probabilidade de obter quatro caras usando a distribuição normal.

Pelas características do experimento, Y tem distribuição binomial com $n = 10$ e $p = 0,5$. Então, a média e o desvio padrão são dados por:

$$\mu = np = 10(0,5) = 5 \text{ e}$$

$$\sigma = \sqrt{np(1 - p)} = \sqrt{10 \cdot (0,5) \cdot (1 - 0,5)} = \sqrt{2,5}$$

Considere o evento: *ocorrer quatro caras*, ou seja $\{Y = 4\}$. Ao expressar este evento em termos de uma variável aleatória contínua $X : N(5, 2,5)$, devemos considerar um intervalo em torno do valor 4, pois para variáveis aleatórias contínuas só faz sentido avaliar probabilidades em *intervalos*. O intervalo adequado, nesse caso, é construído pela subtração e soma de *meia unidade* ao valor quatro, ou seja, $\{3,5 < X < 4,5\}$, como ilustra a Figura 6.14.

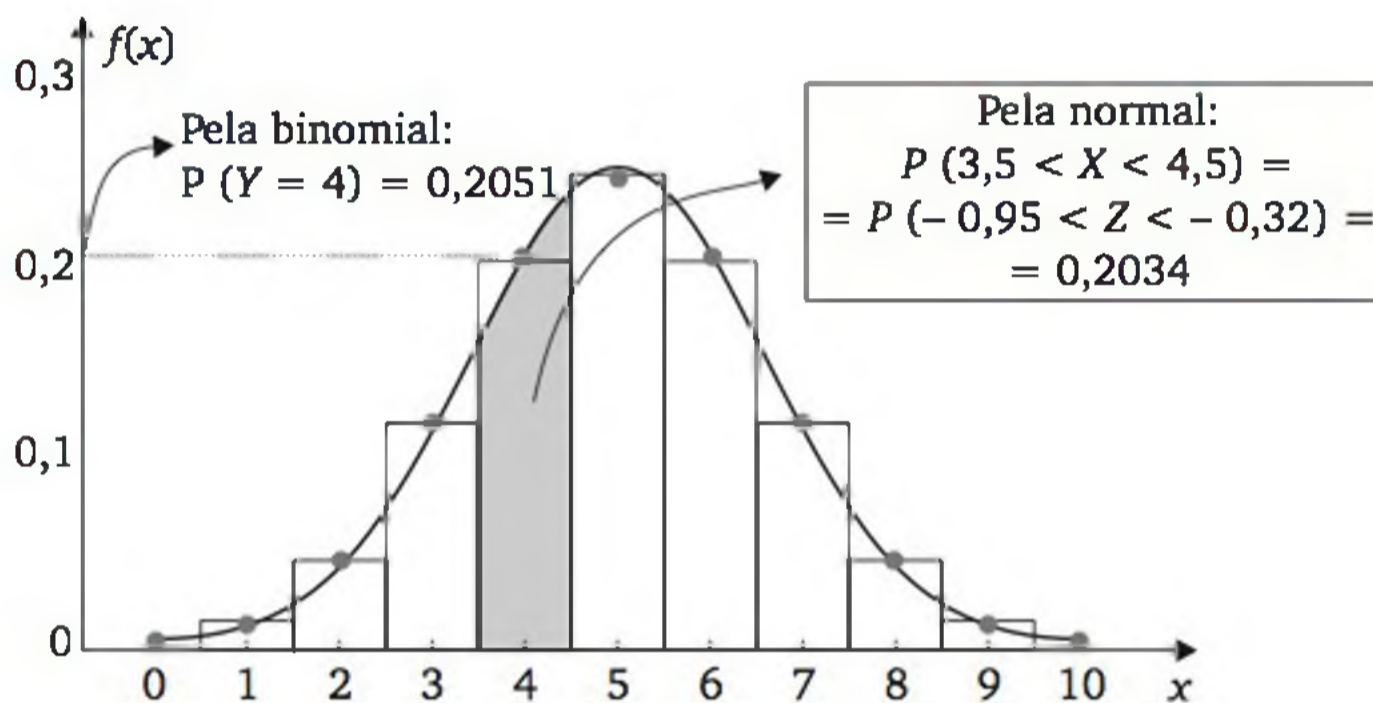


Figura 6.14 Aproximação da probabilidade do evento $\{Y = 4\}$ (em termos da distribuição binomial) para a probabilidade do evento $\{3,5 < X < 4,5\}$ (em termos da distribuição normal).

O procedimento de subtrair e somar meia unidade para construir um intervalo em torno de valores inteiros é conhecido como *correção de continuidade*. Esta correção deve ser usada ao aproximar um evento de uma variável aleatória que só assume valores inteiros para um evento de uma variável aleatória contínua.

A Figura 6.15 ilustra as diversas situações possíveis de probabilidade associada a uma variável aleatória discreta, assumindo valores em $\{0, 1, 2, \dots\}$, aproximada por probabilidade associada a uma variável aleatória contínua, a qual pode assumir qualquer valor real.

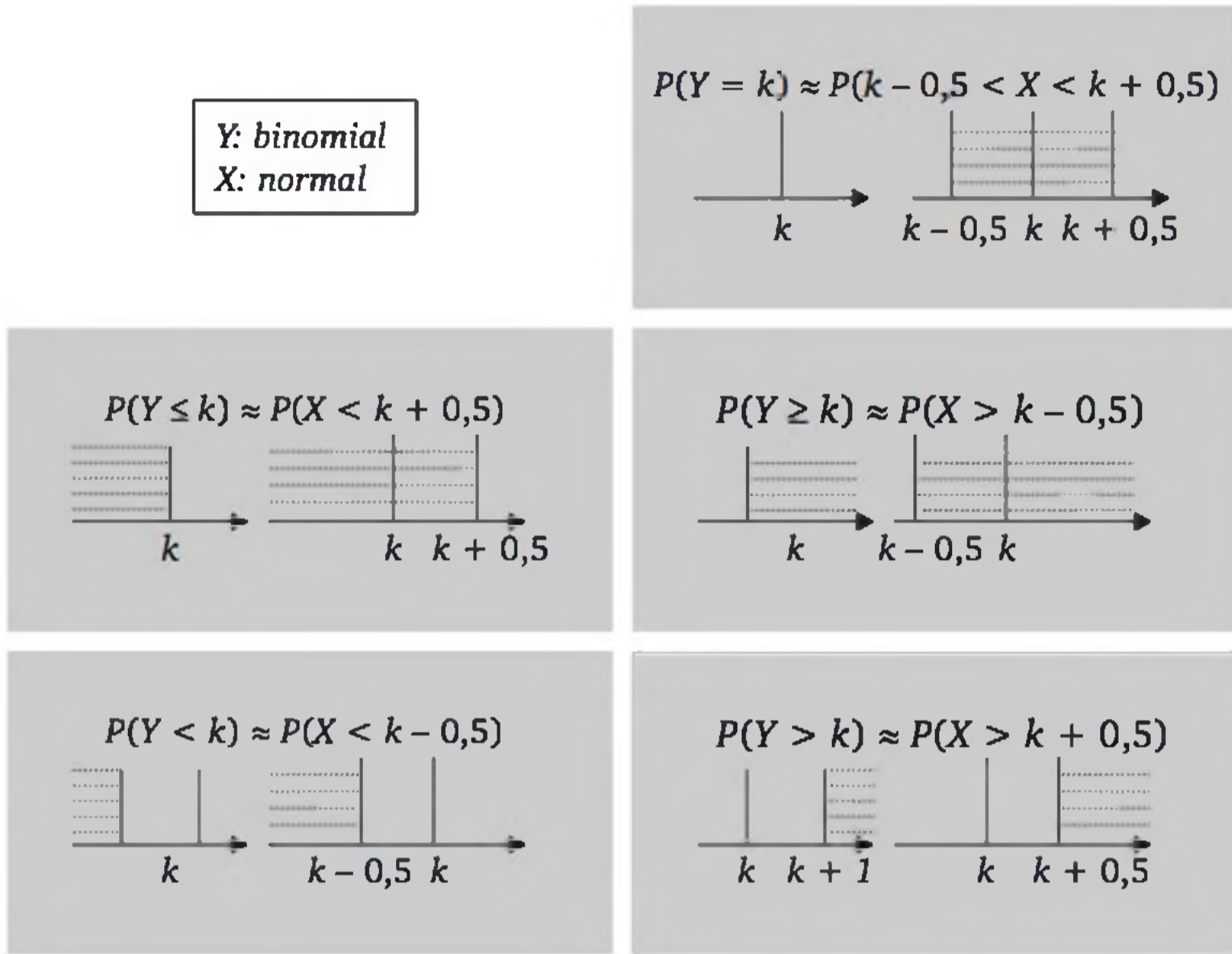


Figura 6.15 Correção de continuidade ao aproximar uma variável aleatória discreta por uma variável aleatória contínua.

6.3.2 Aproximação normal à Poisson

A distribuição de Poisson (Figura 6.16) também se aproxima da normal quando λ é grande. Como o valor esperado e a variância de uma Poisson são ambos iguais a λ , então, na aproximação normal, devemos usar:

$$\mu = \lambda$$

$$\sigma = \sqrt{\lambda}$$

(6.29)

(6.30)

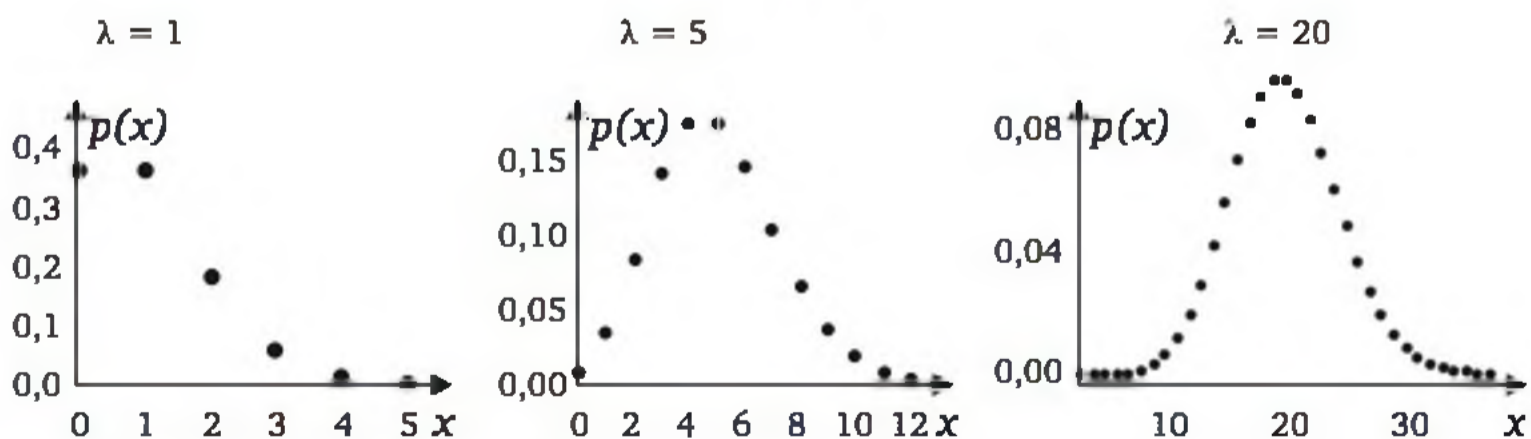


Figura 6.16 Distribuições de Poisson para diferentes valores de λ .

Conforme mostra a Figura 6.16, a aproximação é razoável para $\lambda \geq 5$. Para a aproximação da normal à Poisson, a correção de continuidade, discutida na Seção 6.4.1, também deve ser usada.

EXERCÍCIOS

11. De um lote de produtos manufaturados, extraímos 100 itens ao acaso. Se 10% dos itens do lote são defeituosos, calcular a probabilidade de:
 - a) 12 itens serem defeituosos;
 - b) mais do que 12 itens serem defeituosos.
12. Uma empresa de auxílio à lista telefônica recebe, em média, sete solicitações por minuto, segundo uma distribuição de Poisson. Qual é a probabilidade de ocorrer mais de 80 solicitações nos próximos 10 minutos?

6.4 GRÁFICO DE PROBABILIDADE NORMAL

Como veremos nos capítulos posteriores, muitos métodos estatísticos são desenvolvidos na suposição de que os dados provêm de uma distribuição normal. Quando o número de observações é grande, podemos construir um histograma e verificar se sua forma segue uma curva em forma de sino, sugerindo o modelo normal. É o caso da Figura 6.17.

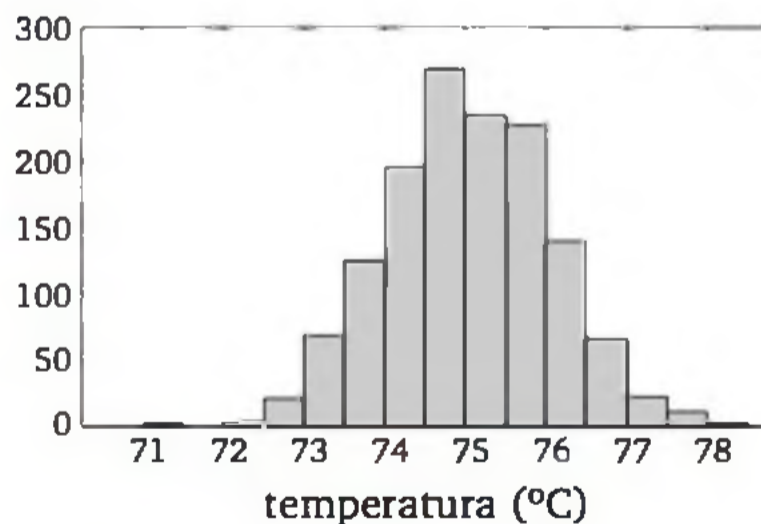


Figura 6.17 *Distribuição de frequências de 1.389 leituras da temperatura de um pasteurizador.*

Podemos, também, calcular a média (\bar{x}) e o desvio padrão (s) dos dados e verificar se os intervalos

$$\bar{x} \pm s, \bar{x} \pm 2s \text{ e } \bar{x} \pm 3s$$

têm percentuais de casos próximos dos esperados por uma distribuição normal (Figura 6.10).

Quando a quantidade de observações for pequena, o histograma pode apresentar uma forma muito diferente da real distribuição do processo que gerou os dados. Além disso, o cálculo de \bar{x} e s podem ser muito influenciados por algum valor discrepante. Em geral, o chamado *gráfico de probabilidade normal* é mais adequado para verificar a suposição de um modelo normal para os dados. Algoritmos computacionais para a construção desse gráfico estão implementados em quase todos pacotes computacionais estatísticos.

Ilustraremos a construção do gráfico de probabilidade normal com apenas cinco observações ($x_i, i = 1, 2, \dots, 5$): 74,8; 74,0; 74,7; 74,4 e 75,9. Sejam $x_{(i)}$ os valores ordenados, isto é, 74,0; 74,4; 74,7; 74,8; 75,9.

Considere uma distribuição normal com a área dividida em cinco partes iguais (mesmo número de partes do número de valores, n). E sejam $q_{(i)}$ ($i = 1, 2, \dots, 5$) os pontos medianos dos intervalos formados pela divisão das cinco áreas iguais (ver a Figura 6.18).

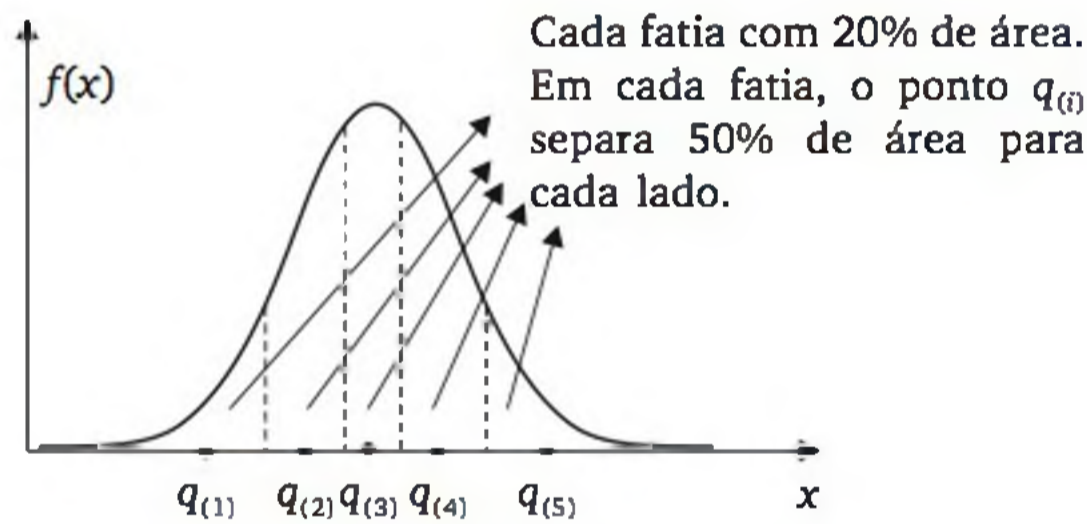
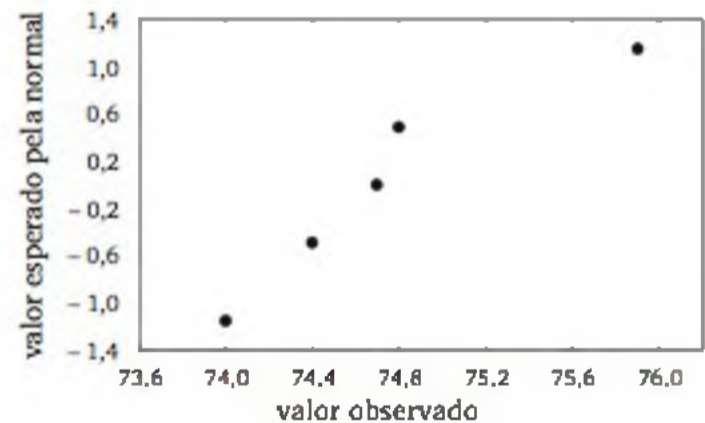


Figura 6.18 Configuração de cinco pontos com as posições relativas mais verossímeis possíveis sob um modelo normal.

Se as cinco observações (74,0; 74,4; 74,7; 74,8; 75,9) provêm de uma distribuição normal, devemos esperar uma relação aproximadamente linear com os valores teóricos $q_{(i)}$. O gráfico de probabilidade normal compreende a apresentação dos pontos $(x_{(i)}, q_{(i)})$, num par de eixos cartesianos, conforme o gráfico ao lado. Nesse gráfico, os valores de $q_{(i)}$ foram padronizados (média zero e variância 1).



A Figura 6.19 apresenta dois gráficos de probabilidade normal. O gráfico da esquerda foi construído com 40 observações que aparentemente seguem

uma distribuição normal. No gráfico da direita introduzimos um valor discrepante.

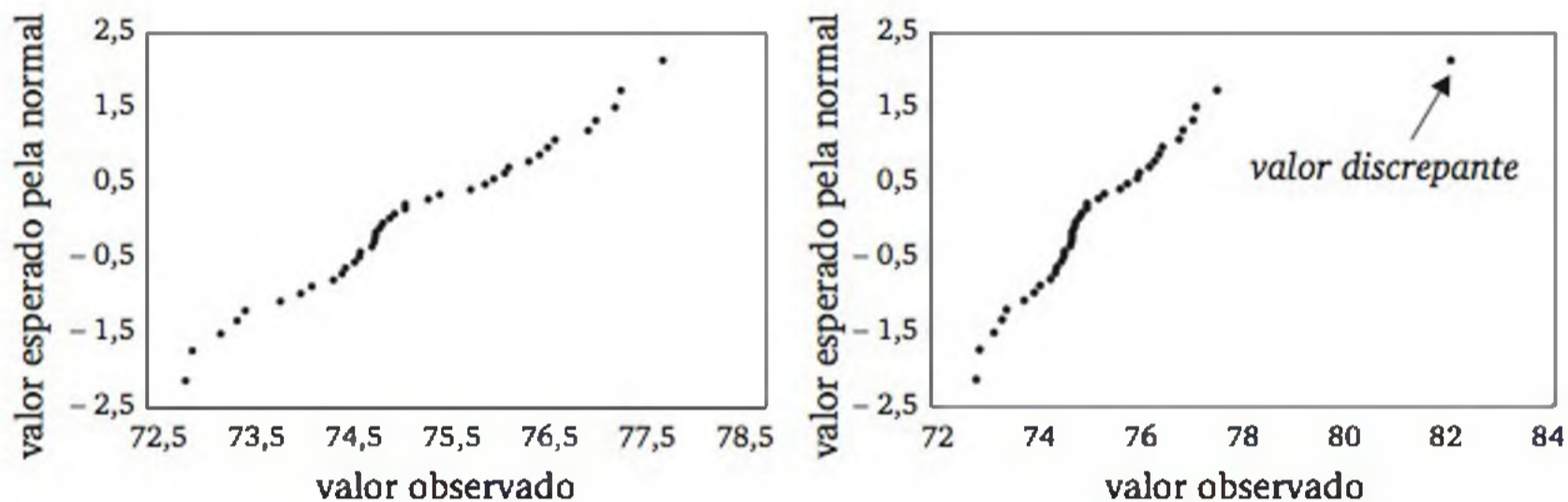


Figura 6.19 Gráfico de probabilidade normal referente a 40 leituras de temperatura de um pasteurizador e o efeito de um valor discrepante.

A Figura 6.20 mostra um gráfico de probabilidade normal construído com dados gerados por uma distribuição assimétrica, como mostrado do lado esquerdo da figura. Note que os pontos não estão aleatoriamente em torno de uma reta.

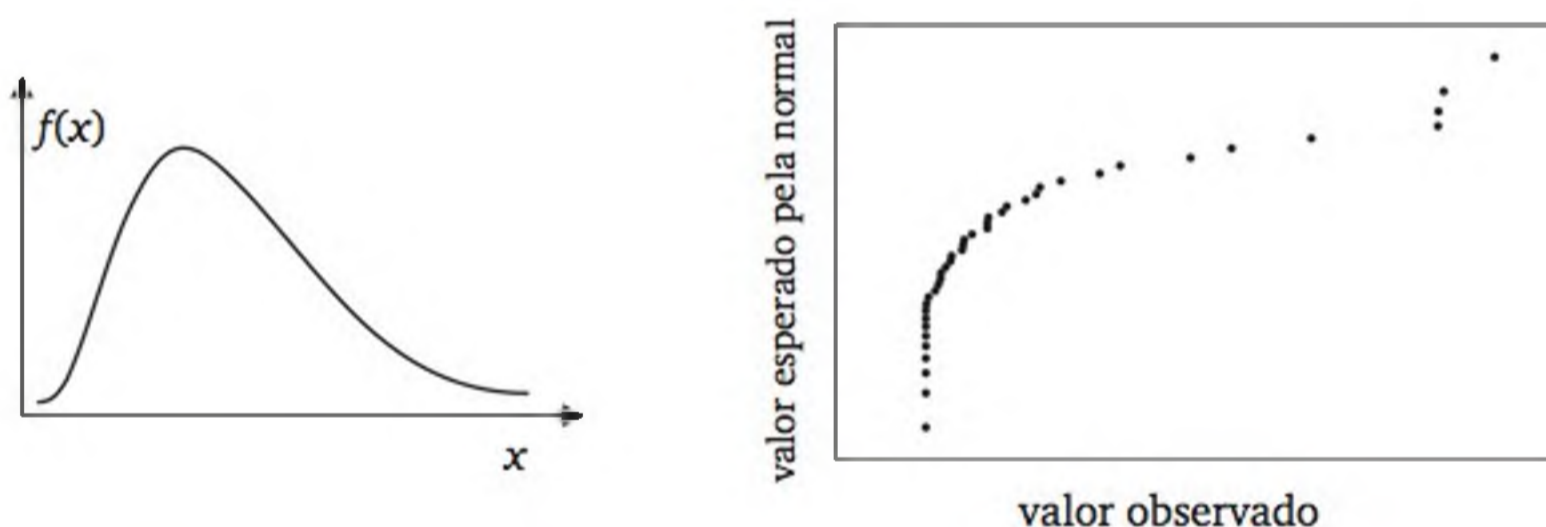


Figura 6.20 Gráfico de probabilidade normal referente a 40 observações geradas por uma distribuição assimétrica.

EXERCÍCIOS COMPLEMENTARES

- 13.** O setor de manutenção de uma empresa fez um levantamento das falhas de um importante equipamento, constatando que há, em média, 0,75 falha por ano e que o tempo entre falhas segue uma distribuição exponencial. Qual é a probabilidade de o equipamento não falhar no próximo ano?

- 14.** A vida útil de certo componente eletrônico é, em média, 10.000 horas e apresenta distribuição exponencial. Qual é a percentagem esperada de componentes que apresentarão falhas em menos de 10.000 horas?
- 15.** A vida útil de certo componente eletrônico é, em média, 10.000 horas e apresenta distribuição exponencial. Após quantas horas se espera que 25% dos componentes tenham falhado?
- 16.** Na manufatura de fios de linha para costura ocorre, em média, um defeito a cada 100 metros de linha, segundo uma distribuição de Poisson.
- Qual é a probabilidade de o próximo defeito ocorrer após 120 metros?
 - Quantos metros de linha poderão ser percorridos para que a probabilidade de aparecimento de algum defeito seja de 10%?
- 17.** Num laticínio, a temperatura do pasteurizador deve ser de 75°C . Se a temperatura ficar inferior a 70°C , o leite poderá ficar com bactérias maléficas ao organismo humano. Observações do processo mostram que valores da temperatura seguem uma distribuição normal com média $75,4^{\circ}\text{C}$ e desvio padrão $2,2^{\circ}\text{C}$.
- Qual é a probabilidade da temperatura ficar inferior a 70°C ?
 - Qual é a probabilidade de que, em 500 utilizações do pasteurizador, em mais do que cinco vezes a temperatura não atinja 70°C ? Precisa supor distribuição normal.
- 18.** O tempo para que um sistema computacional execute determinada tarefa é uma variável aleatória com distribuição normal, com *média* 320 segundos e *desvio padrão* de 7 segundos.
- Qual é a probabilidade de a tarefa ser executada entre 310 e 330 segundos?
 - Se a tarefa é colocada para execução 200 vezes. Qual é a probabilidade de ela demorar mais do que 325 segundos em pelo menos 50 vezes?
- 19.**
- Um exame de múltipla escolha consiste em dez questões, cada uma com quatro possibilidades de escolha. A aprovação exige, no mínimo, 50% de acertos. Qual é a probabilidade de aprovação se o candidato comparece ao exame sem saber absolutamente nada, apelando apenas para o “palpite”?
 - E se o exame tivesse 100 questões?
- 20.** No horário de maior movimento, um sistema de banco de dados recebe, em média, 100 requisições por minuto, segundo uma distribuição de Poisson. Qual é a probabilidade de que no próximo minuto ocorram mais de 120 requisições? Use a aproximação normal com correção de continuidade.

- 21.** Os dados históricos de uma rede de computadores sugerem que as conexões com essa rede, em horário normal, seguem uma distribuição de Poisson com média de cinco conexões por minuto. Calcule t_0 , tal que se tenha probabilidade igual a 0,90 de que ocorra pelo menos uma conexão antes do tempo t_0 .
- 22.** O padrão de qualidade recomenda que os pontos impressos por uma impressora estejam entre 3,7 e 4,3 mm. Uma impressora imprime pontos, cujo diâmetro médio é igual a 4 mm e o desvio padrão é 0,19 mm. Suponha que o diâmetro dos pontos tenha distribuição normal.
- Qual é a probabilidade do diâmetro de um ponto dessa impressora estar dentro do padrão?
 - Qual deveria ser o desvio padrão para que a probabilidade do item (a) atingisse 95%?
- 23.** Certo tipo de cimento tem resistência à compressão com média de 5.800 kg/cm², e desvio padrão de 180 kg/cm², segundo uma distribuição normal. Dada uma amostra desse cimento, calcule as seguintes probabilidades:
- resistência inferior a 5.600 kg/cm²;
 - resistência entre 5.600 kg/cm² e 5.950 kg/cm²;
 - resistência superior a 6.000 kg/cm², sabendo-se que ele já resistiu a 5.600 kg/cm².
 - se quer a garantia de que haja 95% de probabilidade de o cimento resistir a determinada pressão, qual deve ser o valor máximo dessa pressão?
- 24.** Uma empresa fabrica dois tipos de monitores de vídeo. É suposto que as durabilidades deles seguem distribuições normais, sendo o monitor M1 com média de 6 anos e desvio padrão 2,3 anos; e o monitor M2 com média de 8 anos e desvio padrão 2,8 anos. M1 tem 2 anos de garantia e M2 tem 3 anos. A empresa lucra R\$ 100,00 a cada M1 vendido e R\$ 200,00 a cada M2 vendido, mas se deixarem de funcionar no período de garantia, a empresa perde R\$ 300,00 (no caso de M1) e R\$ 800,00 (no caso de M2). Em média, qual é o tipo de monitor que gera mais lucro?

Distribuições Amostrais e Estimação de Parâmetros

Este capítulo apresenta a base para aprendermos a estatística indutiva, a qual fornece procedimentos formais para tirar conclusões sobre uma população, a partir dos dados de uma amostra. Para isso, veremos como se relacionam *estatísticas* (características dos elementos de uma amostra) com *parâmetros* (características dos elementos de uma população).

7.1 PARÂMETROS E ESTATÍSTICAS

Relembremos alguns conceitos básicos:

População: conjunto de elementos que formam o universo de nosso estudo e que são passíveis de ser observados, sob as mesmas condições.

Amostra: uma parte dos elementos de uma população.

Amostragem: o processo de seleção da amostra.

Amostragem aleatória simples: o processo de seleção é feito por sorteio, fazendo com que todos os elementos da população tenham a mesma chance de serem escolhidos e, além disso, todo subconjunto de n elementos tenha a mesma chance de fazer parte da amostra.

Quando a amostragem é aleatória e, em especial, quando é aleatória simples, podemos fazer inferências sobre a população, com base no estudo de uma amostra (ver a Figura 7.1).

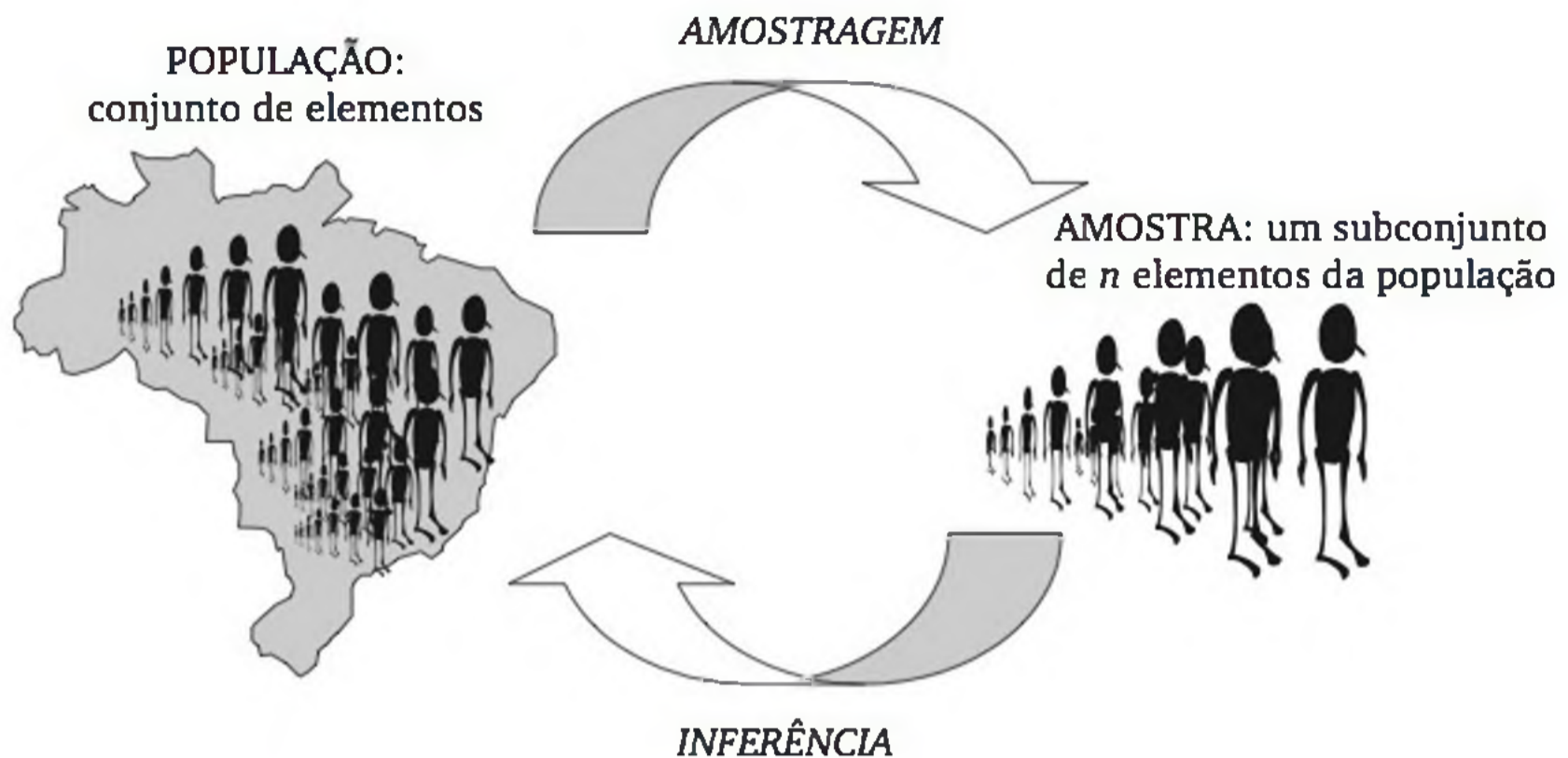


Figura 7.1 Ilustração de conceitos básicos da estatística.

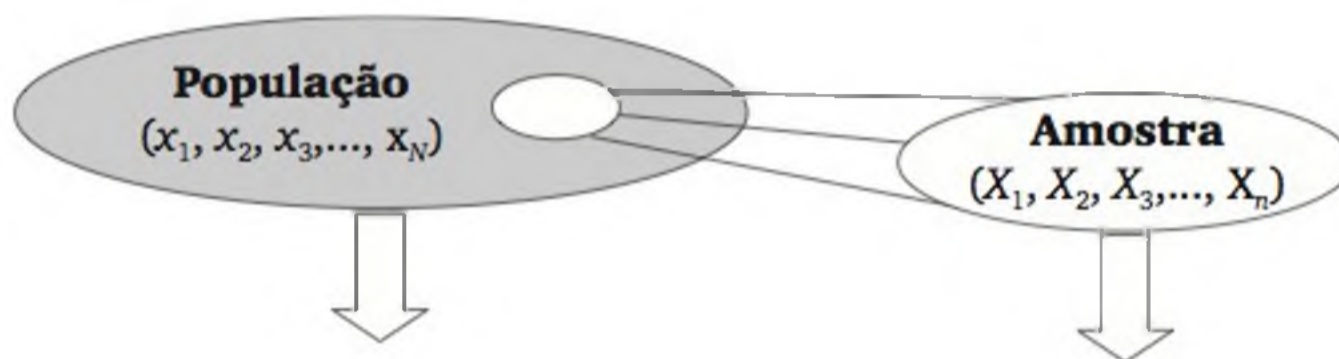
Em geral, estamos pesquisando uma ou mais variáveis associadas aos elementos da população ou da amostra. Nesse contexto, também podemos caracterizar a população e a amostra em termos da variável em estudo. Por exemplo, ao verificar se cada consumidor potencial de uma revendedora de automóveis planeja ($X = 1$) ou não ($X = 0$) comprar um carro novo no próximo ano, podemos representar a população pelo conjunto $\{x_1, x_2, x_3, \dots, x_N\}$, onde $x_j = 0$ ou 1 , dependendo se o j -ésimo indivíduo da população pretende ou não comprar um carro novo no próximo ano. Da mesma forma, a amostra, que será formada por n indivíduos a serem selecionados da população, pode ser representada por $\{X_1, X_2, \dots, X_n\}$, onde X_i é a variável aleatória que corresponde ao valor de X (0 ou 1) na i -ésima observação ($i = 1, 2, \dots, n$).

No presente contexto, definimos:

Parâmetro: alguma medida descritiva (média, variância, proporção etc.) dos valores x_1, x_2, x_3, \dots , associados à população.

Amostra aleatória simples: conjunto de n variáveis aleatórias independentes $\{X_1, X_2, \dots, X_n\}$, cada uma com a mesma distribuição de probabilidades de certa variável aleatória X . Essa distribuição de probabilidades deve corresponder à distribuição de frequências dos valores da população (x_1, x_2, x_3, \dots).

Estatística: alguma medida descritiva (média, variância, proporção etc.) das variáveis aleatórias X_1, X_2, \dots, X_n , associadas à amostra (ver a Figura 7.2).



	Parâmetros	Estatísticas
Proporção	$p = \frac{n^\circ \text{ de elementos com o atributo}}{N}$	$\hat{p} = \frac{n^\circ \text{ de elementos com o atributo}}{n}$
Média	$\mu = \frac{1}{N} \sum_{i=1}^N x_i$	$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$
Variância	$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$	$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$

Figura 7.2 Alguns parâmetros e estatísticas.

Estatísticas e variáveis aleatórias

Vamos considerar uma amostragem aleatória simples, o que faz com que qualquer medida associada à amostra (*estatística*) seja uma variável aleatória. Isso ocorre por causa da aleatoriedade introduzida pelo sorteio, na amostragem. A fim de entender melhor esse conceito, vamos acompanhar os exemplos seguintes.

Exemplo 7.1 Em um estudo sobre emissões de CO₂, definiu-se uma população como sendo composta por quatro ônibus de uma pequena companhia de transporte urbano. Dos quatro ônibus, um deles apresentava alto índice de emissão, enquanto os outros três estavam dentro dos padrões. Assim, a população pode ser descrita por {1, 0, 0, 0}. O parâmetro de interesse é a *proporção de veículos fora do padrão*. Considere as seguintes questões acerca dessa população:

- calcular a proporção populacional (p);
- se for retirada uma amostra aleatória simples, com reposição, de tamanho $n = 2$, qual será a proporção \hat{P} de veículos fora dos padrões na amostra?

Solução: Para a questão (a), a resposta é trivial: $p = 1/4$. Já a questão (b) não pode ser respondida, pois a proporção amostral (\hat{P}) é uma variável aleatória. Assim, não podemos dizer o que *vai* ocorrer, mas tão somente o que *pode* ocorrer.

- c) retificando a questão (b), construir a distribuição de probabilidades da proporção amostral.

Solução: Seja a variável aleatória $X = \text{número de ônibus com alto índice de emissão entre dois ônibus selecionados ao acaso}$. Então, X possui distribuição binomial, já que se trata de dois eventos independentes, com duas possibilidades cada (*sucesso = ônibus com alto índice de emissão; fracasso = ônibus dentro dos padrões*). Os parâmetros da distribuição são $n = 2$ e $p = 0,25$. Assim, pela distribuição binomial (ver a Tabela I do apêndice), temos:

x	$p(x)$
0	0,5625
1	0,3750
2	0,0625
Total	1

A proporção amostral é dada por:

$$\hat{p} = \frac{X}{n} = \frac{X}{2}$$

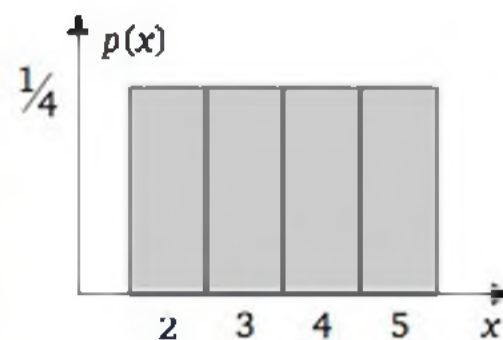
Assim, sua função de probabilidade é dada por:

\hat{p}	$p(\hat{p})$
0	0,5625
$\frac{1}{2}$	0,3750
1	0,0625
Total	1

A função de probabilidade de \hat{P} também é chamada de *distribuição da proporção amostral* ou *distribuição amostral da proporção*, pois apresenta os possíveis resultados de uma proporção, que é calculada sobre os elementos de uma amostra a ser extraída da população em estudo. De maneira geral, temos:

Uma *estatística* é uma variável aleatória e sua distribuição de probabilidades é chamada de *distribuição amostral*.

Exemplo 7.2 Seja a população dos quatro ônibus e a variável $X = \text{número de vezes que o ônibus teve um defeito grave}$. Se um ônibus teve dois defeitos graves, o outro três, o outro quatro e o último cinco defeitos graves, então a população, em termos da variável X , pode ser descrita pelo conjunto $\{2, 3, 4, 5\}$. A população também pode ser descrita pela função de probabilidade ao lado – a *distribuição da população*. Essa distribuição tem os parâmetros valor esperado (média) e variância dados por:



$$\mu = \frac{1}{N} \sum_{i=1}^N x_i = \frac{1}{4} (2 + 3 + 4 + 5) = 3,5 \text{ e}$$

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 = \frac{1}{4} [(2 - 3,5)^2 + (3 - 3,5)^2 + (4 - 3,5)^2 + (5 - 3,5)^2] = 1,25$$

A Tabela 7.1 mostra a construção da distribuição da média amostral, considerando uma amostragem aleatória simples com $n = 2$ elementos, extraída com reposição.

Tabela 7.1 Construção da distribuição de \bar{X} (Exemplo 7.2).

Amostras possíveis	Valor de \bar{X}	Probabilidade
(2, 2)	2,0	$\frac{1}{16}$
(2, 3), (3, 2)	2,5	$\frac{2}{16}$
(2, 4), (3, 3), (4, 2)	3,0	$\frac{3}{16}$
(2, 5), (3, 4), (4, 3), (5, 2)	3,5	$\frac{4}{16}$
(3, 5), (4, 4), (5, 3)	4,0	$\frac{3}{16}$
(4, 5), (5, 4)	4,5	$\frac{2}{16}$
(5, 5)	5,0	$\frac{1}{16}$

O valor esperado e a variância da distribuição de X são:

$$E(\bar{X}) = 2\left(\frac{1}{16}\right) + 2,5\left(\frac{2}{16}\right) + 3\left(\frac{3}{16}\right) + 3,5\left(\frac{4}{16}\right) + 4\left(\frac{3}{16}\right) + 4,5\left(\frac{2}{16}\right) + 5\left(\frac{1}{16}\right) = 3,5$$

$$V(\bar{X}) = (2 - 3,5)^2 \frac{1}{16} + (2,5 - 3,5)^2 \frac{2}{16} + \dots + (5 - 3,5)^2 \frac{1}{16} = 0,625$$

A Figura 7.3 mostra a distribuição da população e a distribuição da média amostral. Note que ambas têm a mesma média (valor esperado), mas a distribuição da média amostral é mais concentrada (menor variância) e tem forma mais parecida com a distribuição normal.

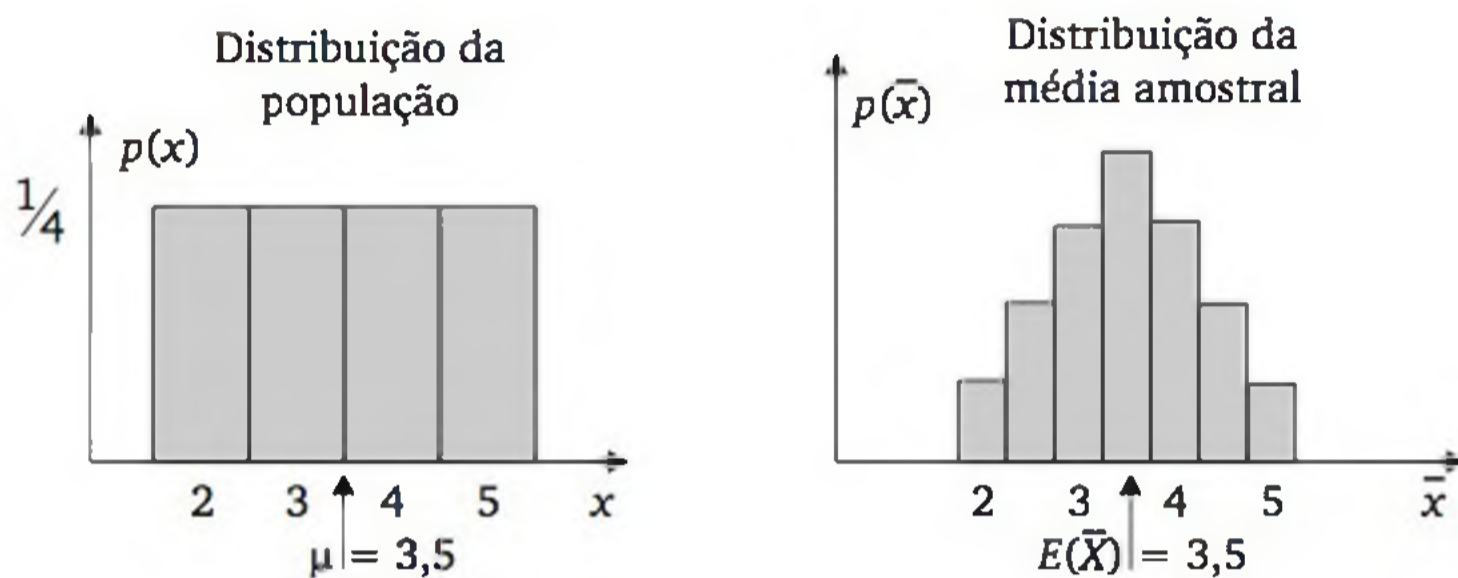


Figura 7.3 *Distribuição da população do Exemplo 7.2 e a distribuição da média amostral, considerando amostragem aleatória simples com $n = 2$ elementos, extraídos com reposição.*

EXERCÍCIOS

1. Refaça o Exemplo 7.1c, considerando que a amostra seja retirada sem reposição.
2. Em um estudo sobre consumo de combustível, definiu-se uma população composta por quatro ônibus de uma pequena companhia de transporte urbano. Os consumos dos ônibus (km/l), em condições padrões de teste, eram 3,8, 3,9, 4,0 e 4,1. Uma amostra de dois elementos será sorteada, com reposição. Verifique todas as amostras possíveis e, em seguida, construa a distribuição amostral para o *consumo médio* da amostra e calcule o valor esperado e a variância.
3. Refaça o exercício anterior considerando amostragem sem reposição.

7.2 DISTRIBUIÇÕES AMOSTRAIS

Quando a amostragem é aleatória simples, várias estatísticas apresentam distribuições amostrais que se aproximam de distribuições contínuas conhecidas, à medida que o tamanho da amostra cresce. É o caso da média e da proporção que apresentam distribuições amostrais aproximadamente normal.

7.2.1 Distribuição amostral da média

Considere o esquema ilustrado na Figura 7.4.

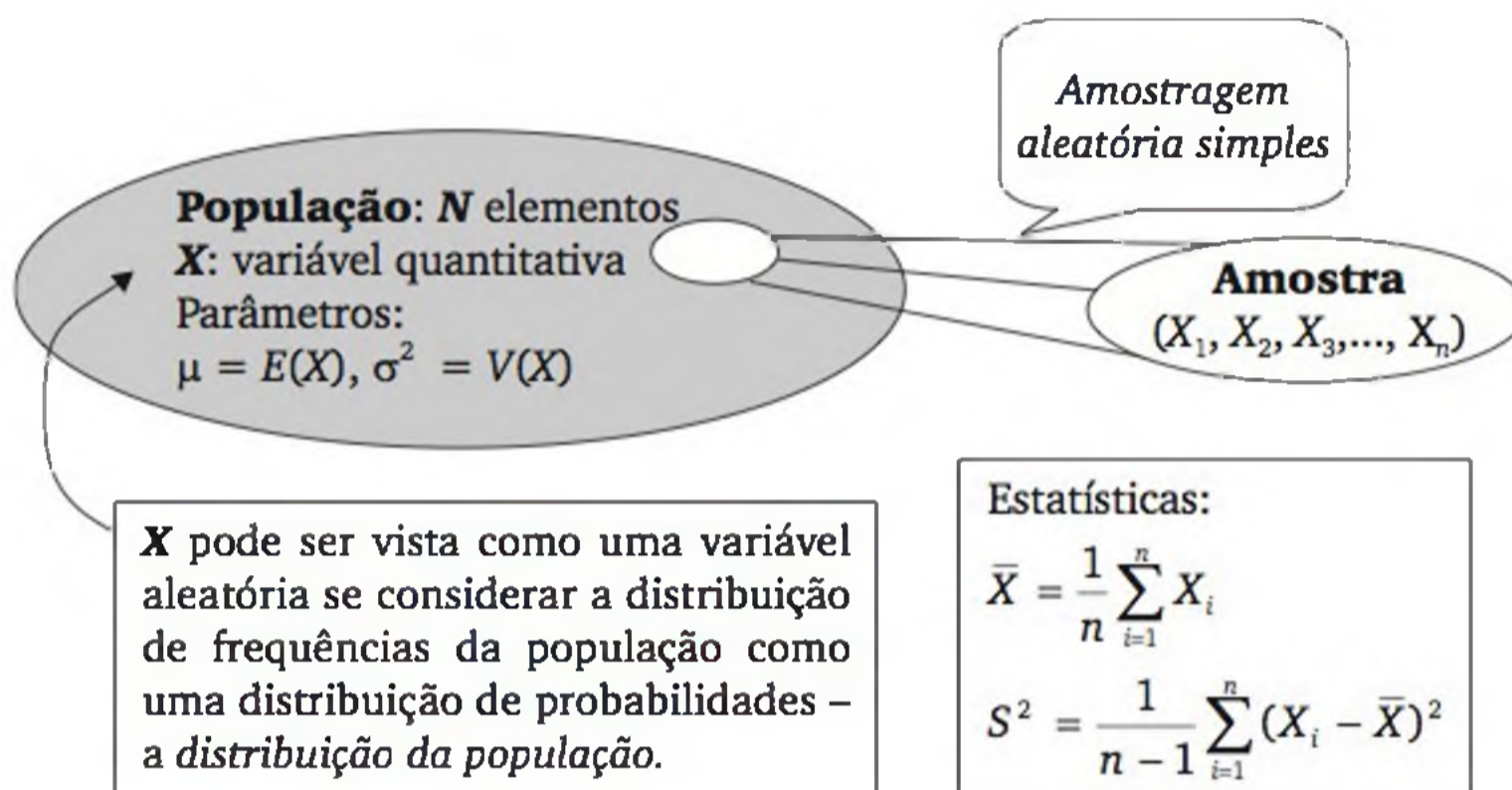


Figura 7.4 Esquema geral de uma amostragem aleatória simples na observação de uma variável quantitativa X .

Seja uma amostra aleatória simples $\{X_1, X_2, \dots, X_n\}$ e a estatística \bar{X} . A distribuição de \bar{X} (*distribuição da média amostral*) apresenta as seguintes propriedades:

- a) O valor esperado da média amostral é igual à média da população, ou seja:¹

$$E(\bar{X}) = \mu \quad (7.1)$$

- b) A variância da média amostral é inferior à variância populacional (σ^2) e a relação é dada por

$$V(\bar{X}) = \frac{\sigma^2}{n} \quad (\text{se a amostragem for com reposição, ou } N \text{ muito grande ou infinito}).^2 \quad (7.2)$$

1 $E(\bar{X}) = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{n\mu}{n} = \mu$

2 Com as condicionantes, podemos supor independência entre X_1, \dots, X_n , donde:

$$V(\bar{X}) = V\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n V(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

ou

$$V(\bar{X}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1} \quad (\text{se a amostragem for sem reposição e } N \text{ não muito grande, } N < 20n) \quad (7.3)$$

- c) (*Teorema do limite central*) Se o tamanho da amostra for razoavelmente grande, então a distribuição amostral da média pode ser aproximada pela *distribuição normal*. Em geral, para $n \geq 30$, a aproximação já é boa, porém, se a distribuição da população não for muito distante de uma normal, a aproximação pode ser usada com n menor.

Exemplo 7.2 (continuação) Dada a população {2, 3, 4, 5}, com parâmetros $\mu = 3,5$ e $\sigma^2 = 1,25$, e considerando o planejamento de uma amostra aleatória simples, com reposição, de $n = 2$ elementos, então podemos obter o valor esperado e a variância da média amostral usando (7.1) e (7.2):

$$E(\bar{X}) = \mu = 3,5 \text{ e}$$

$$V(\bar{X}) = \frac{\sigma^2}{n} = \frac{1,25}{2} = 0,625$$

Observe que são os mesmos valores encontrados anteriormente.

7.2.2 Distribuição amostral da proporção

Quando o interesse é estudar uma proporção, tal como a *proporção dos elementos que têm certo atributo A*, a população pode ser vista como dividida em dois subgrupos:

1. o subgrupo dos elementos que *têm* o atributo A; e
2. o subgrupo dos elementos que *não têm* o atributo A, como mostra a Figura 7.5.

População: $N = N_A + N_{\bar{A}}$ elementos

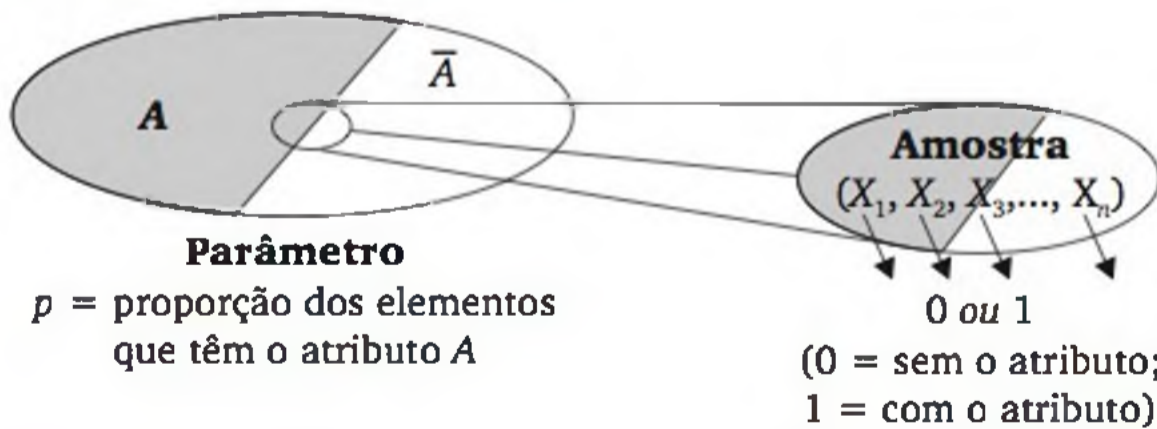


Figura 7.5 Esquema geral de uma amostragem aleatória simples quando se observa a proporção de certo atributo A.

A distribuição da população pode ser representada por uma variável aleatória de Bernoulli (tipo “0–1”), com função de probabilidade:

x	$p(x)$
0	$1 - p$
1	p

Como vimos no Capítulo 5, o valor esperado e a variância de uma distribuição desse tipo são dados, respectivamente, por:

$$\mu = p \text{ e} \quad (7.4)$$

$$\sigma^2 = p(1 - p) \quad (7.5)$$

Representando as observações amostradas por

$$X_i = \begin{cases} 1 & \text{se o } i\text{-ésimo elemento tem o atributo } A \\ 0 & \text{se o } i\text{-ésimo elemento não tem o atributo } A \end{cases}$$

verificamos que

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{\text{número de elementos com o atributo } A}{n} = \hat{p} \quad (7.6)$$

ou seja, a proporção equivale a uma média aritmética para dados de variáveis do tipo “0–1”. Assim, as propriedades da distribuição amostral da média também são aplicadas à distribuição amostral da proporção. Usando as notações próprias da proporção, temos:

- a) O valor esperado da proporção amostral é igual à proporção da população:

$$E(\hat{p}) = p \quad (7.7)$$

- b) A variância da proporção amostral é dada por

$$V(\hat{p}) = \frac{p(1-p)}{n} \quad (\text{se a amostragem for com reposição, ou } N \text{ muito grande ou infinito).} \quad (7.8)$$

ou

$$V(\hat{p}) = \frac{p(1-p)}{n} \cdot \frac{N-n}{N-1} \quad (\text{se a amostragem for sem reposição e } N \text{ não muito grande, } N < 20n) \quad (7.9)$$

- c) Se o tamanho da amostra for razoavelmente *grande*, então a distribuição amostral da proporção pode ser aproximada pela *distribuição normal*.³

As aplicações da distribuição da proporção amostral podem ser feitas no contexto da aproximação da distribuição normal à binomial, inclusive com a correção de continuidade (Seção 6.3.1).

EXERCÍCIOS

4. Uma fundição produz blocos para motor de caminhões. Os furos para as camisas devem ter diâmetro de 100 mm, com tolerância de 5 mm. Para verificar qual é o diâmetro *médio* no processo, a empresa vai retirar uma amostra com 36 blocos e medir os diâmetros de 36 furos (1 a cada bloco). Suponha que o desvio padrão (populacional) dos diâmetros seja conhecido e igual a 3 mm.
- Qual é o desvio padrão da distribuição da média amostral?
 - Qual é a probabilidade de a média amostral diferir da média populacional (desconhecida) em mais do que 0,5 mm (para mais ou para menos)?
 - Qual é a probabilidade de a média amostral diferir da média populacional (desconhecida) em mais do que 1 mm (para mais ou para menos)?
 - Se alguém afirmar que a média amostral não se distanciará da média populacional em mais do que 0,98 mm, qual é a probabilidade de essa pessoa acertar?
 - Se alguém afirmar que a média amostral não se distanciará da média populacional em mais do que 1,085 mm, qual é a probabilidade de essa pessoa errar?
5. Uma empresa fabricante de pastilhas para freios efetua um teste para controle de qualidade de seus produtos. Supondo que 1% das pastilhas fabricadas pelo processo atual apresenta desempenho deficiente quanto ao nível de desgaste, qual é a probabilidade, em uma amostra aleatória simples com 10.000 pastilhas, de serem encontradas 85 ou menos pastilhas com problemas?
6. Sabe-se que 50% dos edifícios construídos em uma grande cidade apresentam problemas estéticos relevantes em menos de cinco anos após a entrega

³ Observamos que a distribuição exata é a binomial (ou a hipergeométrica se a amostragem for feita de população pequena e sem reposição).

da obra. Considerando a seleção de uma amostra aleatória simples com 200 edifícios com cinco anos, qual é a probabilidade de menos de 90 deles apresentarem problemas estéticos relevantes (considerar que não tenha havido obras de reparo nos edifícios selecionados)?

7. Existem vários algoritmos computacionais que permitem gerar números aleatórios (ou, mais apropriadamente, *pseudo-aleatórios*) no intervalo $[0, 1]$, com distribuição uniforme. Considere a geração de 100 números $(X_1, X_2, \dots, X_{100})$ desta forma e seja \bar{X} a média aritmética simples desses 100 números.
- Qual é o valor esperado e a variância de X_1 ?
 - Qual é a probabilidade de X_1 assumir um valor no intervalo $[0,47, 0,53]$?
 - Qual é o valor esperado e a variância de \bar{X} ?
 - Qual é a distribuição de probabilidade de \bar{X} ?
 - Qual é a probabilidade de \bar{X} assumir um valor no intervalo $[0,47, 0,53]$?
8. Um profissional de Computação observou que seu sistema gasta entre 20 e 24 segundos para realizar determinada tarefa. Além disso, o tempo gasto, X , pode ser razoavelmente representado pela seguinte função de densidade:

$$f(x) = \begin{cases} \frac{x}{4} - 5, & \text{para } 20 \leq x < 22 \\ 6 - \frac{x}{4}, & \text{para } 22 \leq x < 24 \\ 0, & \text{para } x \notin [20, 24] \end{cases}$$

- Numa particular rodada, qual é a probabilidade de o sistema gastar mais que 22,4 segundos?
- Em 30 rodadas, qual é a probabilidade de o sistema gastar, em média, mais que 22,4 segundos por rodada?

7.3 ESTIMAÇÃO DE PARÂMETROS

Nesta seção, estudaremos o problema de avaliar parâmetros populacionais, a partir de operações com os dados de uma amostra. É um raciocínio tipicamente indutivo, em que se generalizam resultados *da parte* (amostra) para o *todo* (população), conforme ilustra a Figura 7.6.

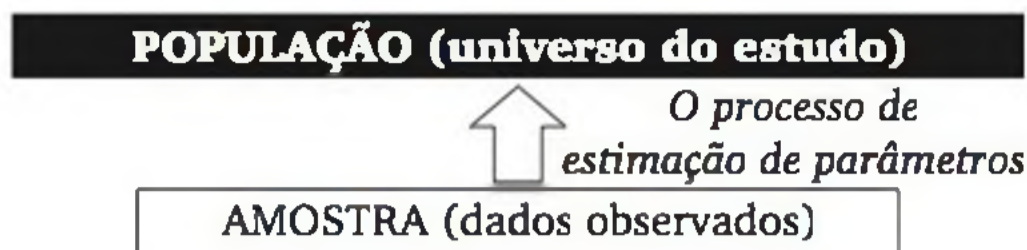


Figura 7.6 O raciocínio indutivo da estimação.

Por exemplo, podemos ter interesse em avaliar a resistência mecânica X de um novo material. Contudo, X não é um número, mas uma variável aleatória, porque há uma infinidade de fatores não controláveis que provocarão variações nas possíveis medidas de resistência mecânica do material. É até razoável admitir que a distribuição de X seja aproximadamente normal, por se tratar de medidas físicas. E o interesse pode estar na avaliação dos *parâmetros populacionais* $\mu = E(X)$ e $\sigma^2 = V(X)$.

Medidas de resistência mecânica (X_1, X_2, \dots, X_n) , a serem realizadas de forma independente e sob as mesmas condições, constituem uma *amostra aleatória simples* de X . Cálculos podem ser feitos sobre essas medidas para *estimar* os parâmetros de interesse. Exemplos desses cálculos podem ser

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{e} \quad (7.10)$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad (7.11)$$

que são *estimadores* dos parâmetros μ e σ^2 , respectivamente.

De forma genérica, considere uma população caracterizada pela distribuição de certa variável aleatória X , com parâmetro θ . E seja (X_1, X_2, \dots, X_n) uma amostra aleatória simples de X .

Uma *estatística* T é uma função dos elementos da amostra, isto é $T = f(X_1, X_2, \dots, X_n)$. Quando ela é usada para avaliar certo parâmetro θ , é também chamada de **estimador** de θ .

Observe que um estimador é uma variável aleatória, pois depende da amostra a ser selecionada. Realizada a amostragem, o estimador assume determinado valor (o resultado do cálculo), o qual denominamos de *estimativa*.⁴

4 Neste texto, as estimativas serão representadas por letras minúsculas, contrastando com os estimadores, os quais serão representados por letras maiúsculas.

Algumas propriedades desejáveis de um estimador serão discutidas, o que permite, pelo menos em tese, escolher o *melhor* estimador para cada situação prática. Usando as distribuições amostrais, é possível avaliar probabilisticamente o erro que se está cometendo por se usar uma amostra e não toda a população – o *erro amostral*. Isso será feito com estimativas em forma de *intervalos de confiança* (Seções 7.3.2 e 7.3.3).

7.3.1 Propriedades de um estimador

Um estimador, por ser uma variável aleatória, pode assumir valores, segundo uma distribuição de probabilidades. Contudo, é desejável que, em média, ele seja igual ao parâmetro que se deseja estimar. Mais formalmente,

T é um estimador **não-viesado** (ou **não tendencioso**) de um parâmetro θ se e só se $E(T) = \theta$.

Por exemplo, \bar{X} e \hat{P} são estimadores *não viesados* dos parâmetros μ e p , respectivamente, porque $E(\bar{X}) = \mu$ e $E(\hat{P}) = p$, conforme foi visto na Seção 7.2. Já o estimador

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^n (X_i - \bar{X})^2 \quad (7.12)$$

é um estimador *viesado* do parâmetro σ^2 , pois $E(\hat{\sigma}^2) = \frac{n-1}{n} \sigma^2$. Por isso, definimos a variância amostral, S^2 , com denominador $(n-1)$ no lugar de n . A diferença

$$E(\hat{\sigma}^2) - \sigma^2 = -\frac{1}{n} \sigma^2 \quad (7.13)$$

é chamada de *viés* do estimador $\hat{\sigma}^2$.

Na prática, retiramos só uma amostra, produzindo um único valor para o estimador – uma *estimativa*. Mesmo que o estimador seja não viesado, o valor da estimativa pode estar longe do valor do parâmetro. Outra propriedade desejável é que o estimador tenha variância pequena, porque isso reduz a chance de a estimativa acusar um valor distante do parâmetro.

Dados dois estimadores não viesados T_1 e T_2 , sendo $V(T_1) < V(T_2)$, então T_1 é dito mais *eficiente* do que T_2 . E a *eficiência relativa* de T_1 em relação a T_2 é dada por:

$$ef(T_1, T_2) = \frac{V(T_2)}{V(T_1)} \tag{6.14}$$

A Figura 7.7 ilustra os conceitos de viés e eficiência, fazendo analogia com tiros ao alvo, realizados por três rifles. Os rifles T_1 e T_2 são *não viesados*, porque, em média, acertam o alvo; enquanto o rifle T_3 é *viesado*. Embora T_1 e T_2 sejam não viesados, T_1 é mais *eficiente* do que T_2 , pois a variância entre os tiros é menor.

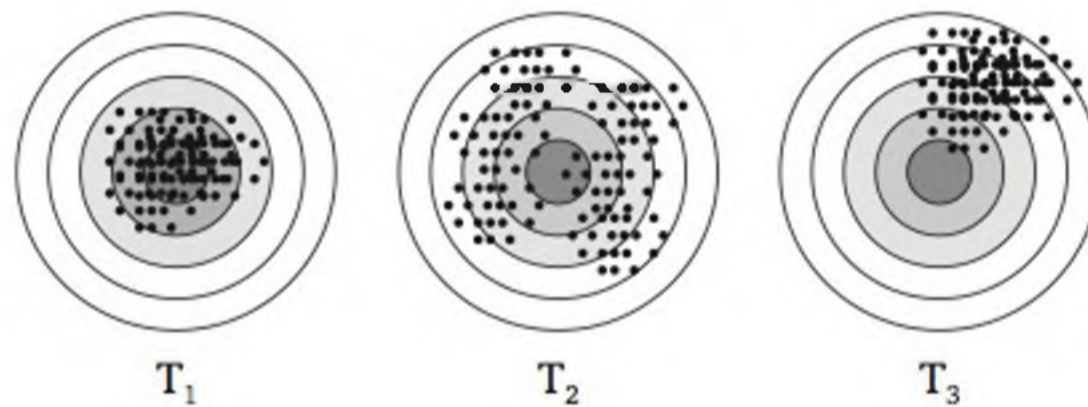


Figura 7.7 Tiros ao alvo com três rifles.

Considere a média amostral (\bar{X}) e a mediana amostral (M_d) como estimadores do parâmetro μ . Sabemos que $V(\bar{X}) = \frac{\sigma^2}{n}$. Se supusermos a população com distribuição normal e a amostra grande, é possível mostrar que $V(M_d) \approx \frac{\pi}{2} \cdot \frac{\sigma^2}{n}$. Como $V(\bar{X}) < V(M_d)$, então \bar{X} é um estimador *mais eficiente* do que M_d na estimação de μ , nas condições estabelecidas. E a *eficiência relativa* de \bar{X} em relação à M_d é dada por

$$\frac{V(M_d)}{V(\bar{X})} \approx \frac{\pi\sigma^2/2n}{\sigma^2/n} \approx 1,57 \tag{7.15}$$

Assim, em amostras grandes de populações normais, a média amostral é cerca de 57% mais eficiente do que a mediana amostral. Isso significa que se formos usar M_d no lugar de \bar{X} , precisamos ter uma amostra 57% maior, para garantir a mesma eficiência na estimação de μ .

De modo geral, a qualidade de um estimador T , na estimação de um parâmetro θ , pode ser avaliada em função de seu *erro quadrático médio*, o qual é definido por

$$EQM(T) = E(T - \theta)^2 \quad (7.16)$$

que pode ser escrito como⁵

$$EQM(T) = V(T) + (\text{viés})^2 \quad (7.17)$$

Então, para um estimador T não viesado, temos $EQM(T) = V(T)$.

Para dois estimadores T_1 e T_2 quaisquer (não necessariamente não viesados), definimos a *eficiência relativa* de T_1 em relação à T_2 por:

$$ef(T_1, T_2) = \frac{EQM(T_2)}{EQM(T_1)} \quad (7.18)$$

7.3.2 Intervalo de confiança para proporção

Em muitas situações, o principal parâmetro de interesse é alguma proporção p . Por exemplo:

- a proporção de itens defeituosos em uma linha de produção;
- a proporção de consumidores que vão comprar certo produto;
- a proporção de mensagens que chegam adequadamente a seu destino etc.

Seja uma população caracterizada por uma variável aleatória X , que assume o valor 0 ou 1, conforme o elemento tenha ou não certo atributo de interesse. Por exemplo, nas peças que saem de uma linha de produção, o código 0 pode identificar peça boa e o código 1 peça defeituosa. Para um elemento tomado ao acaso, seja $p = P(X = 1)$. Note que p representa a proporção de elementos com o atributo, na população.

Já vimos que a proporção amostral \hat{P} é um bom estimador da proporção populacional p . Dada uma amostra aleatória simples de tamanho n , o que se pode dizer sobre o *erro amostral*: $|\hat{P} - p|$?

Conforme discutido na Seção 7.2, se n for grande, a distribuição amostral de \hat{P} é aproximadamente normal com⁶

⁵ $EQM(T) = E\{T - E(T) + E(T) - \theta\}^2 = E\{T - E(T)\}^2 + \{E(T) - \theta\}^2 + 2E\{T - E(T)\}\{E(T) - \theta\} = E\{T - E(T)\}^2 + \{E(T) - \theta\}^2 + 0 = V(T) + (\text{viés})^2$.

⁶ Estamos supondo população infinita ou bastante grande. Caso contrário, deveríamos usar a correção (7.9), como discutido na Seção 7.2.

$$E(\hat{P}) = p \text{ e} \quad (7.19)$$

$$V(\hat{P}) = \frac{p(1-p)}{n} \quad (7.20)$$

Denotaremos por

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \quad (7.21)$$

o desvio padrão da distribuição amostral de \hat{P} , que no presente contexto será chamado de **erro padrão de \hat{P}** .

Seja uma variável aleatória normal padrão, Z . Como vale a relação⁷

$$P\{-1,96 \leq Z \leq 1,96\} = 0,95 \quad (7.22)$$

podemos escrever

$$P\left\{-1,96 \leq \frac{\hat{P} - p}{\sigma_{\hat{p}}} \leq 1,96\right\} = 0,95 \quad (7.23)$$

ou

$$P\{\hat{P} - (1,96)\sigma_{\hat{p}} \leq p \leq \hat{P} + (1,96)\sigma_{\hat{p}}\} = 0,95 \quad (7.24)$$

ou seja, com probabilidade de 95%, temos:

$$|\hat{P} - p| \leq (1,96)\sigma_{\hat{p}} \quad (7.25)$$

Observada efetivamente a amostra, e chamando de \hat{p} a proporção obtida nesta amostra, podemos definir um *intervalo de confiança* para p , com *nível de confiança* de 95%, por

$$IC(p, 95\%) = \hat{p} \pm (1,96)\sigma_{\hat{p}} \quad (7.26)$$

Na prática, $\sigma_{\hat{p}}$ não pode ser calculado por (7.21), porque depende do parâmetro desconhecido p . Então, usamos em seu lugar a estimativa:

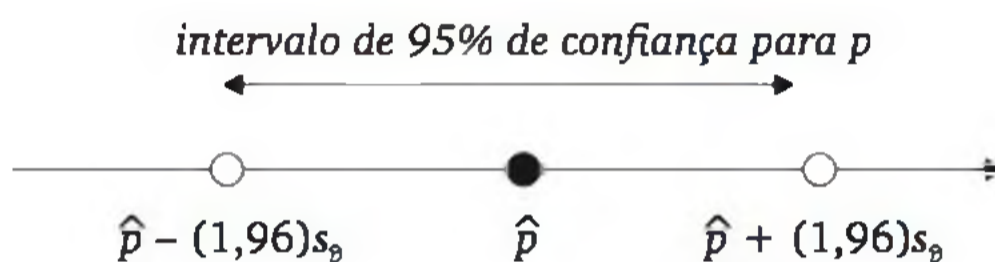
⁷ Ver Exemplo 6.5 (Capítulo 6).

$$s_{\hat{p}} = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \tag{7.27}$$

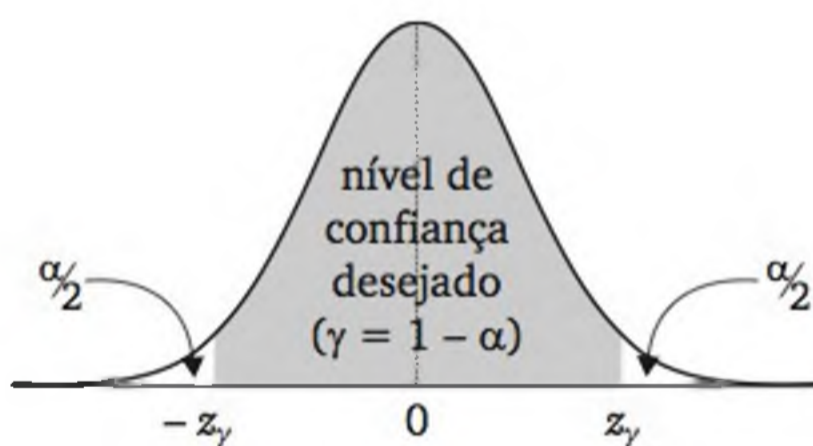
Desde que a amostra seja grande (p. ex., $n \geq 50$), a diferença entre $s_{\hat{p}}$ e $\sigma_{\hat{p}}$ pode ser considerada desprezível, e um *intervalo de confiança* para p , com *nível de confiança* de 95%, pode ser calculado por:

$$IC(p, 95\%) = \hat{p} \pm (1,96) \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \tag{7.28}$$

Esquemáticamente,



Em suma, embora p seja um parâmetro populacional desconhecido, é possível, com base em uma amostra aleatória simples, construir um intervalo que deve conter p com alto nível de confiança. É bastante usual o nível de confiança de 95%, mas o intervalo pode ser construído com um nível γ qualquer, bastando encontrar o valor de z_γ na distribuição normal padrão, conforme mostra a Figura 7.8.



γ	0,800	0,900	0,950	0,980	0,990	0,995	0,998
z_γ	1,282	1,645	1,960	2,326	2,576	2,807	3,090

Figura 7.8 Valores de z_γ para alguns níveis de confiança.

Com z_γ tomado adequadamente, conforme o esquema da Figura 7.8, calculamos o intervalo de confiança para p por:

$$IC(p, \gamma) = \hat{p} \pm z_\gamma \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \quad (7.29)$$

Exemplo 7.3 Na avaliação de dois sistemas computacionais, A e B, foram selecionadas 400 cargas de trabalho (tarefas) – supostamente uma amostra aleatória da infinidade de cargas de trabalho que poderiam ser submetidas a esses sistemas. O sistema A foi melhor que o B em 60% dos casos. Construir intervalos de confiança para p (proporção de vezes que o sistema A é melhor que o sistema B, considerando todas as possíveis cargas de trabalho) usando níveis de confiança de 95% e 99%.

Para o nível de confiança de 95%, temos $z_\gamma = 1,96$, resultando em

$$\begin{aligned} IC(p, 95\%) &= \hat{p} \pm (1,96) \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} = 0,6 \pm (1,96) \sqrt{\frac{0,6(0,4)}{400}} = \\ &= 0,600 \pm 0,048 \end{aligned}$$

ou, em porcentagens:

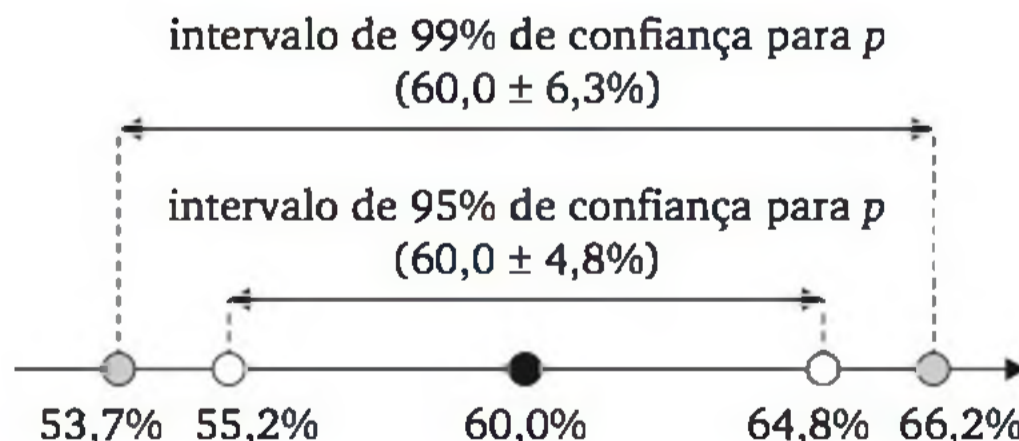
$$IC(p, 95\%) = 60,0\% \pm 4,8\%$$

Concluimos, então, que o intervalo (55,2%; 64,8%) contém o parâmetro p , com nível de confiança de 95%.

Para o nível de confiança de 99%, temos $z_\gamma = 2,576$, resultando em

$$IC(p, 99\%) = 0,6 \pm (2,576) \sqrt{\frac{0,6(0,4)}{400}} = 0,600 \pm 0,063$$

Ou seja, o intervalo (53,7%; 66,2%) contém o parâmetro p , com nível de confiança de 99%. Esquemáticamente:



Observe que, ao exigir maior nível de confiança, o intervalo de confiança aumenta em magnitude. Tente entender o porquê disso! Para *dado nível de confiança*, dizemos que uma estimativa é tão mais *precisa* quanto menor for a am-

plitude de seu intervalo de confiança. A forma natural de aumentar a precisão é aumentando o tamanho da amostra. Voltaremos a esse ponto na Seção 7.4.

7.3.3 Intervalo de confiança para média

Seja uma população caracterizada pela distribuição de uma variável aleatória X com os seguintes parâmetros: $E(X) = \mu$ e $V(X) = \sigma^2$. Por exemplo, X pode representar a mensuração da resistência mecânica de um novo material. Devido ao erro experimental onipresente, X é uma variável aleatória; assim, o interesse recai em seu valor esperado μ .

Considere uma amostra aleatória simples $\{X_1, X_2, \dots, X_n\}$ de X . Supondo X com distribuição aproximadamente normal, então

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad (7.30)$$

é o estimador natural de μ . Vimos na Seção 7.2 que \bar{X} tem distribuição aproximadamente normal com média e variância dadas por⁸

$$E(\bar{X}) = \mu \quad (7.31)$$

$$V(\bar{X}) = \frac{\sigma^2}{n} \quad (7.32)$$

O desvio padrão da distribuição amostral de \bar{X} ,

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \quad (7.33)$$

será chamado de **erro padrão** de \bar{X} .

Escolhendo z_γ em função do nível de confiança γ desejado, tal que $P\{-z_\gamma \leq Z \leq z_\gamma\} = \gamma$, podemos escrever

$$P\left\{-z_\gamma \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_\gamma\right\} = \gamma \quad (7.34)$$

ou

⁸ Novamente, estamos supondo a população infinita ou muito grande. Caso contrário, deveríamos usar a correção (7.3), como visto na Seção 7.2.

$$P\left\{\bar{X} - z_\gamma \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_\gamma \frac{\sigma}{\sqrt{n}}\right\} = \gamma \quad (7.35)$$

Observada efetivamente a amostra, e chamando de \bar{x} a média aritmética dos dados, podemos definir um intervalo de confiança para μ , com nível de confiança γ , por:

$$IC(\mu, \gamma) = \bar{x} \pm z_\gamma \frac{\sigma}{\sqrt{n}} \quad (7.36)$$

Exemplo 7.4 Em uma indústria de cerveja, a quantidade de cerveja inserida em latas tem-se comportado como uma variável aleatória com média 350 ml e desvio padrão 3 ml. Após alguns problemas na linha de produção, suspeita-se que houve alteração na média. Uma amostra de 20 latas acusou média $\bar{x} = 346$ ml. Construa um intervalo de confiança para o novo valor da quantidade média μ de cerveja inserida em latas, com nível de confiança 95%, supondo que não tenha ocorrido alteração no desvio padrão do processo.

Solução:

$$IC(\mu, 95\%) = \bar{x} \pm (1,96) \frac{\sigma}{\sqrt{n}} = 346 \pm (1,96) \frac{3}{\sqrt{20}} = 346 \pm 1,31 \text{ ml}$$

Interpretando: a quantidade média μ de cerveja inserida em latas, após os problemas na linha de produção, é 346 ml, tolerando, com 95% de confiança, uma margem de erro de até 1,31 ml. Assim, o intervalo (344,69; 347,31) contém, com 95% de confiança, o valor μ . Isso mostra que estatisticamente houve alteração na média do processo, pois o valor da média antiga (350 ml) não pertence ao intervalo.

Desvio padrão desconhecido

O intervalo de confiança descrito em (7.36) somente poderá ser usado nas situações em que conhecemos o desvio padrão σ da população, o que não é comum na prática. Caso contrário, o procedimento usual é substituir σ pelo desvio padrão calculado com os dados da amostra:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} = \sqrt{\frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right)} \quad (7.37)$$

Duas situações a considerar:

1. quando a amostra for grande (digamos, $n \geq 50$), a diferença entre σ e s pode ser desprezível, permitindo ainda o uso de (7.36);
2. quando a amostra for pequena, é necessário efetuar uma correção, como veremos a seguir.

A distribuição *t* de Student

Supondo a população com distribuição normal, a estatística

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \tag{7.38}$$

tem distribuição de probabilidades conhecida como *distribuição t de Student*, com $gl = n - 1$ graus de liberdade.

A *distribuição t de Student*, como mostra a Figura 7.9, tem forma parecida com a normal padrão, mas é um pouco mais dispersa. Essa dispersão varia com o tamanho da amostra. É bastante dispersa para amostras pequenas, mas se aproxima da normal padrão para amostras grandes.

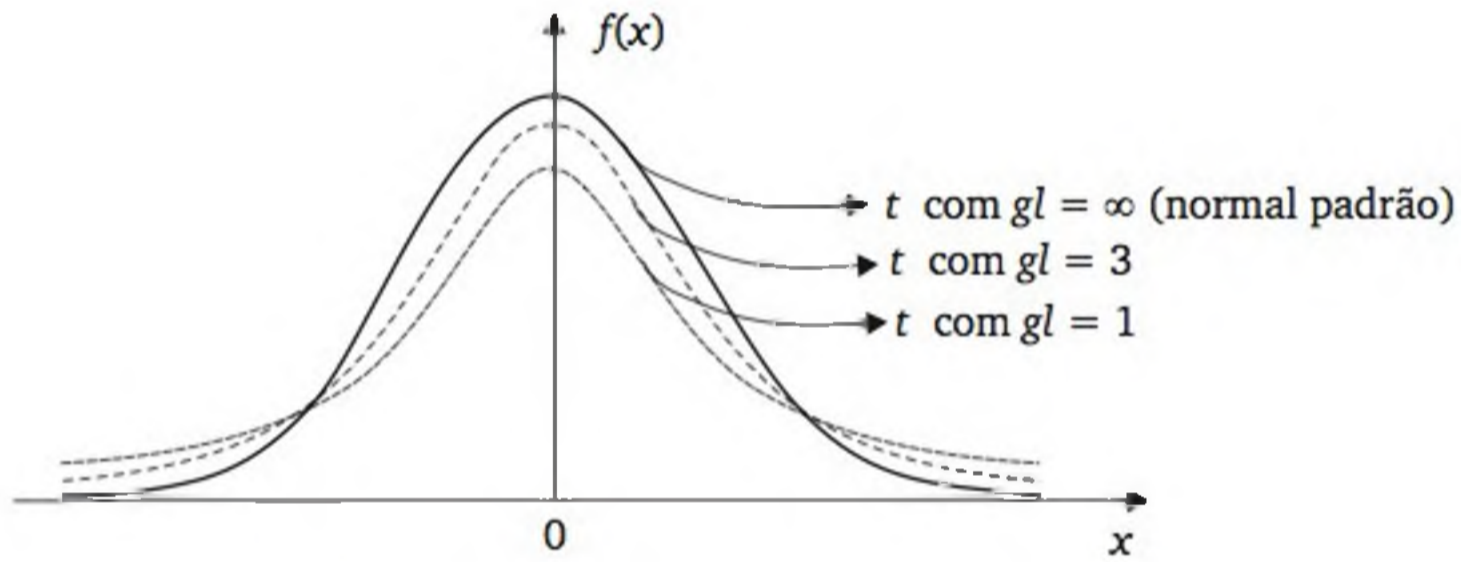


Figura 7.9 Gráficos de distribuições *t* de Student e da normal padrão.

Dado um nível de confiança γ , podemos obter o valor t_γ da distribuição *t* de Student, usando a Tabela 4 do apêndice, na linha correspondente a $gl = n - 1$. A Figura 7.10 ilustra esse processo, com $gl = 9$ e nível de confiança de 95%.

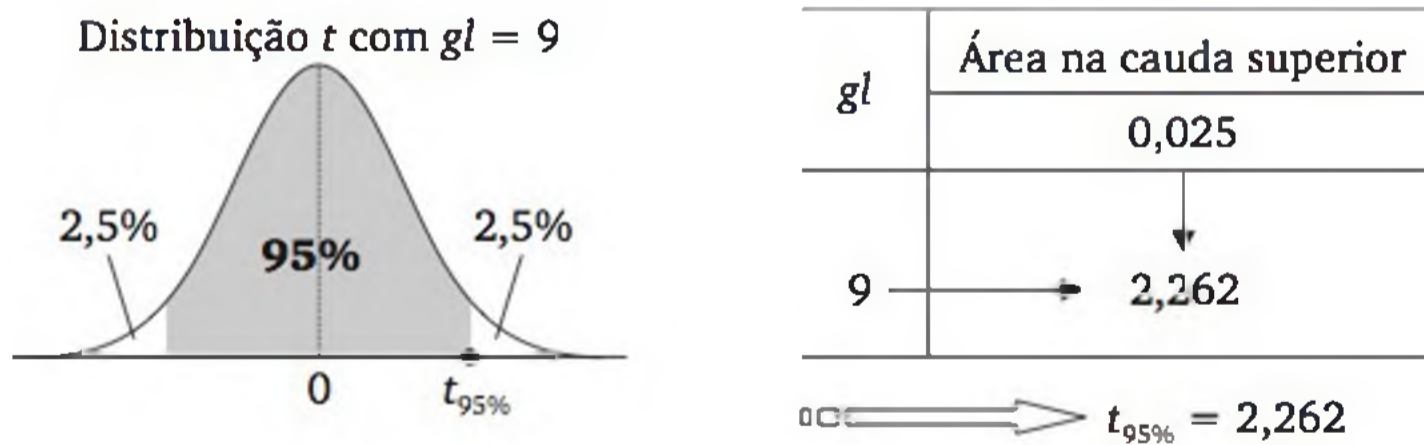


Figura 7.10 Uso da tabela da distribuição *t* de Student: ilustração com $gl = 9$ e nível de confiança de 95%.

Intervalo de confiança para μ com uso da distribuição t de Student

Usando a Tabela 4 com $gl = n - 1$, podemos escolher o valor t_γ em função do nível de confiança γ desejado, tal que

$$P\left\{-t_\gamma \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_\gamma\right\} = \gamma \quad (7.39)$$

ou

$$P\left\{\bar{X} - t_\gamma \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_\gamma \frac{S}{\sqrt{n}}\right\} = \gamma \quad (7.40)$$

Assim, o intervalo de confiança para μ , com a amostra efetivamente observada, é dado por⁹

$$IC(\mu, \gamma) = \bar{x} \pm t_\gamma \frac{s}{\sqrt{n}} \quad (7.41)$$

Exemplo 7.5 Deseja-se avaliar a dureza esperada μ do aço produzido sob um novo processo de têmpera. Uma amostra de dez corpos de prova do aço produziu os seguintes resultados de dureza, em HRc:

36,4 35,7 37,2 36,5 34,9 35,2 36,3 35,8 36,6 36,9

Construir um intervalo de confiança para μ , com nível de confiança de 95%.

Calculando as estatísticas para a amostra observada, temos:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = 36,15$$

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} = \sqrt{\frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right)} = 0,7352$$

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = 0,2325$$

⁹ Para amostras pequenas, a validade do $IC(\mu, \gamma)$ está condicionada à suposição de que os dados provenham de uma distribuição aproximadamente normal. Para $n \geq 30$, o teorema limite central garante a validade de $IC(\mu, \gamma)$ e, além disso, tem-se para n grande: $t_\gamma \approx z_\gamma$.

Como vimos na Figura 7.10, usando nível de confiança $\gamma = 95\%$, temos, pela Tabela 4 com $gl = 9$, o valor $t_{95\%} = 2,262$, resultando em:

$$IC(\mu, 95\%) = \bar{x} \pm t_{95\%} \frac{s}{\sqrt{n}} = 36,15 \pm 0,53$$

Ou seja, a resistência mecânica esperada do aço produzido pelo novo processo de têmpera é 36,15 HRc, tolerando, com 95% de confiança, uma margem de erro de até 0,53 HRc.

EXERCÍCIOS

9. Sejam X_1, X_2, \dots, X_7 uma amostra aleatória simples de uma população com média μ e desvio padrão σ . Considere os seguintes estimadores de μ :
- $$T_1 = (X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7) / 7$$
- $$T_2 = (X_2 + X_3 + X_4 + X_5 + X_6) / 5$$
- $$T_3 = (X_2 + X_3 + X_4 + X_5 + X_6) / 7$$
- a) Quais estimadores são não viciados? Justifique.
- b) Qual estimador é o mais eficiente entre T_1 e T_2 ? Justifique.
10. Em uma amostra aleatória simples com 200 edifícios com cinco anos, em certa cidade, 55% apresentaram problemas estéticos relevantes após a entrega da obra. Construir um intervalo de confiança para a proporção de edifícios da cidade que apresentam problemas estéticos relevantes nos cinco primeiros anos. Use nível de confiança de 95%.
11. Uma empresa fabricante de pastilhas para freios efetua um teste para controle de qualidade de seus produtos. Selecionou-se uma amostra de 600 pastilhas, das quais 18 apresentaram níveis de desgaste acima do tolerado. Construir um intervalo de confiança para a proporção de pastilhas com desgaste acima do tolerado, do atual processo industrial, com nível de confiança de 95%. Interpretar o resultado.
12. Uma fundição produz blocos para motor de caminhões. Os blocos têm furos para as camisas e deseja-se verificar qual é o diâmetro *médio* no processo do furo. A empresa retirou uma amostra de 36 blocos e mediu os diâmetros de 36 furos (1 a cada bloco). A amostra acusou média de 98,0 mm e desvio padrão de 4,0 mm. Construir um intervalo de confiança para a média do processo, com nível de confiança de 99%. Interpretar o resultado. Se o processo deveria ter média 100 mm, há evidência (com 99% de confiança) de que a média do processo não está no valor ideal? Explique.

7.4 TAMANHO DE AMOSTRA

Na seção anterior, aprendemos como estimar um parâmetro mediante observação de uma amostra aleatória simples de tamanho n . A estimação é feita com certa precisão, no sentido de que também avaliamos o *erro amostral* que podemos estar cometendo.

Contudo, ainda na fase do planejamento da pesquisa, muitas vezes precisamos calcular o tamanho n da amostra, para garantir certa precisão desejada, que é descrita em termos do *erro amostral máximo tolerado* (E_0) e do nível de confiança (γ) a ser adotado no processo de estimação. No caso de estimação de μ , podemos exigir

$$|\bar{X} - \mu| \leq E_0 \quad (7.42)$$

Usando (7.34), temos:

$$z_\gamma \frac{\sigma}{\sqrt{n}} \leq E_0 \quad (7.43)$$

Isolando n , temos:

$$n \geq \frac{z_\gamma^2 \sigma^2}{E_0^2} \quad (7.44)$$

O tamanho mínimo da amostra é o menor n que satisfaz a inequação precedente.

A dificuldade operacional para calcularmos o tamanho da amostra é que o cálculo depende da variância populacional σ^2 , que em geral é desconhecida. Em alguns problemas, σ^2 pode ser avaliada por meio de estudos anteriores ou pela experiência do engenheiro; em outras situações, σ^2 é obtida de uma *amostragem piloto*, isto é, alguns elementos da população são examinados e a variância encontrada nesta amostra piloto é usada no lugar de σ^2 . Nesse caso, é melhor usar a expressão (7.44) com t_γ no lugar de z_γ .

Exemplo 7.5 (continuação) Considere que o pesquisador julgou o resultado encontrado, $IC(\mu, 95\%) = 36,15 \pm 0,53$, pouco preciso. Ele tolera um erro amostral máximo de 0,3 HRc. Além disso, ele quer realizar as estimações com nível de confiança de 99%. Qual deve ser o tamanho da amostra?

Solução: Para efetuar o cálculo, é necessário o conhecimento da variância populacional, σ^2 . Usaremos, em seu lugar, a variância calculada sobre as dez observações, isto é $s^2 = (0,7352)^2 \approx 0,54$. Logo,

$$n \geq \frac{t_{99\%}^2 \sigma^2}{E_0^2} = \frac{(3,250)^2 (0,54)}{(0,3)^2} = 63,375$$

Portanto, precisamos de $n = 64$ corpos de prova para satisfazer a precisão desejada.

O Quadro 7.1 apresenta o formulário para o cálculo do tamanho da amostra, em função do parâmetro desejado. Quando a população é infinita ou muito grande, n_0 já é o tamanho da amostra. Porém, se N é conhecido e, especialmente, se não for muito grande, devemos dar sequência ao cálculo, conforme indica a última linha do Quadro 7.1.

Quadro 7.1 *Tamanho mínimo de uma amostra aleatória simples.*

Parâmetro de interesse	Valor inicial do tamanho da amostra
a) uma média (μ):	$n_0 = \frac{z_\gamma^2 \sigma^2}{E_0^2}$ (7.45)
b) uma proporção (p):	$n_0 = \frac{z_\gamma^2 p(1-p)}{E_0^2}$ (7.46)
c) várias proporções (p_1, p_2, \dots):	$n_0 = \frac{z_\gamma^2}{4E_0^2}$ (7.47)
Tamanho da amostra	
População infinita:	$n = n_0$ (arredondamento para o inteiro superior) (7.48)
População de tamanho N :	$n = \frac{N \cdot n_0}{N + n_0 - 1}$ (arredondamento para o inteiro superior) (7.49)

Quando o objetivo é estimar uma proporção p ($0 < p < 1$), podemos usar a seguinte relação (ver a Figura 7.11):

$$\sigma^2 = p \cdot (1 - p) \leq \frac{1}{4} \quad (7.50)$$

A primeira parte da relação (7.50) foi usada para construir a expressão de n_0 em (7.46) (Quadro 7.1b); e a segunda em (7.47) (Quadro 7.1c).

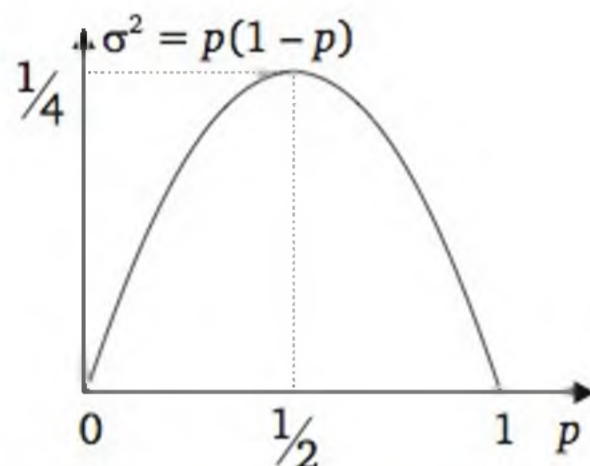


Figura 7.11 O parâmetro σ^2 em função do valor da proporção p .

A expressão (7.47) (Quadro 7.1c) também é usada com o objetivo de estimar uma única proporção p , quando não temos informação *a priori* sobre o valor de p . Contudo, deve resultar em um valor de n maior do que o necessário.

No caso de estarmos usando nível de confiança de 95%, temos $z_\gamma = 1,96 \approx 2$, fazendo com que (7.47) resulte em:

$$n_0 = \frac{1}{E_0^2} \quad (7.51)$$

A expressão (7.51) é muito usada no planejamento de pesquisas de levantamento, com o objetivo de estimar várias proporções, como nos exemplos seguintes:

- numa pesquisa eleitoral, em que é comum a necessidade de avaliar a proporção de cada candidato;
- em pesquisas de mercado, em que normalmente desejam-se avaliar as proporções de vários atributos nos consumidores;
- no levantamento de arquivos que trafegam em uma rede, em que é comum o interesse em verificar a proporção de cada tipo de arquivo.

EXERCÍCIOS

- 13.** Um pesquisador precisa determinar o tempo médio gasto para perfurar três orifícios em uma peça de metal. Qual deve ser o tamanho da amostra para que a média amostral esteja a menos de 15 s da média populacional? Por experiência prévia, pode-se supor o desvio padrão em torno de 40 s. Considere também, que a estimação será realizada com nível de confiança de 95%.
- 14.** Seja a construção de um plano para garantir a qualidade dos parafusos vendidos em caixas com 100 unidades. Um dos requisitos é controlar o comprimento médio dos parafusos. Quer-se saber quantos parafusos de-

ve-se examinar em cada caixa, para garantir que a média da amostra (\bar{x}) não difira do comprimento médio dos parafusos da caixa (μ) em mais que 0,8 mm. Considere que a estimação seja realizada com nível de confiança de 95%. Análises feitas na linha de produção indicam variância em torno de 2 mm².

- 15.** Considerando o Exercício 14, mas supondo a caixa com 1.000 parafusos, qual é o tamanho da amostra necessário?
- 16.** Com o objetivo de avaliar a confiabilidade de um novo sistema de transmissão de dados, torna-se necessário verificar a proporção de *bits* transmitidos com erro em cada lote de 100 Mb. Considere que seja tolerável um erro amostral máximo de 2% e que em sistemas similares a taxa de erro na transmissão é de 10%. Qual deve ser o tamanho da amostra?
- a)** Use $\gamma = 0,95$.
- b)** Use $\gamma = 0,99$.

EXERCÍCIOS COMPLEMENTARES

- 17.** Sob condições normais, realizaram-se dez observações sobre o tempo de resposta de uma consulta a certo banco de dados. Os resultados, em segundos, foram:

28 35 43 23 62 38 34 27 32 37

Construa um intervalo de confiança para o tempo médio de uma consulta, sob condições normais. Use $\gamma = 0,99$.

- 18.** Fixados certos parâmetros de entrada, o tempo de execução de um algoritmo foi medido 12 vezes, obtendo-se os seguintes resultados, em minutos:

15 12 14 15 16 14 16 13 14 11 15 13

- a)** Apresente um intervalo de 95% de confiança para o tempo médio de execução do algoritmo.
- b)** Considerando as 12 mensurações como uma amostra piloto, avalie o número de mensurações (tamanho da amostra) necessário para garantir um erro máximo de 15 segundos (0,25 minutos). Use $\gamma = 0,95$.
- 19.** Uma empresa tem 2.400 empregados. Deseja-se extrair uma amostra de empregados para verificar o grau de satisfação em relação à qualidade da comida no refeitório. Em uma amostra piloto, numa escala de 0 a 10, o grau de satisfação recebeu nota média 6,5 e desvio padrão 2,0.

- a) Determine o tamanho mínimo da amostra, supondo amostragem aleatória simples, com erro máximo de 0,5 unidade e nível de confiança de 99%.
- b) Considere que a amostra planejada no item anterior tenha sido realizada e obteve-se média 5,3 e desvio padrão 1,8 ponto. Construa um intervalo de 99% de confiança para o parâmetro μ .
- c) Considerando o resultado do item anterior, você diria, com nível de confiança de 99%, que a nota média seria superior a cinco se a pesquisa fosse aplicada a todos os 2.400 funcionários? Justifique.
- d) Realizada a amostra planejada no item (a), suponha que 70 atribuíram notas iguais ou superiores a cinco. Apresente um intervalo de 90% de confiança para a porcentagem de indivíduos da população que atribuíram notas iguais ou superiores a cinco.
- 20.** Com os dados históricos sobre a temperatura do pasteurizador de um laticínio, sabe-se que a variância é aproximadamente $1,8 (^{\circ}\text{C})^2$. Planeja-se fazer uma amostragem para avaliar o valor médio da temperatura do pasteurizador. Suponha que as observações sejam feitas sob as mesmas condições e de forma independente. Qual deve ser o tamanho da amostra, para garantir um erro máximo de $0,3^{\circ}\text{C}$, com nível de confiança de 95%?
- 21.** Planeja-se extrair uma amostra aleatória simples dos 2.000 funcionários de uma empresa, para avaliar a satisfação com o trabalho. A satisfação será avaliada através de um questionário com vários itens numa escala de 1 a 5. Pretende-se avaliar o valor médio de cada item. Qual deve ser o tamanho da amostra para garantir um erro máximo de 0,2 unidade, com nível de confiança de 95%?
- Nota:** Use como variância o valor teórico que se obtém ao supor probabilidade igual para cada um dos cinco níveis da escala. Observe que dificilmente algum item terá variância maior do que esta que você está calculando, pois, na prática, a tendência é que as respostas se concentrem em torno de algum nível.
- 22.** Numa pesquisa para estudar a preferência do eleitorado a uma semana da eleição presidencial, qual deve ser o tamanho de uma amostra aleatória simples para garantir, com nível de confiança de 95%, um erro amostral não superior a 2%?
- 23.** Um analista de sistemas está avaliando o desempenho de um novo programa de análise numérica. Forneceu como entrada do programa 14 operações similares e obteve os seguintes tempos de processamento (em milissegundos):

12,0	13,5	16,0	15,7	15,8	16,5	15,0	13,1
15,2	18,1	18,5	12,3	17,5	17,0		

- a) Calcule a média e o desvio padrão da amostra do tempo de processamento.
 - b) Construir um intervalo de confiança para o tempo médio de processamento, com nível de confiança de 95%.
 - c) Qual deve ser o tamanho da amostra para garantir um erro amostral máximo de 0,5 milissegundo, na estimação do tempo médio de processamento, com nível de confiança de 99%?
- 24.** Uma unidade fabril da Intel produziu 500.000 *chips Pentium IV* em certo período. São selecionados, aleatoriamente, 400 *chips* para testes.
- a) Supondo que 20 *chips* não tenham a velocidade de processamento adequada, construir o intervalo de confiança para a proporção de *chips* adequados. Use nível de confiança de 95%.
 - b) Verificar se essa amostra é suficiente para obter um intervalo de 99% de confiança, com erro amostral máximo de 0,5%, para a proporção de *chips* adequados. Caso contrário, qual deveria ser o tamanho da amostra?

8

Testes de Hipóteses

Muitas vezes o pesquisador tem alguma ideia, ou conjetura, sobre o comportamento de uma variável, ou de uma possível associação entre variáveis. Nesses casos, o planejamento da pesquisa deve ser de tal forma que permita, com os dados amostrais, testar a veracidade de suas ideias sobre a população (ou as populações) em estudo. Adotamos que a população seja o mundo real e as ideias sejam as hipóteses de pesquisa, que poderão ser testadas por técnicas estatísticas denominadas *testes de hipóteses* ou *testes de significância*.

8.1 AS HIPÓTESES

Exemplo 8.1 Seguem alguns exemplos de hipóteses:

- a)** substituindo o processador *A* pelo processador *B*, altera-se o tempo de resposta de um computador;
- b)** aumentando a dosagem de cimento, aumenta-se a resistência do concreto;
- c)** uma campanha publicitária produz efeito positivo nas vendas;
- d)** a implementação de um programa de melhoria da qualidade em uma empresa prestadora de serviços melhora a satisfação de seus clientes.

Para verificar estatisticamente a veracidade de uma hipótese, precisamos de um conjunto de dados, observados adequadamente em termos do problema

em questão. No Exemplo 8.1 (a), podemos realizar um experimento, fazendo uma amostra de ensaios com o processador *A* e uma amostra de ensaios com o processador *B*, anotando o tempo de resposta em cada ensaio. Em (b), podemos fazer uma amostra de corpos de prova com a dosagem d_1 de cimento e uma amostra de corpos de prova com a dosagem d_2 de cimento ($d_2 > d_1$), medindo a resistência de cada corpo de prova. Em (c), podemos verificar o nível das vendas em vários pontos de venda antes do início da campanha publicitária e, também, depois desse evento. Em (d), podemos avaliar o percentual de reclamações antes e depois do programa de melhoria da qualidade.

Levando em conta o planejamento da pesquisa, as hipóteses podem ser colocadas de forma mais específica, descritas em termos de parâmetros populacionais. Com respeito ao Exemplo 8.1, temos:

- a)** a *média* dos tempos de resposta do equipamento com o processador *A* é diferente da *média* dos tempos de resposta com o processador *B*;
- b)** a *média* dos valores de resistência do concreto com a dosagem d_2 de cimento é maior do que a *média* dos valores de resistência com a dosagem d_1 ;
- c)** a *média* das vendas depois da campanha publicitária é maior do que a *média* das vendas antes da campanha publicitária;
- d)** a *proporção* de reclamações após a realização do programa de melhoria da qualidade é menor do que antes da realização do programa.

Observe que, nos três primeiros casos, as hipóteses estão descritas em termos da comparação de duas *médias*, enquanto no caso (d), a hipótese está descrita em termos da comparação de duas *proporções*. E, também, no caso (a), queremos verificar se há *diferença* entre as duas condições em estudo, enquanto nos demais casos especificamos na hipótese qual parâmetro deve ser o *maior*. Essas diferenças dependem do problema e do planejamento do experimento ou do levantamento da coleta de dados.

Dado um problema de teste de hipóteses, precisamos formular as chamadas *hipótese nula* e *hipótese alternativa*.

A *hipótese nula* ou *hipótese de trabalho* (H_0) é a hipótese aceita como verdadeira até prova estatística em contrário. É o ponto de partida para a análise dos dados. Em geral, ela é formulada em termos de igualdade entre parâmetros, ou entre um parâmetro e uma constante. Ela geralmente representa o *contrário* do que queremos provar.¹

¹ Lembre-se de que a prova, na estatística, não é exata, pois existe probabilidade de incorrerem em erro, mas é desejável que essa probabilidade seja pequena e conhecida.

Quando os dados mostrarem evidência suficiente de que a hipótese nula (H_0) é falsa, o teste rejeita-a, aceitando em seu lugar a chamada *hipótese alternativa* (H_1). Em geral, a **hipótese alternativa** (H_1) é formulada em termos de desigualdades (\neq , $<$ ou $>$). Ela comumente corresponde ao que se quer provar, ou seja, corresponde à própria *hipótese de pesquisa* formulada em termos de parâmetros.

Com respeito ao Exemplo 8.1, temos:

a) $H_0: \mu_A = \mu_B$ e $H_1: \mu_A \neq \mu_B$

onde: μ_A é o tempo médio de resposta com o processador A; e
 μ_B é o tempo médio de resposta com o processador B.

b) $H_0: \mu_2 = \mu_1$ e $H_1: \mu_2 > \mu_1$

onde: μ_2 é a resistência média do concreto com a dosagem d_2 de cimento; e
 μ_1 é a resistência média do concreto com a dosagem d_1 de cimento.

c) $H_0: \mu_2 = \mu_1$ e $H_1: \mu_2 > \mu_1$

onde: μ_1 é o valor médio das vendas antes da campanha publicitária; e
 μ_2 é o valor médio das vendas depois da campanha publicitária.

d) $H_0: p_2 = p_1$ e $H_1: p_2 < p_1$

onde: p_1 é a proporção de reclamações antes do programa de melhoria da qualidade; e
 p_2 é a proporção de reclamações depois do programa de melhoria da qualidade.

A decisão em aceitar H_0 ou H_1 é feita a partir de amostras extraídas adequadamente das populações envolvidas. No caso (d), por exemplo, podemos observar uma amostra aleatória de n_1 clientes atendidos *antes* do programa de qualidade e outra amostra aleatória de n_2 clientes atendidos *depois* de implantado o programa. É natural que as proporções de reclamações nas duas amostras (\hat{p}_1 e \hat{p}_2) sejam diferentes, mesmo que a hipótese nula ($H_0: p_2 = p_1$) seja verdadeira, pois sempre existe o efeito aleatório na seleção das amostras.

A aplicação de um *teste estatístico* (ou *teste de significância*) serve para verificar se os dados fornecem evidência suficiente para que possamos aceitar como verdadeira a hipótese alternativa (H_1), precavendo-nos, com certa segurança, de que as diferenças observadas nos dados não são meramente casuais.

8.2 CONCEITOS BÁSICOS

Apresentaremos as primeiras ideias sobre testes estatísticos usando como ilustração um experimento binomial. Considere o seguinte problema:

Verificar se uma moeda, usada num jogo de azar, é viciada.

Chamamos de p a probabilidade de cara dessa moeda. Assim, podemos formular as hipóteses da seguinte maneira:

$H_0: p = 0,5$ (a moeda é honesta) e $H_1: p \neq 0,5$ (a moeda é viciada)

Suponha, inicialmente, H_0 como verdadeira. Ela somente vai ser rejeitada em favor de H_1 se houver evidência suficiente que a contradiga. A existência dessa possível evidência será verificada com base num conjunto de observações do problema em estudo. No presente exemplo, o conjunto de observações (amostra) consistirá nos resultados de uma série de lançamentos imparciais da moeda.

Em cada lançamento da moeda, observamos um resultado: *cara* ou *coroa*. Ao observar uma amostra de n lançamentos, podemos computar o valor da estatística:

$Y =$ número total de caras nos n lançamentos

A estatística Y poderá ser usada na definição de um critério de decisão: *aceitar H_0 ou rejeitar H_0 em favor de H_1* . Nesse contexto, a estatística Y é chamada de **estatística do teste**.

Vamos considerar uma amostra de $n = 10$ lançamentos e duas situações:

SITUAÇÃO A: Suponha que, nos 10 lançamentos, observamos $Y = 10$ caras. Podemos rejeitar H_0 , em favor de H_1 ?

SITUAÇÃO B: E se tivéssemos observado $Y = 7$ caras?

Na situação A, é intuitivo que exista mais evidência para rejeitar H_0 . Contudo, em nenhuma das duas situações, podemos rejeitar H_0 com plena certeza de que essa hipótese é realmente falsa, pois estamos trabalhando com um fenômeno aleatório, no qual é plenamente possível, em 10 lançamentos de uma moeda sabidamente honesta (H_0 verdadeira), ocorrerem 7, 8, 9 ou até mesmo 10 caras. Por outro lado, se a ocorrência de um resultado for muito pouco provável para uma moeda honesta, torna-se natural decidirmos por H_1 (moeda viciada).

No presente contexto, é necessário conhecermos a probabilidade de ocorrerem resultados como $Y = 10$ caras (situação A) ou $Y = 7$ caras (situação B), em 10 lançamentos de uma moeda honesta. Como se trata de uma variável aleatória, precisamos da distribuição de probabilidades da estatística do teste Y ,

admitindo H_0 verdadeira. Essa distribuição de probabilidades será a referência básica para analisarmos o resultado da amostra e decidirmos entre H_0 e H_1 .

A distribuição de probabilidades de Y : distribuição de referência do teste

Pelas características do experimento, podemos afirmar que Y tem distribuição binomial, com parâmetros $n = 10$ e $p = 0,5$ (supondo H_0 verdadeira). A Figura 8.1 apresenta essa distribuição sob forma gráfica. As probabilidades $p(y)$ foram extraídas da Tabela 1 do apêndice, ou seja, são derivadas da expressão da distribuição binomial, como visto no Capítulo 5. Para facilitar a exposição, essas probabilidades foram arredondadas para três decimais.

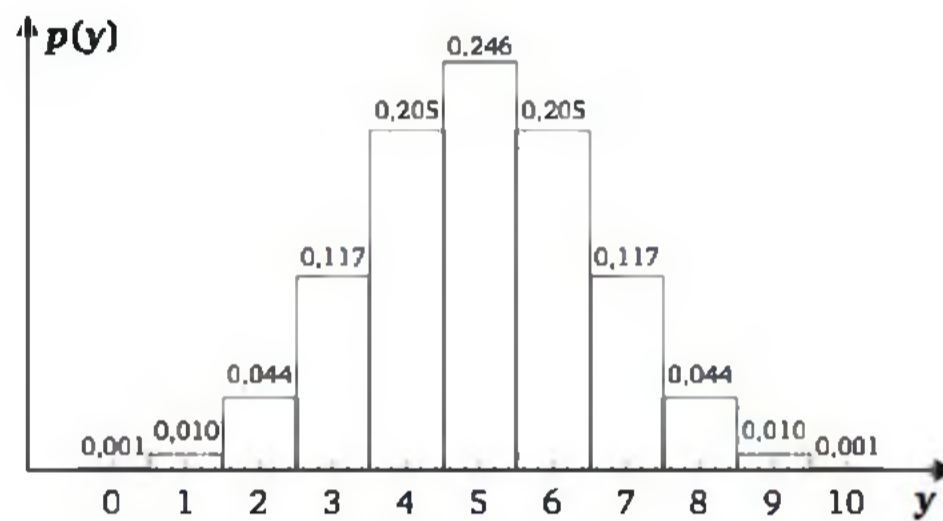


Figura 8.1 Distribuição da estatística $Y =$ número de caras em 10 lançamentos da moeda, sob H_0 (binomial com $n = 10$ e $p = 0,5$).

Com a distribuição de probabilidades da estatística do teste, podemos avaliar melhor a adequação de H_0 com o resultado de Y , calculado com base na amostra. A Figura 8.1 mostra que, se H_0 for verdadeira, os resultados mais prováveis estão em torno do *valor esperado*: $\mu = n.p = 5$ caras.

A distribuição de referência do teste identifica a variável aleatória, ou seja, diz quais resultados são mais prováveis e quais são improváveis. Com base nessa informação, devemos, em seguida, comparar o que *ocorreu* na amostra (resultado da estatística do teste) com o que *esperaríamos* que tivesse ocorrido, considerando a hipótese nula verdadeira.

Valor p

Considerando H_0 verdadeira, podemos calcular a probabilidade de obtermos, por acaso, um resultado tão ou mais longe do esperado do que aquele que verificamos na amostra efetivamente observada. A essa probabilidade damos o nome de *probabilidade de significância*, ou *valor p* .

O valor p (ou **probabilidade de significância**) é definido como a probabilidade de a estatística do teste acusar um resultado tão ou mais distante do esperado, como o resultado ocorrido na particular amostra observada, supondo H_0 como a hipótese verdadeira.²

Exemplo 8.2 Retomemos a situação A, em que observamos $Y = 10$ caras em $n = 10$ lançamentos. Considerando o número esperado de caras sob H_0 ($\mu = 5$) como referência, verificamos que tão ou mais distante do que o valor observado na amostra ($Y = 10$) encontram-se o valor 0 e o próprio valor 10, como ilustra a Figura 8.2.

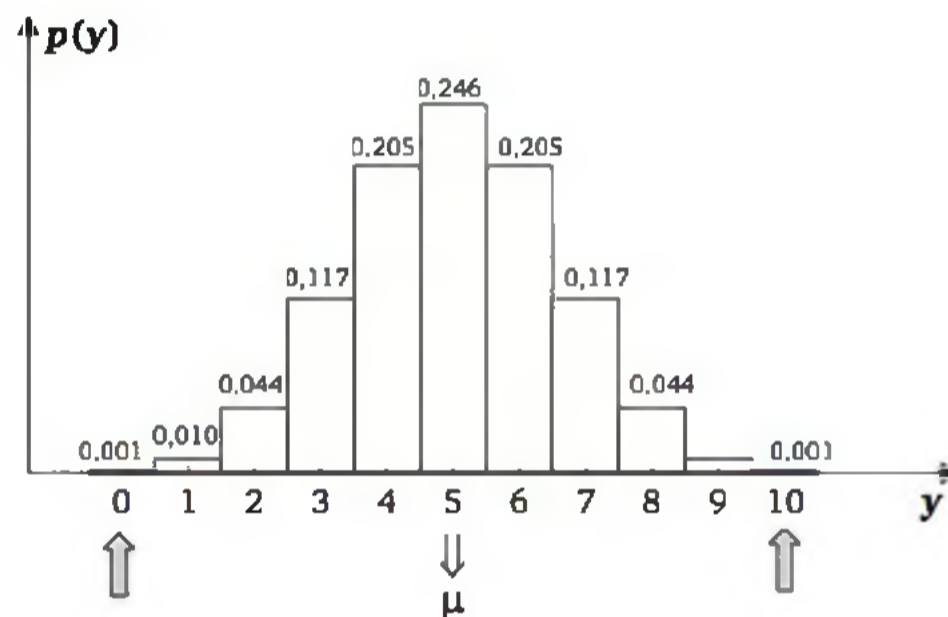


Figura 8.2 Distribuição de Y , sob H_0 . As setas indicam os valores que distam do esperado ($\mu = 5$) tão ou mais do que o valor $Y = 10$, observado na amostra da situação A.

Consequentemente, a probabilidade de significância será:

$$p = p(0) + p(10) = 0,001 + 0,001 = 0,002$$

Assim, para uma moeda honesta (H_0 verdadeira), temos a pequena probabilidade $p = 0,002$ de ocorrer um resultado tão ou mais distante do valor esperado, como o que, de fato, ocorreu na amostra efetivamente observada ($Y = 10$ caras). Como 0,002 é uma probabilidade muito pequena, torna-se natural rejeitar a hipótese de que a moeda é honesta (H_0), decidindo-se pela hipótese de que a moeda é viciada (H_1).

Os dados observados mostram evidência suficiente para dizer que a moeda é viciada!

² Esta definição é apropriada para distribuições de referência simétricas e testes bilaterais (ver seção 8.5).

Exemplo 8.3 Vejamos, agora, a situação B, em que observamos $Y = 7$ caras em $n = 10$ lançamentos. Nesta situação, tão ou mais distante do que o valor $Y = 7$ encontram-se os valores: 7, 8, 9, 10, 0, 1, 2 e 3, como ilustra a Figura 8.3.

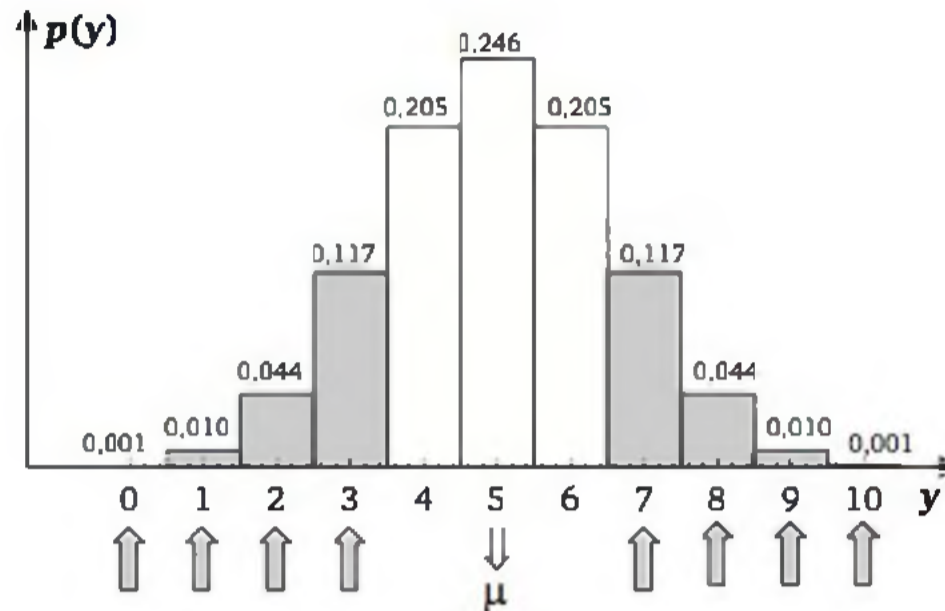


Figura 8.3 Distribuição de Y , sob H_0 . As setas indicam os valores que distam do esperado ($\mu = 5$) tão ou mais do que o valor $Y = 7$, observado na amostra da situação B.

Temos, então, a seguinte probabilidade de significância:

$$\begin{aligned} \mathbf{p} &= p(0) + p(1) + p(2) + p(3) + p(7) + p(8) + p(9) + p(10) = \\ &= 0,001 + 0,010 + 0,044 + 0,117 + 0,117 + 0,044 + 0,010 + \\ &+ 0,001 = 0,344 \end{aligned}$$

Esta segunda situação mostra que, para uma moeda honesta (H_0 verdadeira), temos a probabilidade $\mathbf{p} = 0,344$ de ocorrer um resultado tão ou mais distante do valor esperado, como o que, de fato, ocorreu neste caso ($Y = 7$ caras). Como 0,344 não é uma probabilidade desprezível, torna-se mais prudente **não** rejeitar H_0 .

Não há evidência suficiente para afirmar que a moeda é viciada!

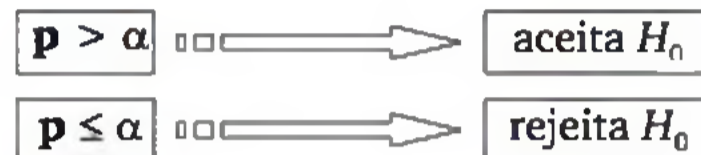
O valor \mathbf{p} aponta o quão *improvável* foi o resultado da amostra à luz de H_0 . Logo, quanto menor for o valor \mathbf{p} , maior a evidência para rejeitar H_0 . O valor \mathbf{p} também pode ser interpretado como o risco de rejeitarmos incorretamente uma H_0 verdadeira, dada a evidência da amostra. Por exemplo, se afirmássemos que a moeda é viciada com a evidência de $Y = 7$ caras em $n = 10$ lançamentos, estaríamos incorrendo numa probabilidade de 34,4% de estarmos fazendo uma afirmação errada.

Nível de significância

Na realização de uma pesquisa, quando desejamos confirmar ou refutar alguma hipótese, é comum estabelecer, ainda na fase do planejamento da pesqui-

sa, a probabilidade tolerável de incorrer no erro de rejeitar H_0 , quando H_0 é verdadeira. Este valor é conhecido como **nível de significância do teste** e é designado pela letra grega α . É comum adotar nível de significância de 5%, isto é, $\alpha = 0,05$. Mas, quando desejamos maior segurança ao afirmar H_1 , podemos adotar níveis de significância menores, como $\alpha = 0,01$.

Estabelecido o nível de significância α , temos a seguinte regra geral de decisão de um teste estatístico:



Exemplo 8.2 (continuação) Na amostra da situação A, quando observamos 10 caras em 10 lançamentos, se estivermos usando o nível de significância de 5% ($\alpha = 0,05$), o teste estatístico *rejeita* H_0 em favor de H_1 , pois a probabilidade de significância, calculada com base na amostra, foi de 0,002, sendo, portanto, *menor* do que o valor adotado para α .

Exemplo 8.3 (continuação) Usando $\alpha = 0,05$ na amostra da situação B, quando observamos 7 caras em 10 lançamentos, o teste estatístico *não* rejeita H_0 , pois a probabilidade de significância, calculada com base na amostra, foi de 0,344, que *não é menor* do que o valor adotado para α .

8.3 TIPOS DE ERRO

Para entendermos os erros associados a um teste de hipóteses, precisa ficar claro que estamos avaliando uma afirmação (hipótese) sobre a população. Essa hipótese pode ou não ser verdadeira, mas *nunca* saberemos com certeza sobre a realidade da população, já que conhecemos apenas uma amostra. O que fazemos é *tomar uma decisão*, considerando evidências na amostra. Antes de observarmos a amostra, mas estabelecidas as hipóteses e o nível de significância α , temos as seguintes possibilidades:

Realidade (desconhecida)	Decisão do teste	
	Aceita H_0	Rejeita H_0
H_0 verdadeira	Decisão correta (probab = $1 - \alpha$)	Erro tipo I (probab = α)
H_0 falsa	Erro tipo II (probab = β)	Decisão correta (probab = $1 - \beta$)

Conforme o esquema, os erros tipos I e II podem ocorrer segundo as seguintes probabilidades condicionais:

$$P(\text{erro tipo I}) = P(\text{rejeitar } H_0 \mid H_0 \text{ é verdadeira}) = \alpha$$

$$P(\text{erro tipo II}) = P(\text{aceitar } H_0 \mid H_0 \text{ é falsa}) = \beta$$

Usando os eventos complementares, temos as probabilidades de decisão correta:

$$P(\text{aceitar } H_0 \mid H_0 \text{ é verdadeira}) = 1 - \alpha \text{ e}$$

$$P(\text{rejeitar } H_0 \mid H_0 \text{ é falsa}) = 1 - \beta$$

Como α é fixado *a priori*, se o teste *rejeita* H_0 em favor de H_1 ($\mathbf{p} \leq \alpha$), o risco de estarmos tomando a decisão errada (erro tipo I) fica limitado pelo nível de significância α adotado. Dessa forma, temos certa garantia da veracidade de H_1 . Em outras palavras, quando o teste *rejeita* H_0 , podemos afirmar (considerando o nível de significância do teste) que H_1 é verdadeira.

Por outro lado, se o teste *aceita* H_0 ($\mathbf{p} > \alpha$), não temos muito controle do risco de estarmos tomando a decisão errada (erro tipo II), pois a probabilidade β , em geral, não é conhecida. Nesse caso, costumamos dizer apenas que *os dados estão em conformidade com a hipótese nula*. Isso não implica que H_0 seja realmente a hipótese verdadeira, mas que os dados não estão mostrando evidência suficiente para rejeitá-la. Sobre isso, R. A. Fisher, considerado o pai da estatística experimental, escreveu:

“A hipótese nula pode ou não ser impugnada pelos resultados de um experimento. Ela nunca pode ser provada, mas pode ser desaprovada no curso da experimentação.”

Em razão do exposto, usamos uma linguagem mais enfática para descrever uma situação em que o teste *rejeita* H_0 (por exemplo, *os dados provaram estatisticamente que a moeda é viciada*) e uma linguagem mais amena quando o teste *aceita* H_0 (por exemplo, *admite-se que a moeda é honesta, pois os dados não mostraram evidência suficiente de que ela é viciada*).

8.4 ABORDAGEM CLÁSSICA

Antes de observar a amostra, mas considerando o planejamento do experimento, podemos montar a regra de decisão em termos da estatística do teste, do nível de significância e da distribuição de probabilidade da estatística do teste sob H_0 . Depois, observamos a amostra e tomamos a decisão em função da regra previamente construída.

Retomando o experimento de lançar 10 vezes a moeda, a regra de decisão para $\alpha = 0,05$ é construída com base na equação:

$$P(\text{erro tipo I}) = P(\text{rejeitar } H_0 \mid H_0 \text{ é verdadeira}) = \alpha$$

Se H_0 é verdadeira, a distribuição para o número de caras Y é binomial com $n = 10$ e $p = 0,5$. E a região de rejeição de H_0 deve estar nas extremidades dessa distribuição, de tal forma que a probabilidade total não ultrapasse o nível de significância estabelecido: $\alpha = 0,05$ (Figura 8.4).³

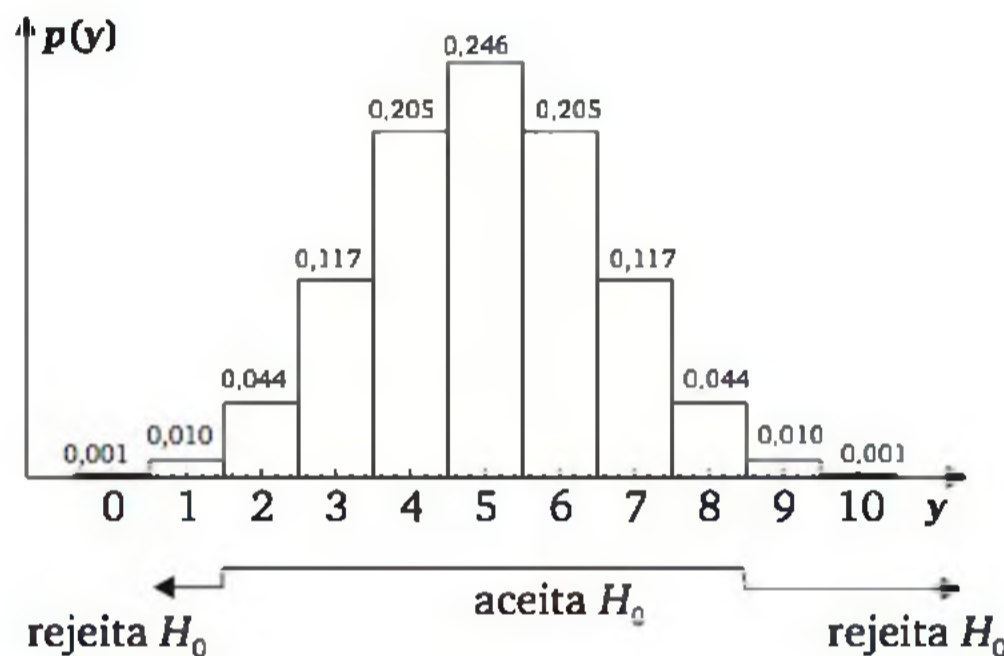


Figura 8.4 Regra de decisão em termos de $Y =$ número de caras em 10 lançamentos da moeda, com $\alpha = 0,05$.

Lançamos a moeda 10 vezes e contamos o número Y de caras. Decidimos por H_0 ou H_1 , conforme a regra ilustrada na Figura 8.4, ou seja:

- se ocorrer de 2 a 8 caras, então o teste aceita H_0 ;
- se ocorrer menos de 2 caras ou mais de 8 caras, então o teste rejeita H_0 em favor de H_1 .

A abordagem clássica é mais fácil quando o trabalho é realizado sem a ajuda do computador, usando apenas as tabelas estatísticas. Por outro lado, quando usamos algum *software* estatístico ou uma planilha eletrônica, o sistema já calcula a estatística do teste e o valor \mathbf{p} , de tal forma que a abordagem discutida na seção 8.2 torna-se mais fácil. Contudo, a decisão sempre será a mesma, independentemente da abordagem utilizada.

³ Como no presente caso a distribuição do teste é discreta, não existe um valor que satisfaz exatamente à equação $P(\text{erro tipo I}) = \alpha$. Então, a regra é construída com o maior α que satisfaça a $P(\text{erro tipo I}) \leq \alpha$.

EXERCÍCIOS

1. Seja p a probabilidade de cara de uma moeda. Sejam $H_0: p = 0,5$ e $H_1: p \neq 0,5$. Lança-se 12 vezes esta moeda, observando-se o número de caras. Usando a tabela da distribuição binomial (Tabela 1 do apêndice), obtenha a probabilidade de significância para cada um dos seguintes resultados:
 - a) 1 cara;
 - b) 4 caras; e
 - c) 11 caras.
2. Adotando o nível de significância de 5%, qual é a conclusão do teste em cada item do Exercício 1?
3. Sejam 15 lançamentos de uma moeda. Considerando as hipóteses H_0 , a moeda é honesta, e H_1 , a moeda é viciada, e o nível de significância de 1%, apresente a regra de decisão (abordagem clássica) em termos da estatística $Y = \text{número de coroas}$.
4. É possível, para uma mesma amostra, aceitar H_0 ao nível de significância de 1%, mas rejeitá-la ao nível de 5%? E o inverso? Exemplifique.
5. Para verificar as hipóteses de seu trabalho, um pesquisador fez vários testes estatísticos (um para cada hipótese de pesquisa), adotando para cada teste o nível de significância de 5%. Responda os seguintes itens:
 - a) Num dado teste, o valor p foi 0,0001. Qual deve ser a conclusão (decide-se pela hipótese nula ou pela hipótese alternativa)? Qual é o risco de o pesquisador estar tomando a decisão errada?
 - b) Em outro teste, o valor p foi 0,25. Qual deve ser a conclusão? Nesse caso, você consegue avaliar o risco de o pesquisador estar tomando a decisão errada?
 - c) Em outros dois testes, as probabilidades de significância foram de 0,0001 e 0,01, respectivamente. Em qual dos testes o pesquisador deve estar mais convicto na decisão de qual hipótese deve ser escolhida? Por quê?

8.5 TESTES UNILATERAIS E BILATERAIS

No exemplo discutido no tópico anterior, rejeita-se $H_0: p = 0,5$, em favor de $H_1: p \neq 0,5$, tanto quando ocorre um valor muito pequeno, quanto muito grande de caras. Essa é uma situação típica de *teste bilateral*.

Existem situações em que pretendemos rejeitar H_0 somente em um dos sentidos. Por exemplo, suspeitamos que a moeda tende a dar mais caras do que coroas. Nesse caso, sendo p a probabilidade de ocorrer cara, o teste pode ser formulado da seguinte maneira:

$H_0: p = 0,5$ (a moeda é honesta) e

$H_1: p > 0,5$ (a moeda tende a dar mais caras do que coroas).

Com essas hipóteses, só faz sentido rejeitar H_0 , em favor de H_1 , se na amostra ocorrer um número significativamente maior de caras do que de coroas, resultando no que chamamos de *teste unilateral*. Assim, nos testes unilaterais, a probabilidade de significância é computada em apenas um dos lados da distribuição de referência.

Exemplo 8.4 Considere que, para testar $H_0: p = 0,5$ contra $H_1: p > 0,5$, tenhamos lançado a moeda $n = 10$ vezes e observado $Y = 7$ caras. A probabilidade de significância será:

$$\mathbf{p} = p(7) + p(8) + p(9) + p(10) = 0,117 + 0,044 + 0,010 + 0,001 = 0,172$$

que corresponde à metade da probabilidade de significância do teste bilateral, discutido no Exemplo 8.3. Com o nível de significância de 5%, o teste *não* rejeita H_0 . A Figura 8.5 ilustra a probabilidade de significância desse teste.

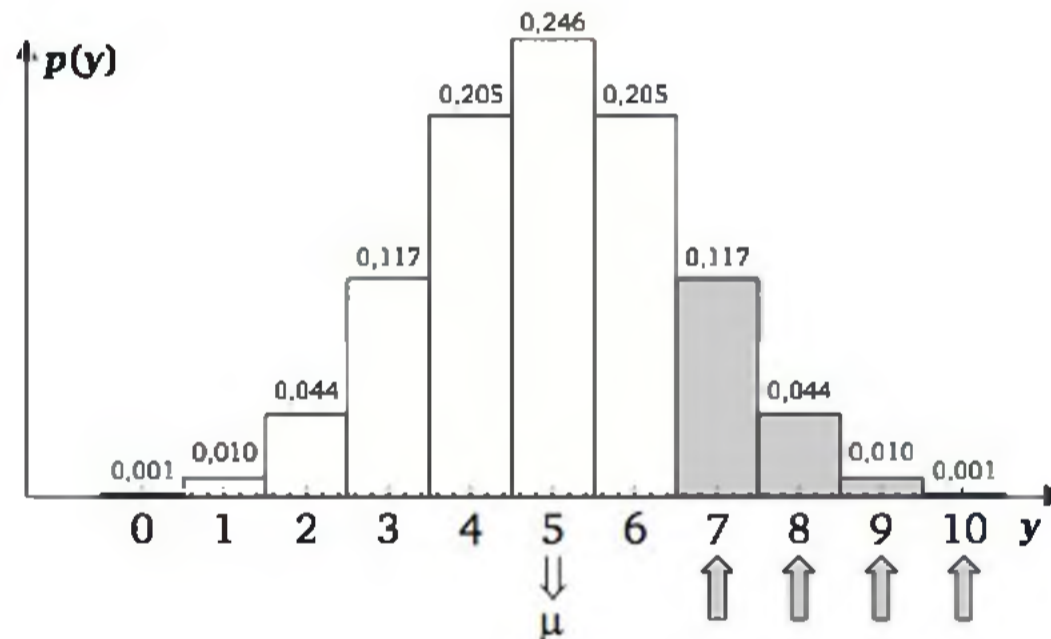


Figura 8.5 Ilustração do cálculo da probabilidade de significância do teste unilateral do Exemplo 8.4.

Uso da abordagem clássica

Se usarmos a abordagem clássica, com $\alpha = 0,05$, montamos a seguinte regra de decisão antes de observarmos a amostra:⁴

⁴ Note que $P(Y \geq 9) = 0,011 < 0,05$; já $P(Y \geq 8) = 0,055$, superando o nível de significância adotado ($\alpha = 0,05$). Assim, nove caras deve estar na região de rejeição e oito caras na região de aceitação de H_0 .

- Se ocorrer oito ou menos caras, então o teste aceita H_0 .
- Se ocorrer mais de oito caras, então o teste rejeita H_0 (aceita H_1).

Exemplo 8.5 Com o objetivo de testar se a diferença de odor em sorvetes de morango é percebida por degustadores, efetuou-se um experimento, como descrito a seguir.

Para cada um dos oito degustadores selecionados para o experimento, foram dadas, em ordem aleatória e sem identificação, duas amostras de sorvete: uma com odor mais forte e outra normal. As amostras de sorvete foram elaboradas de forma tão similar quanto possível, com exceção da intensidade de odor, que é a característica em estudo.

Chamando de p a probabilidade de o degustador acusar corretamente a amostra de sorvete com odor mais intenso, temos interesse em testar as seguintes hipóteses:

- $H_0: p = 0,5$ (o degustador *chuta* a resposta, isto é, o odor mais intenso não é detectado) e
- $H_1: p > 0,5$ (existe uma tendência de o degustador perceber o sorvete que tem o odor mais intenso).

Seja Y o número de degustadores que indicam corretamente o sorvete com odor mais intenso. Pelas características do experimento, podemos deduzir que, se H_0 for correta, a estatística Y tem distribuição binomial com $n = 8$ e $p = 0,5$.

Os resultados do experimento mostraram que, dos oito degustadores, seis indicaram corretamente o sorvete de odor mais intenso ($Y = 6$). Usando a distribuição binomial (Tabela 1 do apêndice), podemos computar a probabilidade de significância:

$$\mathbf{p} = p(6) + p(7) + p(8) = 0,109 + 0,031 + 0,004 = 0,144$$

Assim, se estamos trabalhando com o nível de significância de 5% ($\alpha = 0,05$), a hipótese nula *não* pode ser rejeitada. Portanto, concluímos que os dados resultantes do experimento são insuficientes para se afirmar que a diferença de odor em sorvetes de morango seja percebida pelos degustadores.

EXERCÍCIOS

6. Para cada um dos itens do Exemplo 8.1, descrever qual a abordagem (unilateral ou bilateral) que é mais apropriada.
7. Para cada um dos itens a seguir, apresente as hipóteses nula e alternativa, indicando qual é a abordagem (unilateral ou bilateral) mais adequada.

- a) Um método de treinamento tende a aumentar a produtividade dos funcionários.
- b) A velocidade de um veículo num percurso é, em média, menor do que o valor anunciado.
- c) Dois métodos de treinamento tendem a produzir resultados diferentes na produtividade.
8. Seja p a probabilidade de cara de uma moeda. Sejam $H_0: p = 0,5$ e $H_1: p < 0,5$. Lança-se 12 vezes essa moeda, observando-se o número de caras. Usando a tabela da distribuição binomial (Tabela 2 do apêndice), obtenha a probabilidade de significância para cada um dos seguintes resultados:
- a) 1 cara b) 4 caras c) 6 caras
- Usando nível de significância de 5%, em quais resultados o teste rejeita H_0 ?
9. Sejam 11 lançamentos de uma moeda. Considerando as hipóteses $H_0: a$ moeda é honesta e $H_1: a$ moeda tende a dar mais cara do que coroa e o nível de significância de 1%, apresente a regra de decisão em termos da estatística $Y = \text{número de coroas}$.
10. Para testar se uma criança tem algum conhecimento sobre determinado assunto, elaboraram-se 12 questões do tipo *certo-errado*. A criança acertou 11. Qual é a conclusão ao nível de significância de 5%?
11. Para testar se uma criança tem algum conhecimento sobre determinado assunto, elaboraram-se 12 questões, cada uma com 4 possibilidades de escolha. A criança acertou 5.
- a) Formule as hipóteses em termos do parâmetro $p =$ probabilidade de acerto de cada questão.
- b) Qual é o número esperado de acertos sob H_0 ?
- c) Qual é o valor p ?
- d) Qual é a conclusão ao nível de significância de 5%?

8.6 APLICAÇÃO DE TESTES ESTATÍSTICOS

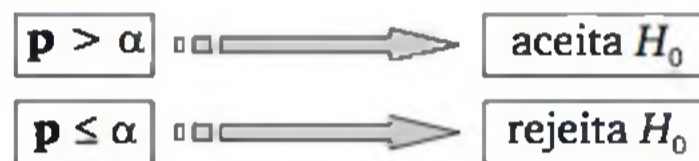
Formulada uma pergunta ou hipótese, o pesquisador precisa planejar a coleta de dados e um teste estatístico adequado à situação. Os testes diferenciam-se, basicamente, pelo tipo de problema que pretendemos resolver e pelo tipo de dados que temos ou que planejamos coletar. Com respeito aos tipos de dados, existem testes voltados para dados quantitativos, em que normalmente as hipóteses são apresentadas em termos de *médias*; e testes voltados para dados quali-

tativos, em que as hipóteses são apresentadas em termos de proporções ou probabilidades de eventos.

Em geral, na aplicação de um teste estatístico, devemos saber:

- a)** formular H_0 e H_1 em termos de parâmetros populacionais;
- b)** como obter a estatística do teste (no exemplo da moeda, $Y = \text{número de caras}$);
- c)** qual é a distribuição de referência adequada para calcular o valor \mathbf{p} (no exemplo da moeda é a distribuição *binomial*);
- d)** quais as suposições básicas para o uso do teste escolhido (no exemplo da moeda, supusemos que os lançamentos da moeda foram imparciais e realizados sob as mesmas condições).

A decisão do teste estatístico é feita pela comparação do valor \mathbf{p} com o nível de significância α preestabelecido, ou seja:



Se for usada a *abordagem clássica*, no passo (c), ao invés de calcular o valor \mathbf{p} , buscamos um valor crítico tabelado, em função do nível de significância adotado *a priori*. A decisão é feita comparando o valor da estatística do teste, calculado com base na amostra observada, com o valor crítico.

A conclusão resultante de um teste estatístico depende da aplicação em questão. Por exemplo, num estudo experimental, normalmente a decisão do teste implica uma relação de causa e efeito, mas, num estudo de levantamento, o resultado do teste usualmente leva apenas a uma conclusão de diferença entre grupos.

Hoje, o cálculo da estatística do teste e a obtenção do valor \mathbf{p} tornaram-se tarefas relativamente fáceis com o auxílio do computador. Ou seja, o pesquisador não mais precisa ter habilidades em cálculos algébricos para realizar testes estatísticos. Por outro lado, a análise do problema de pesquisa, o planejamento da coleta dos dados, a escolha do teste estatístico adequado, a verificação das suposições e a correta interpretação do resultado estatístico exigem conhecimento e raciocínio lógico.

8.7 TESTE PARA PROPORÇÃO

O teste para proporção é aplicado em situações nas quais queremos verificar se a proporção de algum atributo na população pode ser igual a certo va-

lor p_0 . No caso de o problema sugerir um teste bilateral, as hipóteses terão a forma:

$$H_0: p = p_0 \text{ e } H_1: p \neq p_0$$

No caso de teste unilateral, a hipótese alternativa seria H_1' : $p > p_0$ (unilateral à direita) ou H_1'' : $p < p_0$ (unilateral à esquerda).

Suponha que a amostragem ou o experimento garanta as suposições de um experimento binomial.⁵ No caso de a amostra ser pequena, podemos fazer o teste usando a própria distribuição binomial, conforme ilustrado com o exemplo da moeda nas seções anteriores. Nessa seção, enfatizaremos o caso em que a amostra é suficientemente grande para garantir a aproximação normal, ou seja,⁶

$$n \cdot p_0 \geq 5 \text{ e} \quad (8.1)$$

$$n \cdot (1 - p_0) \geq 5 \quad (8.2)$$

Seja:

$$\hat{p} = \frac{y}{n} = \frac{\text{número de elementos com o atributo de interesse}}{n} \quad (8.3)$$

a proporção do atributo de interesse, na amostra observada. O cálculo da estatística do teste pode ser feito por:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 \cdot (1 - p_0)}{n}}} = \frac{y - n \cdot p_0}{\sqrt{n \cdot p_0 \cdot (1 - p_0)}} \quad (8.4)$$

onde: p_0 é o valor da proporção, segundo H_0 ;

n é tamanho da amostra;

y é o número de elementos com o atributo de interesse, na amostra; e

\hat{p} é a proporção de elementos com o atributo de interesse, na amostra.

É recomendável o uso da correção de continuidade, fazendo

$$y' = y - 0,5 \text{ se } y > n \cdot p_0; \text{ ou} \quad (8.5)$$

$$y' = y + 0,5 \text{ se } y < n \cdot p_0 \quad (8.6)$$

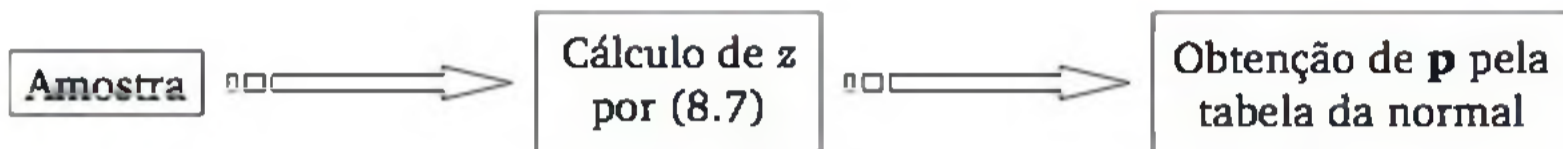
5 Ver Capítulo 5.

6 Ver seção 6.4.1.

Assim, o cálculo da estatística do teste deve ser feito por:

$$z = \frac{y' - n \cdot p_0}{\sqrt{n \cdot p_0 \cdot (1 - p_0)}} \tag{8.7}$$

Abordagem do valor p . O processo segue o esquema descrito abaixo:



Com o auxílio de uma tabela da distribuição normal padrão, obtemos a área acima de $|z|$. Se o teste for unilateral e a relação entre \hat{p} e p_0 for coerente com H_1 , então esta área já é o valor p .⁷ Se o teste for bilateral, o valor p corresponde ao dobro da área da cauda superior (veja Figura 8.6).

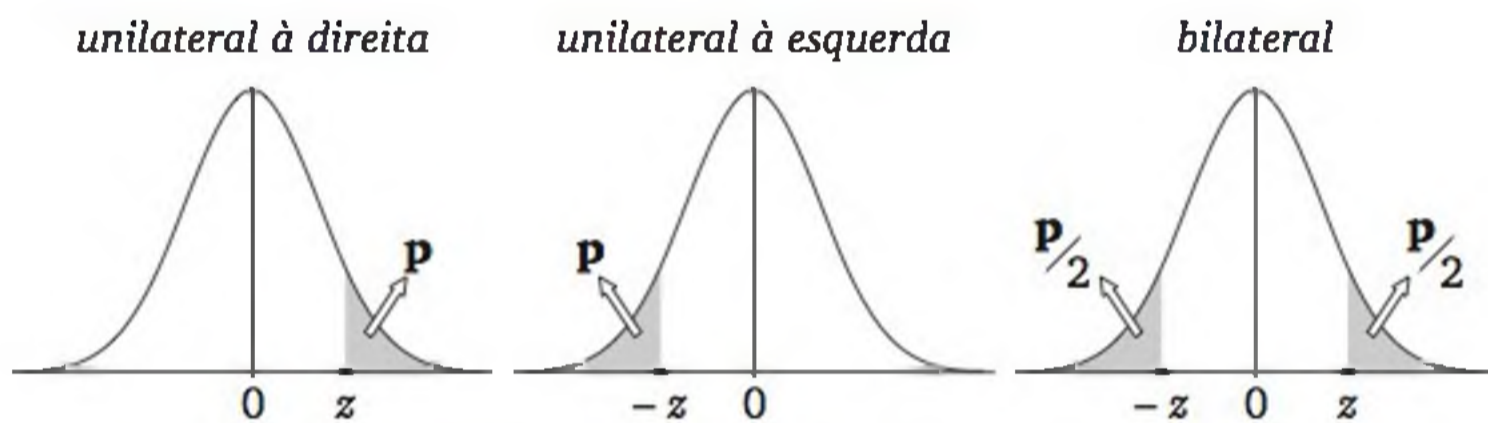
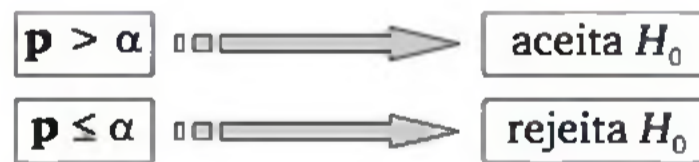
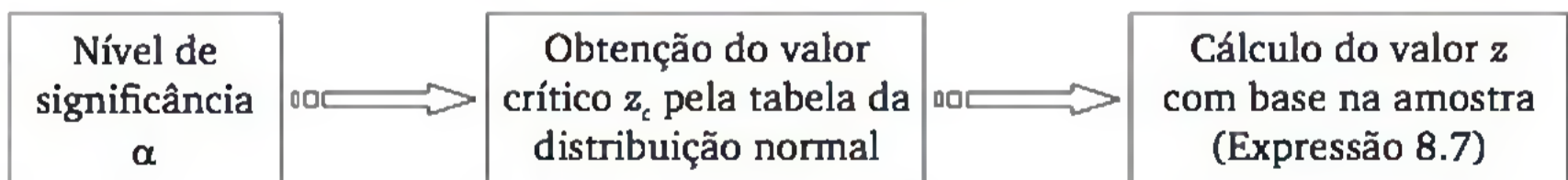


Figura 8.6 *Abordagem do valor p : regra de decisão de testes estatísticos usando o modelo normal padrão.*

A regra de decisão, como em todo teste estatístico feito com a abordagem do valor p , é dada por:



Abordagem clássica. O processo segue o esquema descrito abaixo:



⁷ Se o teste for unilateral à direita e $\hat{p} < p_0$, então não é necessário seguir em frente (o teste aceita H_0). O mesmo acontece quando o teste for unilateral à esquerda e $\hat{p} > p_0$.

Após a formulação das hipóteses e fixado o nível de significância α , podemos construir a regra de decisão, conforme mostra a Figura 8.7 e Tabela 8.1.

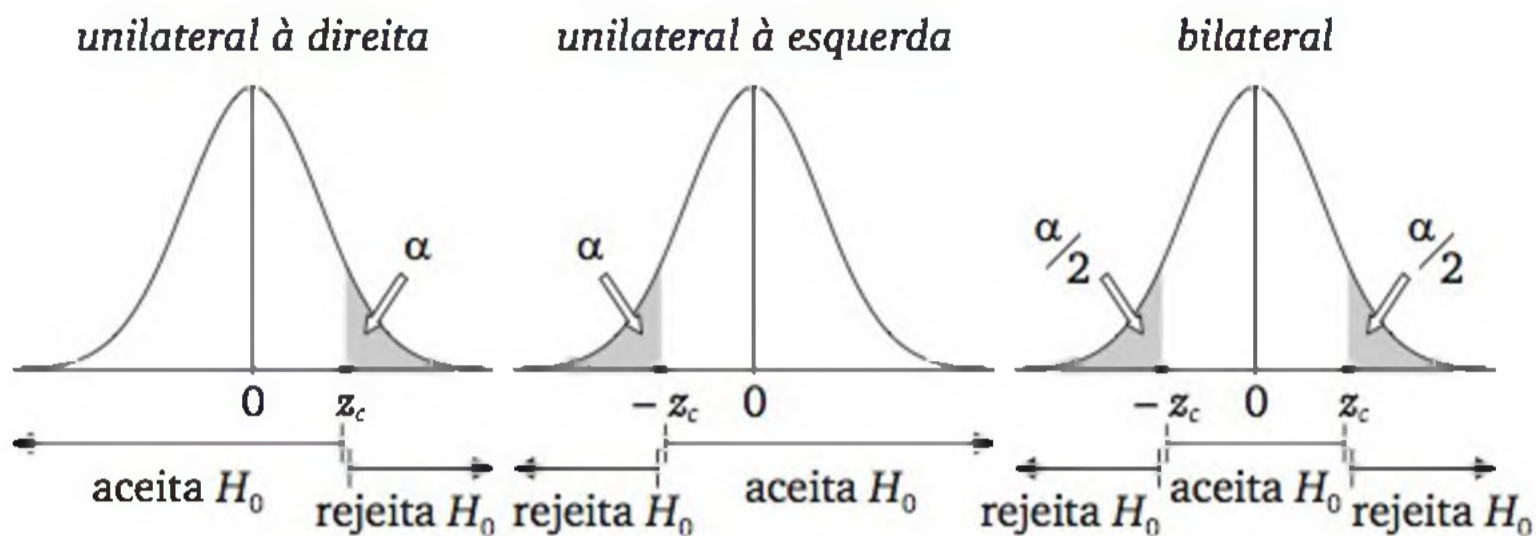


Figura 8.7 Abordagem clássica: regra de decisão de testes estatísticos usando o modelo normal padrão.

Tabela 8.1 Valores usuais de z_c , obtidos da distribuição normal padrão.

teste bilateral, α :	0,20	0,10	0,05	0,02	0,01	0,005
teste unilateral, α :	0,10	0,05	0,025	0,01	0,005	0,0025
valor crítico (z_c):	1,282	1,645	1,960	2,326	2,576	2,807

Com os dados da amostra, calculamos o valor de z , relativo à estatística do teste, e tomamos a decisão em termos da regra apropriada, conforme descrito na Figura 8.7.

Exemplo 8.6 Uma empresa retira periodicamente amostras aleatórias de 500 peças de sua linha de produção para análise da qualidade. As peças da amostra são classificadas como defeituosas ou não, sendo que a política da empresa exige que o processo produtivo seja revisto se houver evidência de mais que 1,5% de peças defeituosas. Na última amostra, foram encontradas nove peças defeituosas. Usando nível de significância de 1%, o processo precisa ser revisto?

Formulação das hipóteses: Como estamos procurando evidência de que a proporção de peças defeituosas é superior a 1,5%, é natural realizar um teste unilateral à direita, ou seja:

$$H_0: p = 1,5\% \quad (p_0 = 0,015)$$

$$H_1: p > 1,5\%$$

Cálculo da estatística do teste: Como foi observada uma amostra de $n = 500$ peças, donde foram encontradas $y = 9$ defeituosas, temos a proporção de peças defeituosas na amostra:

$$\hat{p} = \frac{9}{500} = 0,018$$

Os critérios para a aproximação normal são válidos, pois

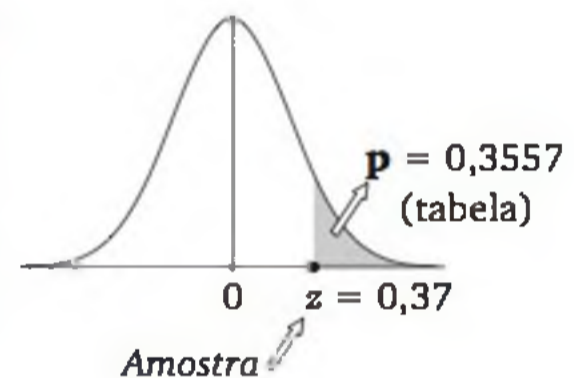
$$n \cdot p_0 = (500) \cdot (0,015) = 7,5 \text{ e}$$

$$n \cdot (1 - p_0) = (500) \cdot (0,985) = 492,5$$

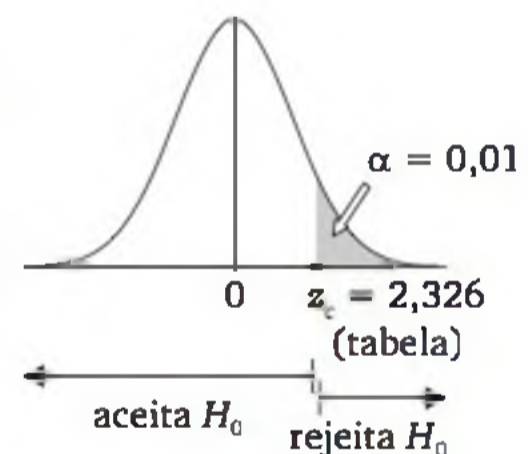
satisfazendo às condições (8.1) e (8.2). Como o valor da proporção amostral \hat{p} é superior ao descrito por H_0 ($p_0 = 0,015$), a correção de continuidade é dada por $y' = y - 0,5 = 8,5$ (ver Expressão 8.5). Logo:

$$z = \frac{y' - n \cdot p_0}{\sqrt{n \cdot p_0 \cdot (1 - p_0)}} = \frac{8,5 - (500) \cdot (0,015)}{\sqrt{(500) \cdot (0,015) \cdot (1 - 0,015)}} = \frac{1}{2,718} \approx 0,37$$

Solução pela abordagem do valor p: Usando a tabela normal padrão (Tabela 3 do apêndice), encontramos área na cauda superior igual a 0,3557. Como o teste é unilateral, este já é o valor da probabilidade de significância ($p = 0,3557$), o qual é maior que o nível de significância adotado ($\alpha = 0,01$). Então, o teste aceita H_0 . Veja a figura ao lado.



Solução pela abordagem clássica: O teste é unilateral à direita e foi adotado o nível de significância $\alpha = 0,01$. Pela tabela da distribuição normal padrão, o valor crítico é $z_c = 2,326$ (ver Tabela 8.1). Como a amostra acusou o valor $z = 0,37$, o qual está na região de aceitação, o teste aceita H_0 . Veja a figura ao lado.



Qualquer que seja a abordagem, o teste mostra que **não** há evidência de que a verdadeira proporção de peças defeituosas superou 1,5%. Como consequência, chegamos à conclusão de que não há provas estatísticas suficientes para recomendar a revisão do processo produtivo.

EXERCÍCIOS

12. Refaça os cálculos do Exercício 1, usando a distribuição normal. Compare os resultados.

- 13.** Seja p a probabilidade de *coroa* de uma moeda. Com o objetivo de testar $H_0: p = 0,5$ contra $H_1: p > 0,5$, fizeram-se 50 lançamentos dessa moeda, obtendo-se 31 coroas.
- O teste rejeita H_0 ao nível de significância de 5%?
 - E se estivéssemos trabalhando com o nível de significância de 1%?
- 14.** Com o objetivo de verificar se uma moeda está *viciada*, decide-se lançá-la várias vezes de forma imparcial e sempre sob as mesmas condições.
- Se em 8 lançamentos obtiverem-se 2 caras (e 6 coroas), qual é a conclusão ao nível de significância de 5%?
 - Se em 80 lançamentos obtiverem-se 20 caras (e 60 coroas), qual é a conclusão ao nível de significância de 5%?
- 15.** Para testar se um sistema computacional “inteligente” adquiriu algum conhecimento sobre determinado assunto, elaboraram-se 60 questões do tipo *certo-errado*. O sistema acertou 40. Qual é a conclusão ao nível de significância de 5%?
- 16.** Um experimento computacional foi repetido 40 vezes. O sistema A mostrou-se superior ao sistema B em 24 das 40 repetições. Há evidência suficiente para dizer que os sistemas têm desempenhos diferentes? Use $\alpha = 5\%$.
- 17.** Um fabricante garante que 90% de seus itens estão dentro das especificações. Um comprador examinou uma amostra aleatória de 50 itens e verificou que apenas 84% estavam dentro das especificações. Há evidência de que o nível de qualidade é menor do que o alegado pelo fabricante? Use $\alpha = 1\%$.

8.8 TESTE PARA MÉDIA

O teste para média é aplicável nas situações em que queremos verificar se uma variável na população pode ser considerada, *em média*, igual a certo valor μ_0 . No caso de o problema sugerir um teste bilateral, as hipóteses terão a forma:

$$H_0: \mu = \mu_0 \quad \text{e} \quad H_1: \mu \neq \mu_0$$

No caso de teste unilateral, a hipótese alternativa seria $H_1': \mu > \mu_0$ (unilateral à direita) ou $H_1'': \mu < \mu_0$ (unilateral à esquerda).

Consideraremos duas situações:

- quando existe alguma informação externa aos dados sobre a variância σ^2 da variável em estudo, na população (variância conhecida); e
- quando não há essa informação (variância desconhecida).

8.8.1 Caso de variância conhecida

Suponha que seja planejada uma amostragem aleatória da população de interesse. Considere que a amostra seja razoavelmente grande ($n \geq 30$) para valer o teorema do limite central; ou que os dados provenham de uma distribuição aproximadamente normal. O cálculo da estatística do teste é realizado por:

$$z = \frac{(\bar{x} - \mu_0) \cdot \sqrt{n}}{\sigma} \quad (8.8)$$

onde: μ_0 é o valor da média segundo H_0 ;

n é o tamanho da amostra;

σ é o desvio padrão populacional; e

\bar{x} é a média da amostra.

Com os dados da amostra, calculamos \bar{x} e z . Seja $|z|$ o valor absoluto de z . Com o auxílio de uma tabela da distribuição normal padrão, obtemos a área acima de $|z|$, dada por $1 - \Phi(|z|)$. Se o teste for unilateral e a relação entre \bar{x} e μ_0 for coerente com H_1 , então essa área já é o valor \mathbf{p} .⁸ Se o teste for bilateral, o valor \mathbf{p} corresponde ao dobro da área da cauda superior (veja a Figura 8.6). Obtido o valor \mathbf{p} , temos a regra de decisão:

Se $\mathbf{p} \leq \alpha$, então o teste rejeita H_0 ;

caso contrário, o teste aceita H_0 .

Alternativamente, podemos usar a abordagem clássica. Com os dados da amostra, calculamos a estatística z e tomamos a decisão em termos da regra da Figura 8.7.

Exemplo 8.7 Na indústria cerâmica, avalia-se sistematicamente a resistência de amostras de massas cerâmicas, após o processo de queima. Dessas avaliações, sabe-se que certo tipo de massa tem resistência mecânica aproximadamente normal, com média 53 MPa e variância 16 MPa². Após a troca de alguns fornecedores de matérias-primas, deseja-se verificar se houve alteração na qualidade. Uma amostra de 15 corpos de prova de massa cerâmica acusou média igual a 50 MPa. Qual é a conclusão ao nível de significância de 5%?

Realizaremos um teste bilateral, porque antes de a amostra ser observada, não sabíamos se a resistência deveria ser maior ou menor que 53. Assim,

⁸ Se o teste for unilateral à direita e $\bar{x} < \mu_0$, então não é necessário seguir em frente (o teste aceita H_0). Da mesma forma, quando o teste for unilateral à esquerda e $\bar{x} > \mu_0$.

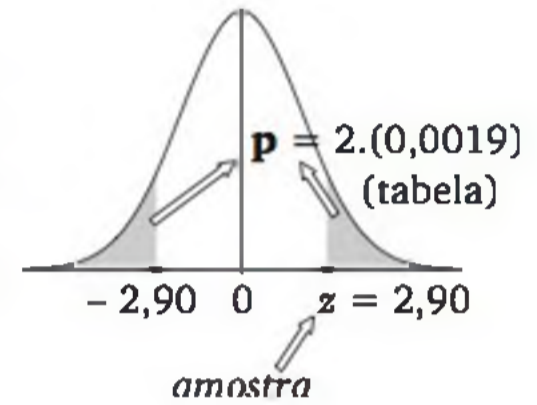
$$H_0: \mu = 53 \text{ MPa} \quad (\mu_0 = 53)$$

$$H_1: \mu \neq 53 \text{ MPa}$$

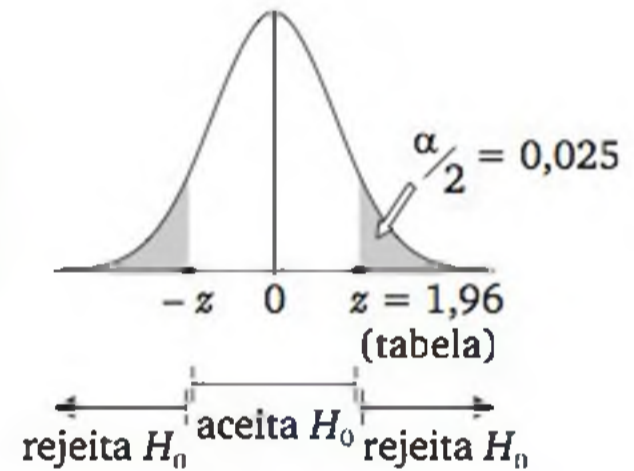
Cálculo da estatística do teste:

$$z = \frac{(\bar{x} - \mu_0) \cdot \sqrt{n}}{\sigma} = \frac{(50 - 53) \cdot \sqrt{15}}{\sqrt{16}} = -2,90$$

Solução pela abordagem do valor p : Usando a tabela normal padrão, encontramos área na cauda superior igual a 0,0019. Como o teste é bilateral, a probabilidade de significância é o dobro deste valor [$p = 2 \cdot (0,0019) = 0,0038$], que é menor do que $\alpha = 0,05$. Portanto, o teste rejeita H_0 . Veja a figura ao lado.



Solução pela abordagem clássica: Encontramos, na tabela normal padrão, o valor crítico $z_c = 1,96$ (ver Tabela 8.1). Como a amostra acusou o valor $z = -2,90$, o qual está na região de rejeição, o teste rejeita H_0 . Veja a figura ao lado.



Como consequência do resultado do teste estatístico, chegamos à conclusão de que **há** evidência de redução na resistência média da massa cerâmica.

8.8.2 Caso de variância desconhecida

Uma situação mais comum ocorre quando não se tem informação sobre a variância. Nesse caso, a variância é estimada com base nos dados da amostra e o cálculo da estatística do teste é feito por:

$$t = \frac{(\bar{x} - \mu_0) \cdot \sqrt{n}}{s} \quad (8.9)$$

onde: μ_0 é o valor da média, segundo H_0 ;

n é o tamanho da amostra;

\bar{x} é a média da amostra; e

s é o desvio padrão da amostra.

Para esse teste, a distribuição de referência é a t de *Student* com $gl = n - 1$ graus de liberdade (Tabela 4 do apêndice).⁹ Para a validade do teste, supõe-se que os dados provenham de uma distribuição aproximadamente normal, especialmente quando $n < 30$.

A realização do teste é feita de forma análoga aos casos anteriores, mas usando t no lugar de z , para se obter o valor \mathbf{p} ; ou t_c no lugar de z_c , para estabelecer a regra clássica de decisão.

Exemplo 8.8 O tempo para transmitir 10 MB em determinada rede de computadores varia segundo um modelo normal, com média 7,4 s e variância 1,3 s². Depois de algumas mudanças na rede, acredita-se numa redução no tempo de transmissão de dados, além de uma possível alteração na variabilidade. Foram realizados 10 ensaios independentes com um arquivo de 10 MB e foram anotados os tempos de transmissão, em segundos:

6,8 7,1 5,9 7,5 6,3 6,9 7,2 7,6 6,6 6,3

Existe evidência suficiente de que o tempo médio de transmissão foi reduzido? Use nível de significância de 1%.

As hipóteses são:

$$H_0: \mu = 7,4 \text{ s}$$

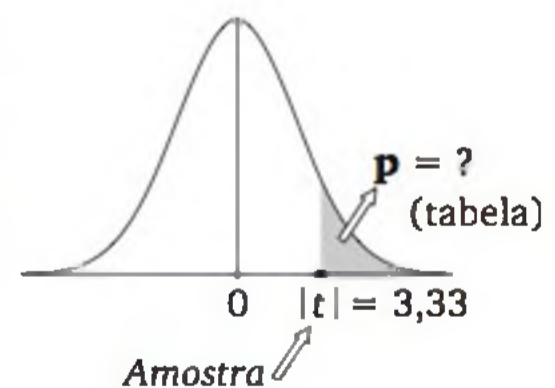
$$H_1: \mu < 7,4 \text{ s}$$

Considerando que a variabilidade possa ter sido alterada, vamos usar a variância da amostra. Fazendo os cálculos com as 10 observações, temos: $\bar{x} = 6,82$ e $s = 0,551$ (ver Expressões 3.4 e 3.8). E o cálculo da estatística do teste:

$$t = \frac{(\bar{x} - \mu_0) \cdot \sqrt{n}}{s} = \frac{(6,82 - 7,4) \cdot \sqrt{10}}{0,551} = -3,33$$

Temos $gl = n - 1 = 9$. Como a tabela fornece áreas na cauda superior e a distribuição t de *Student* é simétrica, usaremos o valor absoluto de t , ou seja: $|t| = 3,33$.

Solução pela abordagem do valor \mathbf{p} : Através da distribuição t de *Student* com $gl = n - 1 = 9$, devemos encontrar o valor \mathbf{p} , como ilustra o esquema ao lado. A Tabela 4 do apêndice permite obter este valor apenas de forma aproximada. Veja o esquema abaixo.

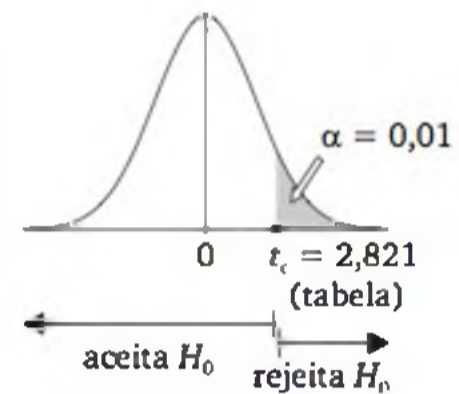


⁹ Para $n \geq 50$, as distribuições t de *Student* e normal padrão são bastante parecidas e, portanto, pode ser usada qualquer das duas.

		Área na cauda superior					
		0,25	...	0,01	0,005	0,0025	...
dados observados	↓						
$ t = 3,33$	→	3,250	3,690	...
		9	...	2,821			...

Ou seja, o valor $|t| = 3,33$ aponta para uma área na cauda superior entre 0,0025 e 0,005. Como o teste é unilateral, essa área é o próprio valor p , isto é, $0,0025 < p < 0,005$. E como $p < \alpha = 0,01$, o teste rejeita H_0 em favor de H_1 .

Solução pela abordagem clássica: o teste é unilateral e foi adotado o nível de significância $\alpha = 0,01$. Assim, encontramos, na tabela da distribuição t , o valor crítico $t_c = 2,821$. Como a amostra acusou $|t| = 3,33$, o qual está na região de rejeição, o teste rejeita H_0 em favor de H_1 . Veja a figura ao lado.¹⁰



Como consequência do resultado do teste estatístico, chegamos à conclusão de que *houve* redução no tempo médio de transmissão com as alterações na rede de computadores.

EXERCÍCIOS

18. Em certo banco de dados, o tempo para a realização das buscas é aproximadamente normal, com média de 53 s e desvio padrão de 14 s. Depois de realizadas algumas modificações no sistema, observou-se que, em 30 consultas, o tempo médio caiu para 45 s. Há evidência de melhora? Admita que as 30 observações possam ser consideradas uma amostra aleatória e que não houve alteração na variância. Use $\alpha = 1\%$.
19. Certo tipo de pneu dura, em média, 50.000 km. O fabricante investiu em uma nova composição de borracha para pneus. Vinte pneus, fabricados com essa nova composição, duraram, em média, 55.000 km, com desvio padrão de 4.000 km. Supondo que a durabilidade dos pneus segue uma distribuição aproximadamente normal, verificar se os dados provam que os novos pneus são mais duráveis. Use $\alpha = 1\%$.

¹⁰ Embora o teste seja unilateral à esquerda, usamos o valor de t em módulo, o que permitiu fazer a comparação do lado direito, em conformidade com a Tabela 4.

8.9 TESTE PARA VARIÂNCIA

Muitas vezes, há interesse em verificar possíveis alterações na variabilidade. Nesses casos, o teste pode ser feito com as hipóteses:

$$H_0: \sigma^2 = \sigma_0^2 \quad \text{e} \quad H_1: \sigma^2 \neq \sigma_0^2$$

No caso de teste unilateral, a hipótese alternativa seria $H_1': \sigma^2 > \sigma_0^2$ (unilateral à direita) ou $H_1'': \sigma^2 < \sigma_0^2$ (unilateral à esquerda).

A estatística do teste é calculada por:

$$q^2 = \frac{(n-1) \cdot s^2}{\sigma_0^2} \quad (8.10)$$

onde:

σ_0^2 é a variância segundo H_0 ;

n é o tamanho da amostra; e

s^2 é a variância da amostra.

Desde que os dados possam ser assumidos com distribuição normal, a distribuição de referência para o teste é a distribuição *qui-quadrado* com $gl = n - 1$ (Tabela 5 do apêndice). Um cuidado adicional é que a distribuição *qui-quadrado* não é simétrica, ou seja, a relação entre área na cauda e valor da abscissa precisa ser feita em cada lado da distribuição.

Exemplo 8.8 (continuação) Existe evidência suficiente de que as mudanças na rede de computadores alteraram a variabilidade no tempo de transmissão de dados? Use nível de significância de 5%.

As hipóteses são:

$$H_0: \sigma^2 = 1,3$$

$$H_1: \sigma^2 \neq 1,3$$

Como uma amostra de 10 observações produziu a variância $s^2 = 0,304$, temos:

$$q^2 = \frac{(n-1) \cdot s^2}{\sigma_0^2} = \frac{9 \cdot (0,304)}{1,3} = 2,10$$

Abordagem clássica: para construir a regra de decisão, precisamos obter os pontos críticos (χ_{c1}^2 e χ_{c2}^2), os quais separam áreas iguais a $\alpha/2$ em cada cauda da distribuição. Usando a Tabela 5 do apêndice, obtemos $\chi_{c1}^2 = 2,700$ e $\chi_{c2}^2 = 19,023$. Veja a Figura 8.8.

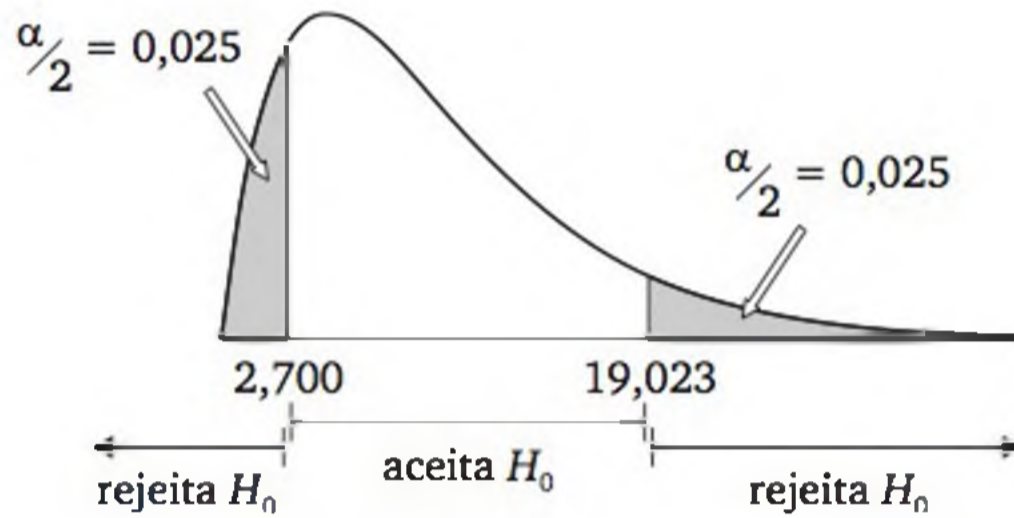


Figura 8.8 Regra de decisão de um teste bilateral usando o modelo qui-quadrado com $gl = 9$.

Como o resultado da estatística do teste ($q^2 = 2,10$) cai na região de rejeição, o teste rejeita H_0 em favor de H_1 . Ou seja, as mudanças realizadas na rede de computadores, além de reduzirem o tempo médio de resposta, também reduziram a variabilidade.

Abordagem do valor p: Podemos observar na Tabela 5 que o valor $q^2 = 2,10$ separa cerca de 1% de área na cauda inferior. Considerando que o teste é bilateral, temos $p \approx 2\%$. Como $p \leq \alpha = 5\%$, o teste rejeita H_0 .

EXERCÍCIOS

20. Usuários de uma rede de transmissão de energia elétrica têm reclamado da alta variação na tensão (desvio padrão de 12 V). A empresa encarregada da transmissão de energia elétrica na região instalou novos transformadores. O desvio padrão calculado sobre 30 observações independentes foi de 8 V e a distribuição de frequências dos valores da amostra sugere uma distribuição normal. Há evidência de redução na variação da tensão? Use $\alpha = 5\%$.
21. Com respeito ao Exercício 18, suponha que, nas 30 observações, o desvio padrão do tempo para a realização das buscas no banco de dados foi de 12 s. Há evidência de alteração na variância? Use $\alpha = 1\%$.

8.10 PODER DE UM TESTE E TAMANHO DA AMOSTRA

Como discutido anteriormente, em um teste de hipóteses temos as seguintes possibilidades:

Realidade (desconhecida)	Decisão do teste	
	Aceita H_0	Rejeita H_0
H_0 verdadeira	Decisão correta (probab = $1 - \alpha$)	Erro tipo I (probab = α)
H_0 falsa	Erro tipo II (probab = β)	Decisão correta (probab = $1 - \beta$)

Fixado o nível de significância α e construída a regra de decisão do teste, podemos estudar $P(\text{erro tipo II}) = P(\text{aceitar } H_0 \mid H_0 \text{ é falsa}) = \beta$.

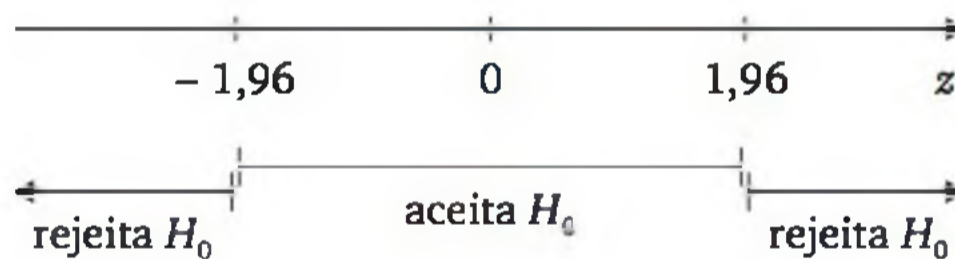
Vamos reconsiderar o Exemplo 8.7, onde tínhamos as hipóteses:

$$H_0: \mu = 53 \text{ e } H_1: \mu \neq 53$$

e, com base na estatística

$$z = \frac{(\bar{X} - \mu_0) \cdot \sqrt{n}}{\sigma} \quad (8.11)$$

construímos a regra de decisão ao nível de significância de 5%:



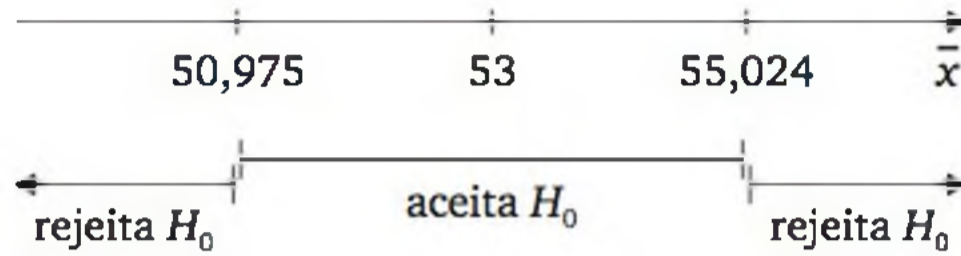
Analogamente, a regra de decisão pode ser colocada em termos da estatística \bar{X} . De (8.11), podemos obter os valores críticos por:

$$\bar{x}_c = \mu_0 \pm z_c \cdot \frac{\sigma}{\sqrt{n}} \quad (8.12)$$

Em termos do Exemplo 8.7, onde $n = 15$, $\sigma = 4$ e $\alpha = 0,05$, temos os valores críticos:

$$\bar{x}_{c1} = 53 - 1,96 \cdot \frac{4}{\sqrt{15}} = 50,975 \quad \text{e} \quad \bar{x}_{c2} = 53 + 1,96 \cdot \frac{4}{\sqrt{15}} = 55,024$$

Assim, a regra de decisão para \bar{x} é dada por:



Vamos considerar que a verdadeira média seja $\mu = 56$; portanto, H_1 é a hipótese verdadeira. Nesse caso, a probabilidade de se cometer o erro tipo II é dada por $\beta = P(\text{aceitar } H_0 \mid \mu = 56)$, ou seja,

$$\begin{aligned} \beta &= P\{50,975 \leq \bar{X} \leq 55,024 \mid \mu = 56\} = \\ &= P\left\{ \frac{(50,975 - 56) \cdot \sqrt{15}}{4} \leq \frac{(\bar{X} - 56) \cdot \sqrt{n}}{\sigma} \leq \frac{(55,024 - 56) \cdot \sqrt{15}}{4} \right\} = \\ &= P\{-4,865 \leq Z \leq -0,945\} = \Phi(-0,945) - \Phi(-4,865) \approx 0,17 \end{aligned}$$

Definimos **poder** de um teste estatístico como a probabilidade de o teste rejeitar H_0 quando H_0 é realmente falsa, ou seja, o poder de um teste é igual a $1 - \beta$.

No Exemplo 8.7, se na realidade $\mu = 56$, o poder de o teste detectar que H_0 é falsa é dado por

$$1 - \beta \approx 1 - 0,17 = 0,83$$

conforme ilustra a Figura 8.9.

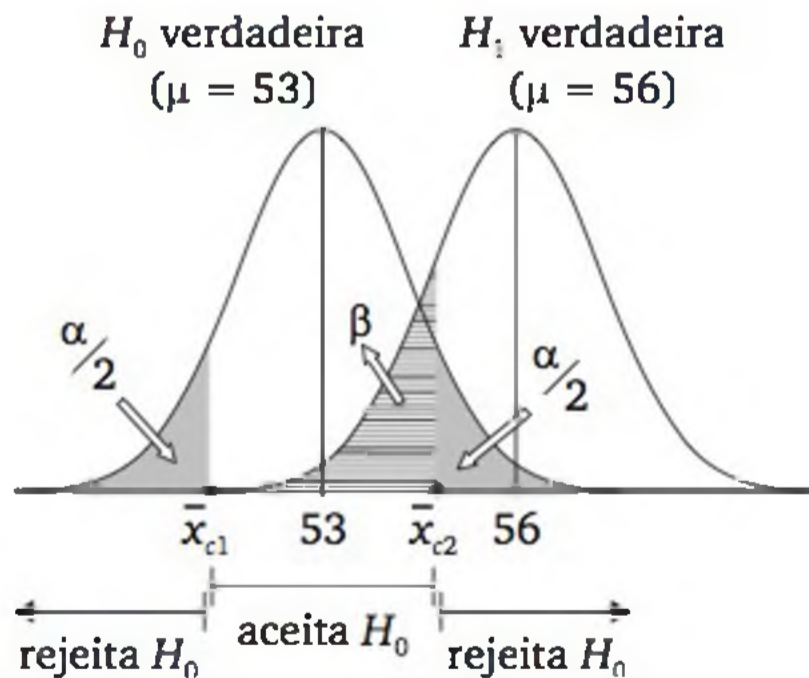


Figura 8.9 Ilustração do poder do teste do Exemplo 8.7, considerando $\alpha = 5\%$ e $\mu = 56$.

Mas a hipótese alternativa não é especificamente $\mu = 56$, mas $H_1: \mu \neq 53$. E, para cada valor de μ , o poder do teste $(1 - \beta)$ terá um valor diferente. Ou seja, o poder de um teste é, na verdade, uma função do parâmetro μ .¹¹ A Figura 8.10 ilustra curvas de poder de um teste bilateral.

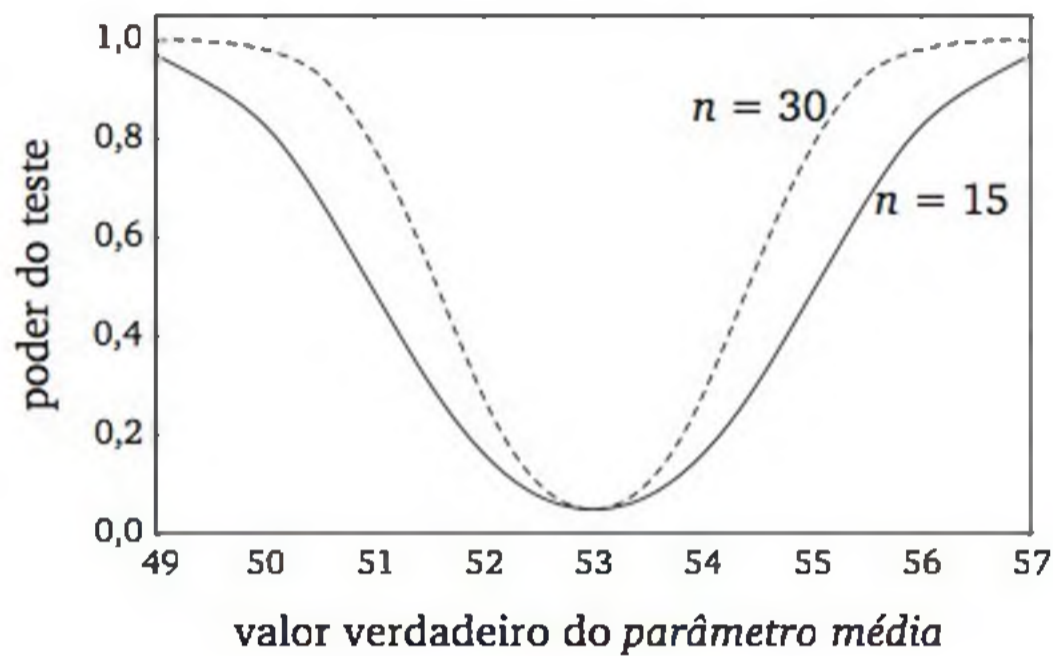


Figura 8.10 Curvas de poder do teste do Exemplo 8.7 para $n = 15$ e $n = 30$, considerando $\alpha = 5\%$.

Para o teste de média, o poder do teste depende da diferença $\mu - \mu_0$, do tamanho da amostra n , do tipo de teste (se bilateral ou unilateral) e do nível de significância α adotado. Observamos na Figura 8.10 que o poder do teste aumenta à medida que o verdadeiro parâmetro se distancia do valor alegado por H_0 e, também, quando se aumenta o tamanho da amostra. Dessa forma, se fixarmos o poder desejado para certo afastamento em relação ao alegado por H_0 , podemos calcular o tamanho da amostra necessário.

Teste para média com variância conhecida

Seja a distância entre a verdadeira média (μ) e o valor alegado por H_0 , em unidade de desvio padrão; ou seja:

$$\delta = \frac{|\mu - \mu_0|}{\sigma} \tag{8.13}$$

A probabilidade β pode ser calculada, em testes bilaterais, por:

$$\beta = \Phi\left(z_{\alpha/2} - \delta\sqrt{n}\right) - \Phi\left(-z_{\alpha/2} - \delta\sqrt{n}\right) \tag{8.14}$$

11 Mais diretamente, o poder do teste é uma função da diferença $\mu - \mu_0$.

onde: Φ é a função de distribuição acumulada normal padrão;

δ definido em (8.13); e

$z_{\alpha/2}$ é a abscissa da normal padrão que deixa $\alpha/2$ de área na cauda superior.¹²

O tamanho mínimo da amostra que garante poder igual a $(1 - \beta)$, em testes bilaterais, é dado aproximadamente por:

$$n \approx \left(\frac{z_{\alpha/2} + z_{\beta}}{\delta} \right)^2 \quad (8.15)$$

onde z_{β} é a abscissa da normal padrão que deixa β de área na cauda superior.

Para testes unilaterais,

$$n \approx \left(\frac{z_{\alpha} + z_{\beta}}{\delta} \right)^2 \quad (8.16)$$

Exemplo 8.9 As especificações de certas lâmpadas eletrônicas afirmam que elas resistem, em média, à tensão nominal de 127 V, com desvio padrão de 10 V. Um comprador desconfia dessa afirmação e resolve fazer um teste estatístico para testar as hipóteses $H_0: \mu = 127$ e $H_1: \mu < 127$, ao nível de significância de 5%. O comprador considera bastante grave se as lâmpadas resistirem, em média, apenas 120 V ou menos. Por isso, se $\mu = 120$, ele quer que o teste detecte que H_0 é falsa com 90% de probabilidade. Mas ele também não quer testar muitas lâmpadas porque o teste é destrutivo, causando perda financeira. Qual é o tamanho mínimo da amostra?

$$\text{Solução: } \delta = \frac{|\mu - \mu_0|}{\sigma} = \frac{|127 - 120|}{10} = 0,7$$

$$n \approx \left(\frac{z_{\alpha} + z_{\beta}}{\delta} \right)^2 = \left(\frac{1,645 + 1,282}{0,7} \right)^2 = 17,48$$

Logo, o tamanho mínimo da amostra deve ser $n = 18$ lâmpadas.

12 Em geral, uma das parcelas de (8.14) é aproximadamente nula.

Teste para média com variância desconhecida

Podemos calcular, aproximadamente, o tamanho mínimo da amostra através das expressões apresentadas no caso anterior, substituindo a abscissa z pela abscissa t da distribuição t de *Student* com $gl = n - 1$.¹³ Contudo, para o cálculo do tamanho da amostra, não se conhece, *a priori*, o valor de n para obter gl . Uma alternativa é fazer a amostragem em dois estágios, ou seja:

- retira-se, primeiramente, uma amostra-piloto de tamanho n_0 ;
- calcula-se o desvio padrão dessa amostra (s_0);
- obtém-se δ com s_0 no lugar de σ ;
- calcula-se o tamanho n da amostra usando $gl = n_0 - 1$; e
- completa-se a amostragem com mais $n - n_0$ observações (supondo $n > n_0$, pois, em caso contrário, a amostra-piloto já seria suficiente). Ver Exercício 23.

Teste para proporção

Em testes bilaterais ($H_0: p = p_0$ e $H_1: p \neq p_0$), o tamanho da amostra para garantir poder igual a $(1 - \beta)$ é dado, em função de p , por:

$$n = \left(\frac{z_{\alpha/2} \cdot \sqrt{p_0 \cdot (1 - p_0)} + z_{\beta} \cdot \sqrt{p \cdot (1 - p)}}{p - p_0} \right)^2 \quad (8.17)$$

Se o teste for unilateral, o cálculo é feito por:

$$n = \left(\frac{z_{\alpha} \cdot \sqrt{p_0 \cdot (1 - p_0)} + z_{\beta} \cdot \sqrt{p \cdot (1 - p)}}{p - p_0} \right)^2 \quad (8.18)$$

Exemplo 8.10 Uma indústria de cerâmicas admite que até 4% de seus itens podem conter defeitos leves. Um grande comprador resolve testar a garantia do fabricante considerando as hipóteses $H_0: p = 0,04$ e $H_1: p > 0,04$. Qual deve ser o tamanho mínimo da amostra para garantir um poder de 90% no teste se a verdadeira proporção for de 8%?

¹³ Formalmente, nesse caso deveríamos usar a chamada *distribuição t não central*, mas isso torna o problema demasiadamente complexo.

$$n = \left(\frac{z_{\alpha} \cdot \sqrt{p_0 \cdot (1 - p_0)} + z_{\beta} \cdot \sqrt{p \cdot (1 - p)}}{p - p_0} \right)^2 =$$

$$= \left(\frac{(1,645) \cdot \sqrt{0,04(0,96)} + (1,282) \cdot \sqrt{0,08(0,92)}}{0,08 - 0,04} \right)^2 = 280,69$$

Logo, o tamanho mínimo da amostra é $n = 281$.

EXERCÍCIOS

- 22.** Certa rede de computadores transmite dados a uma velocidade média de 200 MB/s, com desvio padrão de 30 MB/s. Algumas alterações estão sendo realizadas com o objetivo de aumentar a velocidade de transmissão de dados. Qual deve ser o tamanho mínimo da amostra para detectar um aumento na média de $0,5 \sigma$, com 90% de probabilidade? Use nível de significância de 1%.
- 23.** As embalagens de óleo de cozinha devem conter 900 ml de conteúdo líquido. Deseja-se fazer uma pesquisa com determinada marca para verificar se, em média, o conteúdo líquido não é menor do que o valor estipulado. Com nível de significância de 5%, qual é o tamanho de amostra necessário para identificar, com 90% de probabilidade, peso médio inferior a 892 ml? Considere que já foi realizada uma pesquisa preliminar com 8 unidades, obtendo desvio padrão igual a 10 ml.

EXERCÍCIOS COMPLEMENTARES

- 24.** Padrões técnicos exigem que o nível de ruído em CPDs seja de, no máximo, 70 dB. Foram analisados 16 CPDs de várias organizações, obtendo-se os seguintes valores máximos de ruído:
- 78 73 68 65 72 64 77 80 82 78 65 72 61 79 58 65
- a)** Calcule a intensidade de ruído média e o desvio padrão para esses 16 CPDs.
- b)** Suponha que os 16 CPDs analisados são uma amostra aleatória de CPDs do país. Para verificar se na média os CPDs atendem aos padrões técnicos, como você construiria as hipóteses?

- c)** Você pode concluir que a intensidade de ruído média dos CPDs nos horários críticos é superior ao especificado? Faça o teste adequado ao nível de significância de 5%.
- d)** Sob o ponto de vista dos que trabalham nos CPDs, qual é o pior erro? Explique.
- e)** Se a verdadeira intensidade média de ruído dos CPDs fosse de 73 dB, qual é a probabilidade de você tomar uma decisão errada no teste do item (c)? Suponha que o verdadeiro desvio padrão é 7 dB.
- 25.** Um cliente de uma torrefação de café suspeita que os pesos dos pacotes, que deveriam ser de 500 g, não estão corretos. Resolveu, então, retirar uma amostra dos pesos de 16 pacotes (supondo que provenham de uma distribuição normal):
- 510, 495, 498, 500, 501, 499, 503, 500, 495, 492, 499, 499, 497, 495, 499, 501
- a)** Calcule o peso médio e o desvio padrão dos elementos da amostra.
- b)** O cliente tem razão na suspeita? Use $\alpha = 0,05$.
- c)** Pode-se afirmar, com nível de significância de 10%, que a variância do processo é superior a 10 g^2 ?
- 26.** O tempo médio de vida de um tipo de lâmpada é de 2.000 horas e desvio padrão de 200 horas, segundo o fabricante. Afirma também que esse tempo médio de vida segue uma distribuição normal. Um possível comprador resolve retirar uma amostra aleatória de 15 lâmpadas. Testando-as, obteve média de 1.950 horas.
- a)** A 5% de significância, há evidência de que a afirmação do fornecedor quanto à média é falsa (prejudicando o comprador)?
- b)** Assumindo um risco de 5% para o fornecedor, a amostra retirada é suficiente para que o comprador tenha um risco de 10% de que a média do tempo de vida esteja 50 horas abaixo do especificado? Qual deveria ser o valor de n ?
- c)** Se o tempo de vida médio real fosse de 1.950 horas, qual é a probabilidade de se tomar uma decisão errada no item (a)?
- 27.** O controle estatístico de certo processo estabeleceu que pelo menos 94% dos produtos têm que estar sem defeitos. Para verificar a validade desta afirmação, foi coletada uma amostra de 150 produtos, obtendo uma proporção sem defeito igual a 92%.
- a)** Com 1% de significância, há evidência de que o processo está em desacordo com o esperado?

- b)** Se o percentual real sem defeito fosse 91%, qual é a probabilidade de se tomar uma decisão errada no item (a)?
 - c)** Suponha que se queira identificar, com 95% de probabilidade, a falsidade de H_0 , quando a proporção de sem defeito for, na realidade, igual a 93%. Considerando que o teste será realizado com 1% de significância, qual é o tamanho de amostra necessário?
- 28.** O resíduo da queima de carvão mineral (cinza) pode ser usado na composição do cimento pozolânico. Deseja-se verificar se a substituição da verdadeira rocha pozolânica por cinza de carvão altera a resistência à compressão do cimento, após 28 dias de hidratação. Suponha que esse cimento com a verdadeira rocha pozolânica tem resistência média de 40 MPa e desvio padrão de 4 MPa (após 28 dias de hidratação). Use $\alpha = 0,05$.
- a)** Quantos corpos de prova devem ser usados para que seja detectada uma redução média de 3 MPa, com 95% de probabilidade?
 - b)** Se foi feito um estudo experimental com o tamanho de amostra calculado no item (a), encontrando $\bar{x} = 39,2$ e $s = 6$ MPa, pode-se dizer que houve redução na resistência média?
 - c)** Houve aumento na variabilidade? (Suponha que os dados provenham de distribuição normal.)

Comparação Entre Tratamentos

É muito comum o interesse em comparar dois ou mais *tratamentos*, como, por exemplo, dois processos de têmpera na produção de aço, três tipos de cimento-e-cola para fixar azulejos, dois sistemas computacionais para a informatização de um processo etc. Para realizar as comparações, podemos planejar experimentos com amostras submetidas a cada tratamento. Em cada ensaio, observamos uma *resposta* adequada. Por exemplo, no caso do aço, a resposta pode ser a resistência mecânica; no cimento-e-cola, o grau de aderência; no sistema computacional, o tempo de resposta; e assim por diante.

Na comparação dos tratamentos, é natural o interesse em verificar se há evidência de diferenças entre os efeitos dos tratamentos, o que pode ser feito através de testes estatísticos. Neste capítulo, apresentaremos *testes paramétricos*, os quais se caracterizam por suporem certa distribuição de probabilidades para a variável resposta. Em consequência, a comparação entre os efeitos dos tratamentos pode ser feita em termos dos parâmetros da distribuição de probabilidades suposta. Nós nos restringiremos à comparação de parâmetros de distribuições normais, com ênfase em testes de médias.

9.1 AMOSTRAS INDEPENDENTES E EM BLOCOS

Na comparação entre g tratamentos, muitas vezes o experimento pode ser conduzido dividindo-se aleatoriamente as unidades experimentais em g grupos, sendo cada grupo submetido a um tratamento – *projeto de experimento comple-*

tamente aleatorizado. Como resultado da aplicação desse projeto, temos g amostras independentes.

Alternativamente, podemos construir h blocos de unidades experimentais relativamente similares. Se em cada bloco temos g unidades experimentais, alocamos, por sorteio, todos os g tratamentos em cada bloco – projeto de experimento em blocos aleatorizados. No caso de $g = 2$, o resultado da aplicação desse projeto leva a dados pareados.

Exemplo 9.1 Considere o problema de comparar dois materiais (A e B), para sola de tênis, em termos do grau de desgaste após certo período de uso.¹ Seguem dois projetos de experimentos alternativos:

Projeto I: Um grupo de indivíduos usa tênis com solas feitas com o material A ; e outro grupo usa tênis com solas feitas com o material B , conforme ilustra a Figura 9.1.

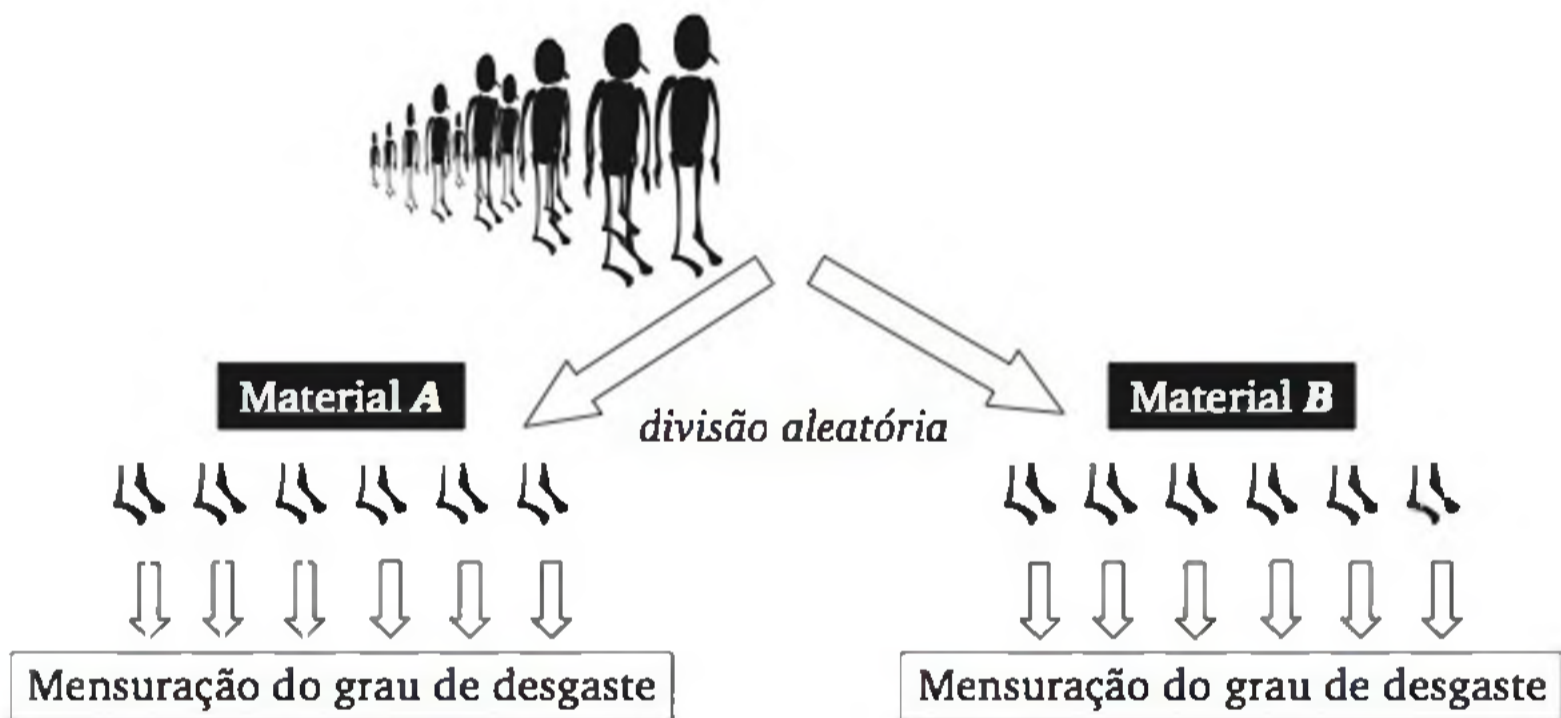


Figura 9.1 Esquema de planejamento de experimento completamente aleatorizado com $g = 2$ grupos. Em cada grupo é alocado um tratamento.

Projeto II: Fabricam-se, para a realização do experimento, pares de tênis com os dois tipos de sola, isto é, um dos pés com o material A e o outro pé com o material B . Em cada par, o material usado em cada pé (direito ou esquerdo) é decidido por sorteio (ver Figura 9.2).

¹ Baseado em exemplo do livro de BOX, G. E. P.; HUNTER, W. G.; HUNTER, J. S. *Statistics for experimenters*. New York: John Wiley, 1978.

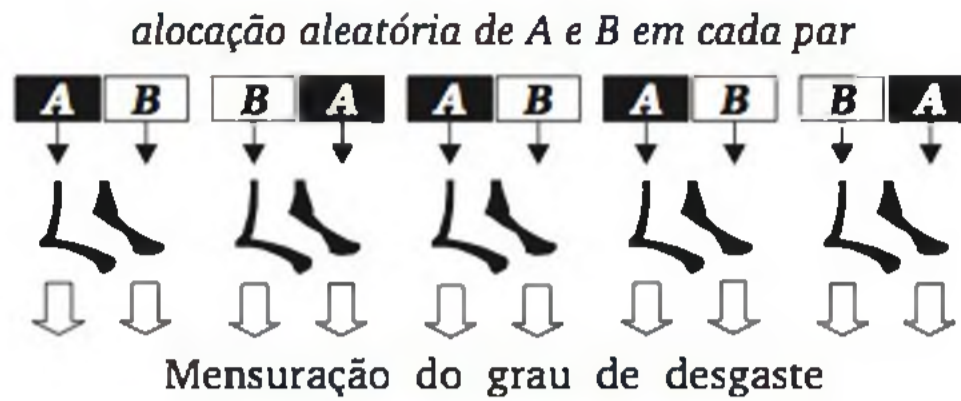


Figura 9.2 Esquema de planejamento de experimento em blocos aleatorizados, com $g = 2$ tratamentos e $h = 5$ blocos.

Para a análise dos resultados do Projeto II, a Figura 9.3 ilustra duas situações: (a) quando desconsideramos os pares, tratando os dados como se fossem duas amostras independentes; e (b) quando analisamos os pares.

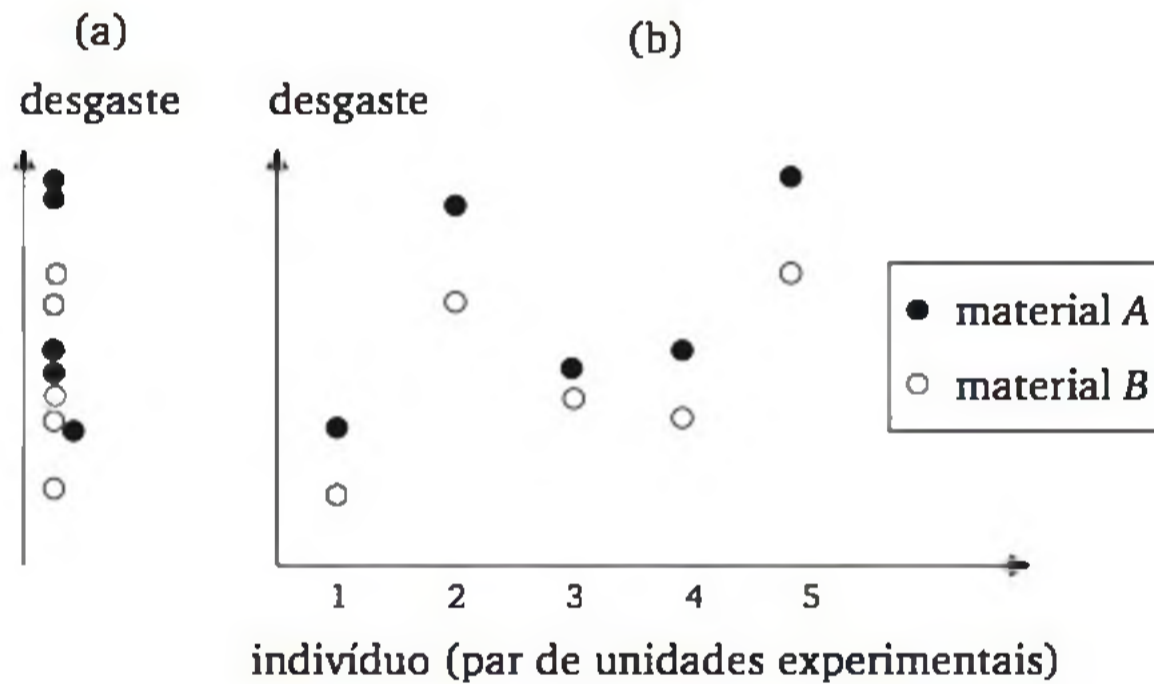


Figura 9.3 Ilustração de um conjunto de dados visto de forma pareada (à direita) e de forma independente (à esquerda).

Analisando a Figura 9.3, fica evidente que, ao olharmos os dados de forma pareada, temos mais informação sobre uma possível diferença entre os dois materiais – os pares até o quinto indivíduo mostraram que o material A teve maior nível de desgaste. Observando as amostras de forma independente, as diferenças entre os dois materiais ficaram ofuscadas pelas diferenças entre os indivíduos.

Em suma, para o particular problema, o Projeto II destaca melhor uma possível diferença entre os materiais. Em geral, quando é possível formar *blocos de unidades relativamente similares*, temos um projeto de experimento melhor.

Na prática, porém, muitas vezes não temos total liberdade de escolher o projeto de experimento mais adequado, seja por questões financeiras, seja por-

que os grupos já estão naturalmente divididos, como no caso de comparar alguma característica entre homens e mulheres (amostras independentes). Ou, ainda, quando queremos comparar o efeito de certo evento, em que é natural observar as unidades experimentais antes e depois do evento, resultando em dados pareados.

9.2 TESTE t PARA DUAS AMOSTRAS PAREADAS

O chamado *teste t* é apropriado para comparar dois conjuntos de dados quantitativos, em termos de seus valores médios. Mais especificamente:

$$H_0: \mu_1 = \mu_2 \text{ e } H_1: \mu_1 \neq \mu_2$$

onde: μ_1 é o valor esperado da resposta sob o tratamento 1 e
 μ_2 é o valor esperado da resposta sob o tratamento 2.

Na abordagem unilateral, a hipótese alternativa é do tipo H_1' : $\mu_1 > \mu_2$ ou H_1'' : $\mu_1 < \mu_2$.

Nesta seção, trataremos do caso em que os dois conjuntos de dados são pareados. Contudo, as observações entre os pares devem ser independentes.

Exemplo 9.2 Seja o problema de verificar se um novo algoritmo de busca em um banco de dados é mais rápido que o algoritmo atualmente usado. Para fazer a comparação dos dois algoritmos, planeja-se realizar uma amostra aleatória de dez buscas experimentais (ensaios). Em cada ensaio, uma dada busca é realizada pelos dois algoritmos e o tempo de resposta de cada algoritmo anotado. Observamos que em cada ensaio os dois algoritmos são usados em condições idênticas, caracterizando dez pares de observações.

As hipóteses podem ser formuladas da seguinte maneira:

H_0 : em média, os dois algoritmos são *igualmente* rápidos e

H_1 : em média, o algoritmo novo é *mais* rápido do que o algoritmo em uso.

ou

$$H_0: \mu_1 = \mu_2 \text{ e } H_1: \mu_1 < \mu_2$$

onde: μ_1 é o tempo esperado de resposta do algoritmo novo e

μ_2 é o tempo esperado de resposta do algoritmo atualmente usado.

Suponha que o experimento tenha sido realizado, gerando os dados apresentados na Tabela 9.1.

Tabela 9.1 *Tempos de resposta dos algoritmos de busca 1 e 2, em dez ensaios pareados.*

Ensaio	Tempo de resposta (s)		
	Novo X_1	Antigo X_2	Diferença $D = X_2 - X_1$
1	22	25	3
2	21	28	7
3	28	26	- 2
4	30	36	6
5	33	32	- 1
6	33	39	6
7	26	28	2
8	24	33	9
9	31	30	- 1
10	22	27	5

Como os dados são pareados, podemos verificar em cada ensaio o quanto um tratamento (algoritmo) foi melhor do que o outro, ou seja, analisar a variável:

$$D = X_2 - X_1 \quad (9.1)$$

Em termos da variável diferença D , as hipóteses são descritas como:

$$H_0: \mu_D = 0 \text{ e } H_1: \mu_D > 0$$

onde μ_D é o valor esperado de D . E o problema torna-se semelhante ao teste de uma média, discutido na seção 8.9. Assim, dada uma amostra, calculamos a estatística do teste por:

$$t = \frac{\bar{d} \cdot \sqrt{n}}{s_d} \quad (9.2)$$

onde: n é o tamanho da amostra (número de pares);

\bar{d} é a média das diferenças observadas; e

s_d é o desvio padrão das diferenças observadas.²

² Note que estamos considerando o tamanho da amostra (n) como sendo igual ao número de pares (h), para evidenciar que esse teste é um caso especial do teste de uma média.

Supondo que os valores de D provenham de distribuição aproximadamente normal, o teste pode ser realizado com a distribuição t de Student com $gl = n - 1$ graus de liberdade.

Exemplo 9.2 (continuação) Valores de D (última coluna da Tabela 9.1):

3, 7, -2, 6, -1, 6, 2, 9, -1, 5

Donde:

$$n = 10, \bar{d} = 3,4 \text{ e}$$

$$s_d = \sqrt{\frac{1}{n-1} \cdot \left(\sum_i d_i^2 - n \cdot \bar{d}^2 \right)} = \sqrt{\frac{246 - (10)(3,4)^2}{9}} = 3,81$$

E, portanto:

$$t = \frac{\bar{d} \cdot \sqrt{n}}{s_d} = \frac{3,4 \cdot \sqrt{10}}{3,81} = 2,82$$

Abordagem do valor p : Como $n = 10$, temos $gl = 9$ graus de liberdade. Tomemos, então, a linha de $gl = 9$ da Tabela 4 do apêndice (tabela da *distribuição t de Student*), como mostra a Figura 9.4. Por essa tabela, obtemos a área associada a um valor maior ou igual a $t = 2,82$. Como o teste é unilateral, essa área corresponde ao valor p .

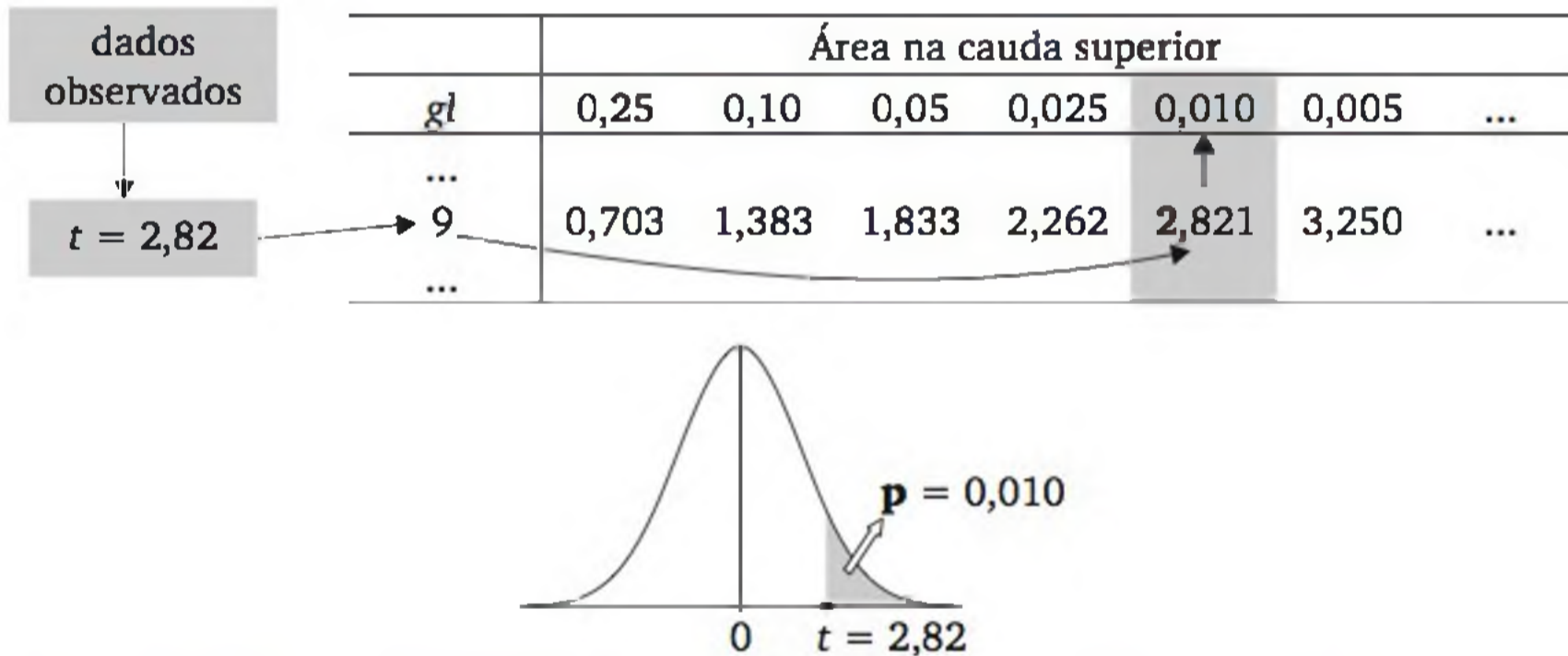
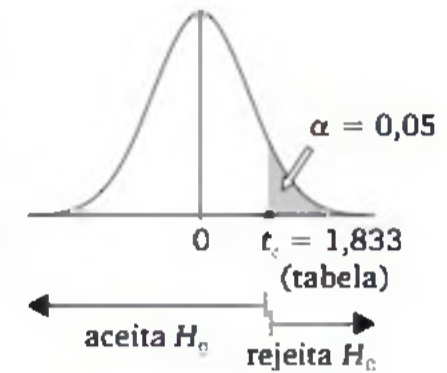


Figura 9.4 Uso da distribuição t de Student com $gl = 9$ para a obtenção do valor p , num teste unilateral, com $n = 10$ e $t = 2,82$.

Observando a linha correspondente a $gl = 9$, verificamos, na tabela, que o valor $t = 2,82$ (calculado com base na amostra) está próximo do valor tabulado 2,821. Logo, como ilustra a Figura 9.4, temos $p \approx 0,010$.

Considerando o nível de significância de 5% ($\alpha = 0,05$), o teste leva a conclusão de que os dados mostram evidência suficiente de que H_0 é falsa (pois $p < \alpha = 0,05$), detectando, então, que o algoritmo novo é, em média, mais rápido do que o algoritmo atualmente em uso. Essa mesma conclusão pode ser obtida se usarmos a abordagem clássica, conforme descrita a seguir.

Abordagem clássica: Nesse caso, busca-se o valor crítico t_c , que deixa área $\alpha = 0,05$ na cauda superior da distribuição t de student com $gl = 9$. Pela Tabela 4 do apêndice temos $t_c = 1,833$, levando à regra de decisão apresentada ao lado, a qual pode ser construída mesmo antes de se observar a amostra.



Como os dados produziram o valor $t = 2,82$, então o teste rejeita H_0 ao nível de significância de 5%.

9.3 TESTE t PARA DUAS AMOSTRAS INDEPENDENTES

Nesta seção, apresentaremos como realizar o teste t quando as amostras são independentes.

Exemplo 9.3 Desejamos verificar se os catalisadores A e B têm efeitos diferentes no rendimento de certa reação química. As hipóteses são:

H_0 : em média, os dois catalisadores são *iguais* em termos de rendimento; e

H_1 : em média, os dois catalisadores são *diferentes* em termos de rendimento.

Ou, ainda:

$$H_0: \mu_1 = \mu_2 \text{ e } H_1: \mu_1 \neq \mu_2,$$

onde

μ_1 : rendimento esperado com o catalisador A; e

μ_2 : rendimento esperado com o catalisador B.

Para testar essas hipóteses, realizamos dez ensaios com cada catalisador, em ordem aleatória. Como ressaltava R. A. Fisher, que construiu as bases da estatística experimental: *a aleatorização dos grupos é fundamental para resguardar a validade de um teste de significância.*

A Tabela 9.2 mostra os resultados do experimento e a Figura 9.5 apresenta o diagrama de pontos de cada amostra.

Tabela 9.2 Rendimentos (%) de uma reação química em função do catalisador utilizado.

Catalisador A					Catalisador B				
45	51	50	62	43	45	35	43	59	48
42	53	50	48	55	45	41	43	49	39

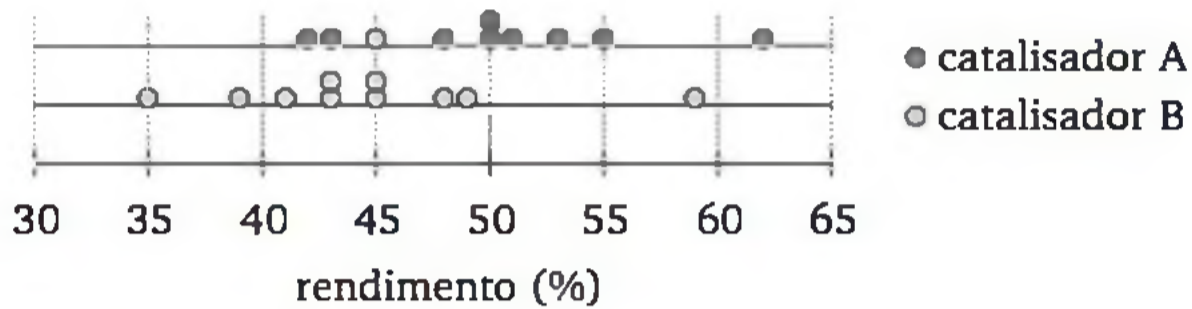


Figura 9.5 Diagrama de pontos dos resultados do experimento.

A análise descritiva das amostras sugere que o catalisador A é melhor, mas precisamos fazer um teste estatístico apropriado para verificar se essa diferença não pode ser explicada por mero acaso do particular experimento realizado.

Estatística do teste para amostras de tamanhos iguais

Dada as duas amostras, a estatística do teste toma como base a diferença entre as médias, $\bar{x}_1 - \bar{x}_2$, mas leva também em consideração o número de elementos em cada amostra e a variabilidade interna das amostras. Quanto maior as amostras, maior a evidência de uma possível diferença real. Por outro lado, se há muita variabilidade entre os elementos de cada amostra, diferenças reais podem ficar nebulosas (veja a Figura 9.6).

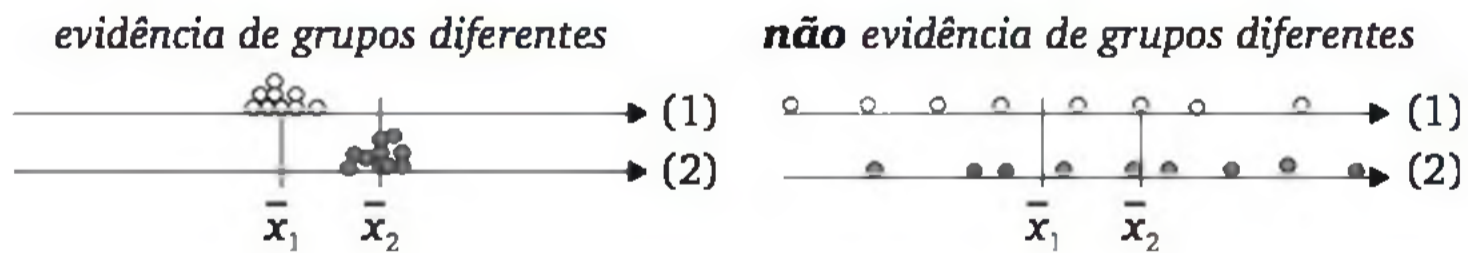


Figura 9.6 Importância de se considerar a variância interna dos grupos.

Considerando o mesmo número de elementos, n , em cada amostra, a chamada *variância agregada* é obtida pela média aritmética das variâncias de cada grupo, ou seja:

$$s_a^2 = \frac{s_1^2 + s_2^2}{2} \tag{9.3}$$

E a estatística do teste é calculada por:

$$t = (\bar{x}_1 - \bar{x}_2) \cdot \sqrt{\frac{n}{2s_a^2}} \quad (9.4)$$

onde

n : tamanho da amostra em cada grupo;

\bar{x}_1 : média da amostra 1; \bar{x}_2 : média da amostra 2;

s_1^2 : variância da amostra 1; s_2^2 : variância da amostra 2; e

s_a^2 : variância agregada das duas amostras.

Estatística do teste para amostras de tamanhos diferentes

No caso de as amostras terem tamanhos diferentes ($n_1 \neq n_2$), os cálculos devem ser feitos por:

$$s_a^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \quad (9.5)$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_a \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (9.6)$$

Suposições básicas para a aplicação do teste:

1. as observações devem ser independentes;
2. as variâncias populacionais devem ser iguais nos dois grupos;
3. os dois conjuntos de dados devem provir de distribuições normais.

A suposição (1) refere-se ao planejamento do experimento, enquanto as suposições (2) e (3) à variável e às populações em estudo.³ Na prática, não é fácil verificar a veracidade dessas suposições. Aconselhamos, contudo, fazer uma análise exploratória, tal como um diagrama de pontos para cada amostra. Esses gráficos permitem avaliar se existem fortes violações das suposições, tais como

³ Observamos que o teste t é razoavelmente robusto às suposições (2) e (3), isto é, o teste somente deixará de ser válido se houver violações fortes dessas suposições, como a presença de valor discrepante, distribuições muito assimétricas ou uma variância muito superior à outra.

a presença de pontos discrepantes, distribuições com formas assimétricas ou, ainda, uma distribuição bem mais dispersa do que a outra.⁴

Distribuição de referência. Se as médias populacionais forem iguais (H_0 verdadeira) e as suposições básicas puderem ser admitidas, então usaremos a distribuição t de *Student*, com $gl = n_1 + n_2 - 2$ graus de liberdade, para a realização do teste.

Exemplo 9.3 (continuação)

Amostra 1: $n = 10, \bar{x}_1 = 49,900$ e $s_1^2 = 35,656$

Amostra 2: $n = 10, \bar{x}_2 = 44,700$ e $s_2^2 = 42,233$

Variância agregada: $s_a^2 = \frac{s_1^2 + s_2^2}{2} = \frac{35,656 + 42,233}{2} = 38,945$

Resultado da estatística do teste:

$$t = (\bar{x}_1 - \bar{x}_2) \cdot \sqrt{\frac{n}{2s_a^2}} = (49,90 - 44,70) \cdot \sqrt{\frac{10}{2(38,94)}} = (5,2) \cdot \sqrt{0,1284} = 1,86$$

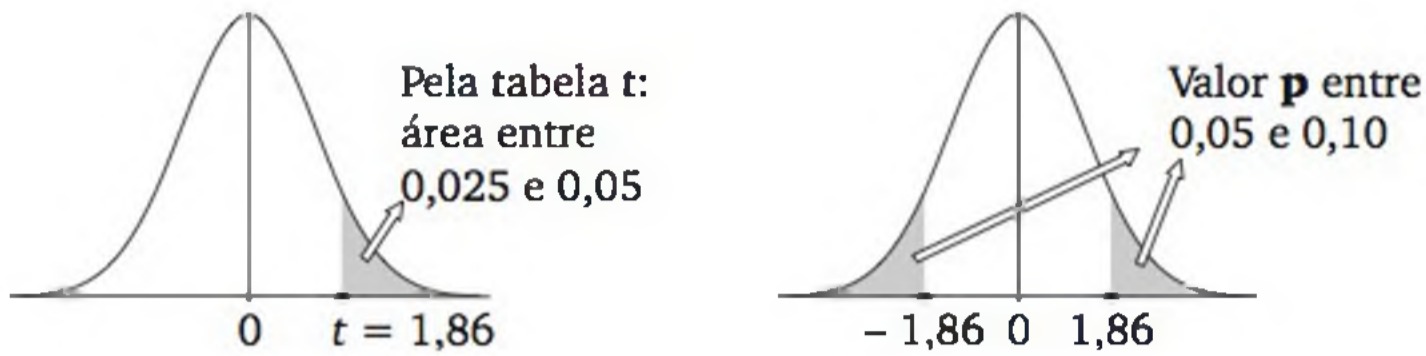
Graus de liberdade: $gl = 2n - 2 = 2(10) - 2 = 18$.

Abordagem do valor p : O esquema seguinte ilustra o uso da Tabela 4 do apêndice para se obter o valor p associado ao t calculado.

dados observados	Área na cauda superior							
	gl	0,25	0,10	0,05	0,025	0,010	0,005	...
$t = 1,86$...			↑	↑			
	18	0,688	1,330	1,734	2,101	2,552	2,878	...
	...							

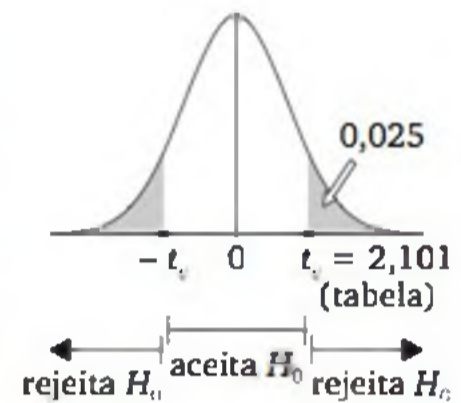
Os dados observados levaram ao valor $t = 1,86$, apontando para uma área na cauda superior da curva entre 0,025 e 0,05. Mas, como o teste é bilateral ($H_1: \mu_1 \neq \mu_2$), a área deve ser dobrada para se ter o valor p correto. Veja o esquema a seguir:

4 Podemos, ainda, realizar um teste de igualdade de variâncias (seção 9.5) e um teste de aderência à distribuição normal (Capítulo 10).



Portanto, $0,05 < p < 0,10$, o que leva à aceitação de H_0 ao nível de significância de 5% (pois $p > \alpha = 0,05$).

Abordagem clássica: Mesmo antes de realizar o experimento, podemos buscar na tabela da distribuição t de *student* com $gl = 18$ o valor crítico t_c , o qual deixa uma área igual a $\alpha = 0,05/2 = 0,025$ em cada calda da distribuição (pois o teste é bilateral). Pela Tabela 4, temos $t_c = 2,101$, levando à regra de decisão apresentada ao lado.



Como os dados produziram o valor $t = 1,86$, o qual pertence à região de aceitação, o teste aceita H_0 ao nível de significância de 5%.

Concluimos, então, ao nível de significância de 5%, que os dados não comprovam uma diferença entre os dois catalisadores. Existe uma probabilidade razoável (superior a 5%) de que as diferenças observadas nos dados experimentais são provenientes de fatores casuais.

9.4 TAMANHO DAS AMOSTRAS

No planejamento de um experimento para comparar dois tratamentos, surge a questão de qual deve ser o número n de ensaios para cada tratamento. Para responder a essa questão, vamos relembrar alguns conceitos. Quando o teste rejeita a hipótese de igualdade entre os tratamentos (H_0), concluindo que existem diferenças significativas entre eles, podemos estar cometendo o chamado erro tipo I: rejeitar H_0 quando verdadeira. Os testes são construídos com a probabilidade deste erro fixada num nível bastante baixo, designada por α (nível de significância do teste) e usualmente $\alpha = 0,05$. Por outro lado, quando o teste aceita H_0 , pode ocorrer o chamado erro tipo II: aceitar H_0 quando falsa. A probabilidade de se cometer este erro é designada por β . É desejável que, quando a diferença entre os tratamentos for grande em termos práticos, a probabilidade β seja pequena e, para que isso aconteça, a quantidade n de elementos em cada grupo deve ser suficientemente grande.

A discussão que segue restringe-se ao problema de comparar duas amostras independentes em termos de médias. Sejam μ_1 e μ_2 as médias das duas populações em estudo e seja

$$\delta = \frac{|\mu_1 - \mu_2|}{\sigma} \quad (9.7)$$

A quantidade δ é a diferença de magnitude entre as verdadeiras médias em unidade de desvio padrão (σ). Supomos aqui que as duas populações tenham o mesmo desvio padrão.

Para avaliar a quantidade n de elementos em cada grupo, o pesquisador precisa ser capaz de fornecer o valor mínimo de δ que leva a consequências práticas. Em geral, o pesquisador tem maior facilidade em raciocinar em termos da unidade em que se está medindo a variável em análise, mas, nesse caso, torna-se necessário ter uma avaliação de σ .

A Figura 9.7 indica o número mínimo n para que uma diferença δ seja detectada pelo teste estatístico com probabilidade 0,80 ($\beta = 0,20$) e com probabilidade 0,90 ($\beta = 0,10$).

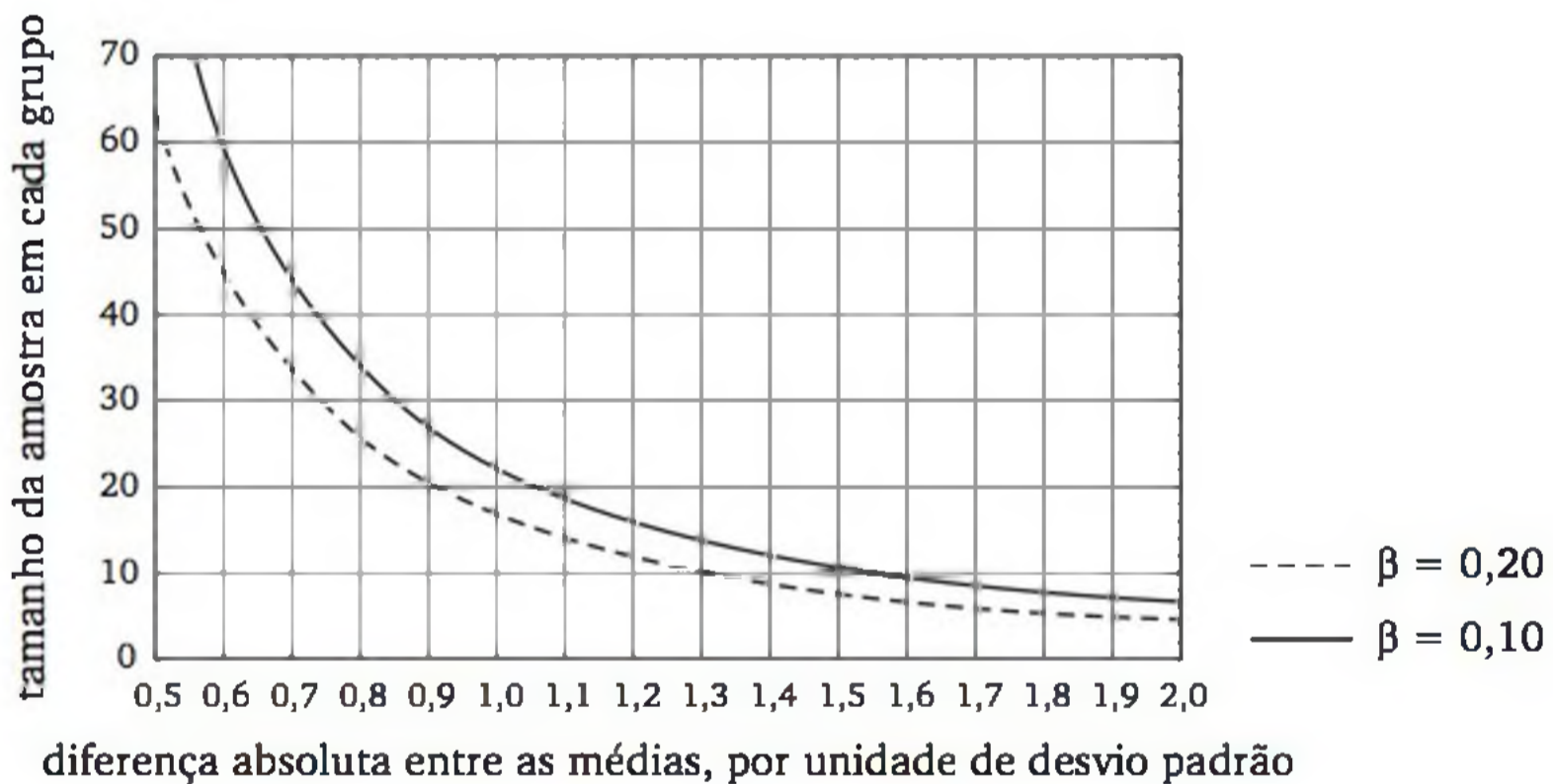


Figura 9.7 *Tamanho mínimo da amostra, n , em cada grupo, em função da distância $\delta = |\mu_1 - \mu_2| : \sigma$ que se deseja detectar no teste estatístico.*

Como exemplo, considere o problema de comparar dois catalisadores, como no Exemplo 9.7. Sabe-se que os engenheiros consideram relevante uma diferença de 4 unidades entre as médias e, com base em estudos anteriores, o desvio padrão do rendimento do processo químico em estudo não deve superar 5 unidades. Logo, $\delta = \frac{4}{5} = 0,8$. Pelo gráfico da Figura 9.7, o número mínimo

de ensaios para cada catalisador deve ser de, aproximadamente, $n = 34$ para $\beta = 0,10$; ou $n = 25$ para $\beta = 0,20$.

EXERCÍCIOS

1. Uma empresa de cerveja, após uma grande fusão, estuda a possibilidade de alterar o rótulo de uma de suas marcas, usando formas e cores mais vivas. Para avaliar se existe vantagem em alterar o rótulo, a empresa levou a cabo uma pesquisa de marketing. Enlatou a cerveja com o rótulo tradicional e com o rótulo novo. A pesquisa foi feita em 8 estabelecimentos comerciais. Em 4 deles, extraídos por sorteio, colocou-se o produto com o rótulo novo e, nos outros 4, manteve-se o produto com o rótulo tradicional. Após um mês, avaliou-se a quantidade vendida em cada estabelecimento. Os estabelecimentos que usaram o rótulo tradicional tiveram os seguintes resultados nas vendas (em milhares de unidades): 6, 5, 2, 2. Os estabelecimentos que usaram o rótulo novo tiveram os seguintes resultados nas vendas (em milhares de unidades): 4, 9, 5, 6. Os dados mostram evidência suficiente de que a média de vendas é superior com o rótulo novo? Responda usando um teste estatístico apropriado ao nível de significância de 5%.
2. Para o mesmo problema da questão anterior, outro instituto de pesquisa, que tem uma equipe com melhor preparação em estatística, elaborou um projeto um pouco diferente. Com seis estabelecimentos comerciais dispostos a colaborar com a pesquisa, colocaram-se as duas embalagens (de rótulo tradicional e de rótulo novo) da mesma cerveja. Tomou-se o cuidado para que em cada estabelecimento a apresentação das duas embalagens do produto fosse feita de forma idêntica. Os resultados das vendas mensais (em milhares de unidades), para cada estabelecimento e cada embalagem, foram os seguintes:

Estabelecimento:	1	2	3	4	5	6
Rótulo tradicional:	16	12	28	32	19	25
Rótulo novo:	20	11	33	40	21	31

Os dados mostram evidência suficiente de que a média de vendas é superior com o rótulo novo? Responda usando um teste estatístico apropriado ao nível de significância de 5%.

3. Para avaliar o efeito de um brinde nas vendas de determinado produto, planeja-se comparar as vendas em lojas que vendem o produto com o brinde, com as vendas em lojas que não oferecem o brinde. Para reduzir o efeito de variações devidas a outros fatores, as lojas foram grupadas em pares, de tal forma que as lojas de um mesmo par são o mais similar possível, em

termos, por exemplo, do volume de vendas, localidade, identidade de preços etc. Em cada par de lojas, uma passou a oferecer o brinde e a outra, não.

- a) Apresente as hipóteses nula e alternativa.
 b) Os resultados das vendas, em quantidade de unidades vendidas, foram os seguintes:

Par de loja	Vendas sem brinde	Vendas com brinde
1	33	43
2	43	39
3	26	33
4	19	32
5	37	43
6	27	46

Os dados mostram evidência suficiente para se afirmar que a oferta do brinde aumenta as vendas? Use nível de significância de 5%.

4. Para resolver o mesmo problema do exercício anterior, decidiu-se fazer um planejamento do tipo *antes-e-depois*. Observou-se a venda mensal do produto em questão nas 12 lojas. Depois, passou-se a oferecer um brinde e voltou-se a avaliar a venda mensal desse produto nas 12 lojas. Os incrementos (ou reduções) nas vendas foram os seguintes:

7 10 5 -2 9 0 3 -4 8 9 1 3

- a) Os dados mostram evidência suficiente para se afirmar que a oferta do brinde aumenta as vendas? Use nível de significância de 5%.
 b) Aponte as vantagens e as desvantagens desse planejamento de pesquisa, em relação ao apresentado no Exercício 4, considerando o particular problema em discussão.
 c) Apresente um terceiro planejamento de pesquisa para esse problema, tentando aproveitar as vantagens dos dois procedimentos apresentados.
5. Um produto fabricado por injeção de plástico é analisado em dois níveis de percentual de talco. Os dados seguintes apresentam os resultados da dureza (HRc), segundo o percentual de talco utilizado:

Baixo:	51,7	49,4	65,9	60,0	71,1	72,9	71,9	75,1
Alto:	75,2	76,0	63,7	69,6	67,1	69,1	52,8	57,6

Os dados mostram evidência suficiente para afirmar que a dureza média do produto é diferente nos dois níveis de percentual de talco? Use $\alpha = 0,05$.

6. Deseja-se verificar se há alteração no rendimento médio de um processo de reação química, ao reduzir a temperatura de 80°C para 70°C. Realizaram-se 12 ensaios em cada temperatura, encontrando os seguintes resultados de rendimento (%):

Temperatura de 80°C: média 40,61 e variância 12,86.

Temperatura de 70°C: média 36,61 e variância 9,34.

Qual é a conclusão? Use $\alpha = 0,01$.

7. Na comparação de duas topologias de rede de computadores, C1 e C2, avaliou-se o tempo de transmissão de pacotes de dados entre duas máquinas. Foram realizados 32 ensaios em C1 e 24 ensaios em C2, cujos resultados são apresentados a seguir:

Topologia	Tempo (em décimos de segundo)	Média	Variância
C1	09 12 10 12 11 09 08 12 13	10,625	6,371
	09 13 08 17 09 09 08 09 08		
	14 08 08 08 08 13 10 10 15		
	13 13 12 14 08		
C2	14 15 08 13 16 12 14 17 14	13,458	4,781
	10 13 12 13 14 10 15 12 17		
	16 12 15 13 14 14		

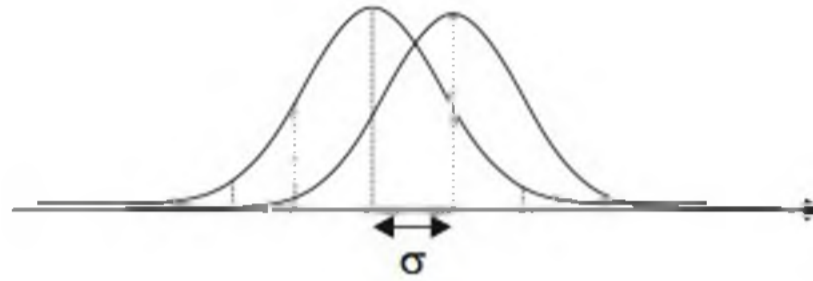
Existe diferença significativa entre o tempo médio de transmissão nas duas topologias?

8. Para comparar dois algoritmos de otimização, foi realizado um experimento com seis ensaios. Em cada ensaio, foram usados separadamente os dois algoritmos em estudo, mas sob as mesmas condições (dados pareados). Os tempos de resposta ao usuário foram:

Ensaio:	1	2	3	4	5	6
Algoritmo 1:	8,1	8,9	9,3	9,6	8,1	11,2
Algoritmo 2:	9,2	9,8	9,9	10,3	8,9	13,1

Os tempos de resposta dos dois algoritmos são, em média, diferentes? Use $\alpha = 0,05$.

9. Quantos ensaios devem ser realizados com cada tratamento para garantir que um teste t para amostras independentes, ao nível de significância de 5%, detecte uma diferença de 1 desvio padrão, com 90% de probabilidade? Admitindo distribuição normal, a diferença mínima que se quer detectar está representada na figura a seguir:



9.5 TESTE F PARA DUAS VARIÂNCIAS

Suponha que queremos comparar se duas populações, supostamente com distribuições normais, têm a mesma variância. Formulamos as hipóteses por:

$$H_0: \sigma_1^2 = \sigma_2^2 \quad \text{e} \quad H_1: \sigma_1^2 \neq \sigma_2^2$$

onde:

σ_1^2 : variância da população 1; e

σ_2^2 : variância da população 2

A hipótese alternativa também pode ser $H_1': \sigma_1^2 > \sigma_2^2$. Com as amostras da população 1 e da população 2, a estatística do teste é calculada por:

$$f = \frac{s_1^2}{s_2^2} \quad (9.8)$$

onde:

s_1^2 : variância da amostra de n_1 elementos; e

s_2^2 : variância da amostra de n_2 elementos, considerando $s_1^2 > s_2^2$, ou seja, a maior variância deve ser colocada no numerador.

A distribuição de referência para este teste é a chamada *distribuição F* com $gl = n_1 - 1$ no numerador e $gl = n_2 - 1$ no denominador, apresentada na Tabela 6 do apêndice para os níveis de significância de 0,10, 0,05, 0,025 e 0,01. Assim, estabelecido o nível de significância α , podemos obter o valor f_c , que deixa área igual a $\alpha/2$ na cauda superior da distribuição (teste bilateral) ou, no caso de

teste unilateral, área igual a α . A regra de decisão, na abordagem clássica, é dada por⁵

se $f < f_c$, então aceita H_0 ;

se $f \geq f_c$, então rejeita H_0 .

Exemplo 9.3 (continuação) Verificamos que não há evidências de que os catalisadores A e B tenham efeitos médios diferentes no rendimento de certa reação química. Vamos verificar, agora, se eles produzem efeitos diferentes nas variâncias. As hipóteses podem ser:

H_0 : as variâncias do rendimento são iguais para os dois catalisadores; e

H_1 : as variâncias do rendimento são diferentes para os dois catalisadores.

Dados os resultados do experimento:

Amostra 1: $n_1 = 10$, $\bar{x}_1 = 49,900$ e $s_1^2 = 35,656$

Amostra 2: $n_2 = 10$, $\bar{x}_2 = 44,700$ e $s_2^2 = 42,233$

No cálculo de f , colocamos a maior variância no numerador, assim:

$$f = \frac{s_2^2}{s_1^2} = \frac{42,233}{35,656} = 1,18$$

Para obter o valor crítico f_c , ao nível de significância de 5%, devemos obter área igual a 2,5% na cauda superior da distribuição F com $gl = 9$ no numerador e $gl = 9$ no denominador, o que acarreta $f_c = 4,03$. Como $f < f_c$, o teste aceita H_0 .

9.6 COMPARAÇÃO DE VÁRIAS MÉDIAS

Nas seções 9.2 e 9.3, aprendemos a testar a significância de duas médias através de testes t . Nesta seção, aprenderemos um teste para verificar se há diferenças significativas entre as médias de g ($g \geq 2$) grupos de observações, sendo cada grupo formado pelos resultados de um tratamento. Dessa forma, podemos comparar dois ou mais algoritmos computacionais em termos do tempo médio de resposta ao usuário; dois ou mais catalisadores em termos do rendimento médio da reação química etc.

5 Estamos usando a abordagem clássica para facilitar o uso da tabela, mas, se o teste for feito com auxílio do computador, a abordagem do valor p é mais fácil.

9.6.1 Amostras independentes

A análise estatística para a comparação de g grupos independentes é tradicionalmente feita por uma análise de variância (ANOVA), acompanhada de um teste F , que, da mesma forma como o teste t , supõe:

1. as observações devem ser independentes;
2. as variâncias populacionais devem ser iguais nos g grupos; e
3. a distribuição das observações em cada grupo deve ser normal.⁶

Formalmente, temos as seguintes hipóteses:

$$H_0: \mu_1 = \mu_2 = \dots = \mu_g \text{ e } H_1: \mu_i \neq \mu_j, \text{ para algum } i \neq j$$

onde μ_i representa o valor esperado da resposta sob o tratamento i ($i = 1, 2, \dots, g$) (ver Figura 9.8).

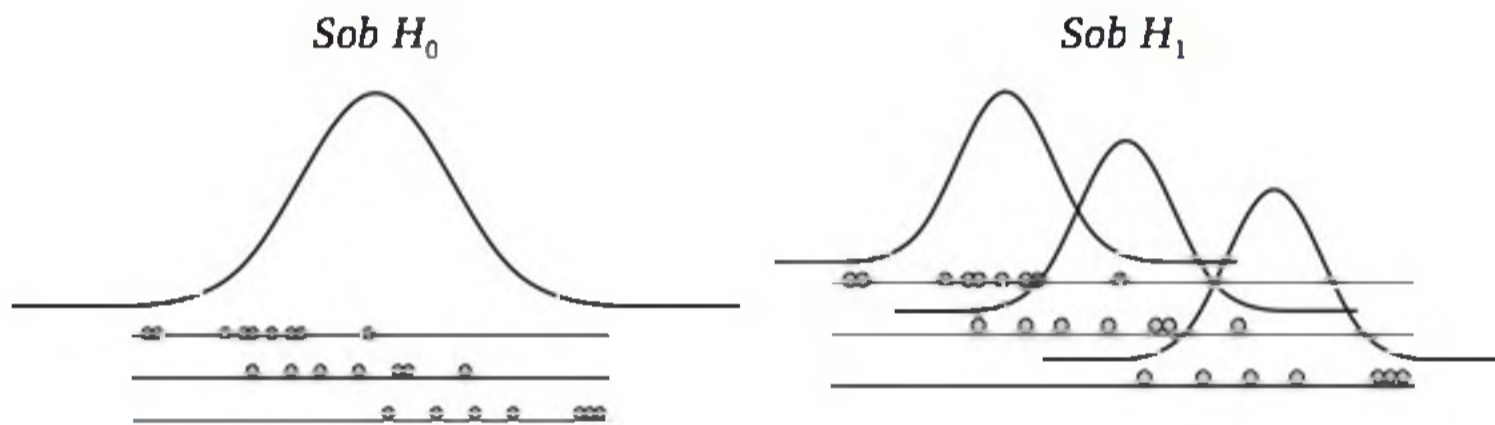


Figura 9.8 Suposições sobre as observações em termos de H_0 e H_1 .

Considerando n replicações sob cada tratamento (amostra de n elementos em cada grupo, totalizando $N = ng$ observações), podemos representar os dados pelo seguinte modelo estatístico:

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad (i = 1, 2, \dots, g; j = 1, 2, \dots, n) \quad (9.9)$$

onde: Y_{ij} é a variável aleatória associada à j -ésima observação do i -ésimo tratamento;

μ é a média global da resposta (independentemente do tratamento);

τ_i é o efeito do i -ésimo tratamento;

ε_{ij} é o efeito aleatório ou erro experimental, o qual é suposto com distribuição aproximadamente normal, média zero e variância constante.

⁶ Para $g = 2$, o teste F é equivalente ao teste t bilateral. Observamos que o teste F é robusto com respeito às suposições (2) e (3), ou seja, ele ainda é válido com pequenas violações destas suposições.

Considerando o modelo (9.9), o valor esperado da resposta no i -ésimo tratamento é dado por $\mu_i = \mu + \tau_i$. E as hipóteses podem ser escritas por:

$$H_0: \tau_1 = \tau_2 = \dots = \tau_g = 0 \quad \text{e} \quad H_1: \tau_i \neq 0, \text{ para algum } i = 1, 2, \dots, g$$

As observações, as somas e as médias por tratamento são representadas por⁷

Replicação	Tratamento				
	1	2	...	g	
1	y_{11}	y_{21}	...	y_{g1}	
2	y_{12}	y_{22}	...	y_{g2}	
...	
n	y_{1n}	y_{2n}	...	y_{gn}	
Soma	$y_{1.}$	$y_{2.}$...	$y_{g.}$	$y_{..} = \sum_i y_{i.}$
Média	$\bar{y}_{1.}$	$\bar{y}_{2.}$...	$\bar{y}_{g.}$	$\bar{y}_{..} = \frac{1}{g} \sum_i \bar{y}_{i.}$

Considere a seguinte soma de quadrados:

$$SQ_{Tot} = \sum_{i=1}^g \sum_{j=1}^n (y_{ij} - \bar{y}_{..})^2 \tag{9.10}$$

Se H_0 for verdadeira e, portanto, todas as observações provêm de uma mesma população, então SQ_{Tot} é o numerador do cálculo da variância, s^2 , de todas as $N = ng$ observações. Pode-se mostrar que SQ_{Tot} (soma de quadrados total) é decomposta na soma de quadrados dos tratamentos e na soma de quadrados do erro dadas, respectivamente, por:

$$SQ_{Trat} = \sum_{i=1}^g \sum_{j=1}^n (\bar{y}_{i.} - \bar{y}_{..})^2 = n \sum_{i=1}^g (\bar{y}_{i.} - \bar{y}_{..})^2 \tag{9.11}$$

$$SQ_{Erro} = \sum_{i=1}^g \sum_{j=1}^n (y_{ij} - \bar{y}_{i.})^2 \tag{9.12}$$

Os g diferentes desvios de SQ_{Trat} são feitos em relação a uma única média amostral ($\bar{y}_{..}$). Por isso, dizemos que SQ_{Trat} tem $g - 1$ graus de liberdade. Enquan-

7 Se os tamanhos das amostras forem diferentes, é necessário adequar a formulação apresentada nesta seção, substituindo n_i (tamanho da amostra no grupo i) por n .

to os N desvios de SQ_{Erro} são feitos em relação a g médias amostrais ($\bar{y}_{i.}$, $i = 1, 2, \dots, g$) e, por isso, SQ_{Erro} tem $N - g$ graus de liberdade. A divisão das somas de quadrados pelos correspondentes graus de liberdade leva aos chamados *quadrados médios*. Assim:

$$QM_{\text{Trat}} = \frac{SQ_{\text{Trat}}}{gl_{\text{Trat}}} = \frac{n \sum_{i=1}^g (\bar{y}_{i.} - \bar{y}_{..})^2}{g - 1} \quad (9.13)$$

$$QM_{\text{Erro}} = \frac{SQ_{\text{Erro}}}{gl_{\text{Erro}}} = \frac{\sum_{i=1}^g \sum_{j=1}^n (y_{ij} - \bar{y}_{i.})^2}{N - g} \quad (9.14)$$

onde $N = ng$.

Observe que QM_{Trat} é uma medida da variância *entre* as médias dos grupos, enquanto QM_{Erro} é uma medida da variância *dentro* dos grupos. Define-se a razão:

$$f = \frac{QM_{\text{Trat}}}{QM_{\text{Erro}}} \quad (9.15)$$

que pode ser interpretada como uma medida de discriminação entre os g grupos. Para testar a hipótese $H_0: \mu_1 = \mu_2 = \dots = \mu_g$, usamos a *distribuição F*, com $gl = g - 1$ no numerador e $gl = N - g$ no denominador. Assim, estabelecido o nível de significância α , podemos obter na Tabela 6 do apêndice o valor f_c , que deixa área igual a α na cauda superior da distribuição. A regra de decisão é dada por:

se $f < f_c$, então aceita H_0 ;

se $f \geq f_c$, então rejeita H_0 .

As somas de quadrados são mais facilmente calculadas, conforme as expressões apresentadas no Quadro 9.1.

Quadro 9.1 Cálculos básicos da ANOVA com um fator.

Fonte de variação	Somas de quadrados	gl	Quadrados médios	Razão <i>f</i>
Entre tratamentos	$SQ_{Trat} = \sum_{i=1}^g \frac{y_i^2}{n} - \frac{y_{..}^2}{N}$	$g - 1$	$QM_{Trat} = \frac{SQ_{Trat}}{gl_{Trat}}$	$f = \frac{QM_{Trat}}{QM_{Erro}}$
Dentro de trat. (Erro)	$SQ_{Erro} = SQ_{Tot} - SQ_{Trat}$	$N - g$	$QM_{Erro} = \frac{SQ_{Erro}}{gl_{Erro}}$	
Total	$SQ_{Tot} = \sum_{i=1}^g \sum_{j=1}^n y_{ij}^2 - \frac{y_{..}^2}{N}$	$N - 1$		

Exemplo 9.4 Considere o problema de comparar três tipos de rede de computadores, C1, C2 e C3, em termos do tempo médio de transmissão de pacotes de dados entre duas máquinas. Realizou-se um experimento com oito replicações com cada tipo de rede, aleatorizando a ordem dos 24 ensaios e mantendo fixos os demais fatores controláveis (ver resultados na Tabela 9.3). Deseja-se testar as hipóteses:

- H_0 : os tempos esperados de transmissão são iguais para os três tipos de rede; e
- H_1 : os tempos esperados de transmissão não são todos iguais (depende do tipo de rede).

Tabela 9.3 Resultados do experimento do Exemplo 9.4.

Replicação	Tipo de rede		
	C1	C2	C3
1	7,2	7,8	6,3
2	9,3	8,2	6,0
3	8,7	7,1	5,3
4	8,9	8,6	5,1
5	7,6	8,7	6,2
6	7,2	8,2	5,2
7	8,8	7,1	7,2
8	8,0	7,8	6,8
Soma	65,7	63,5	48,1
Média	8,21	7,94	6,01

Soma global: $y_{..} = 177,3$

Soma de quadrados: $\sum_{i=1}^g \sum_{j=1}^n y_{ij}^2 = (7,2)^2 + (9,3)^2 + \dots = 1344,25$

$$SQ_{Trat} = \sum_{i=1}^g \frac{y_i^2}{n} - \frac{y_{..}^2}{N} = \frac{(65,7)^2 + (63,5)^2 + (48,1)^2}{8} - \frac{(177,3)^2}{24} = 22,99$$

$$SQ_{Total} = \sum_{i=1}^g \sum_{j=1}^n y_{ij}^2 - \frac{y_{..}^2}{N} = 1344,25 - \frac{(177,3)^2}{24} = 34,45$$

$$SQ_{Erro} = SQ_{Total} - SQ_{Trat} = 34,45 - 22,29 = 11,46$$

Resultando no quadro da ANOVA:

Fonte da variação	SQ	gl	QM	f
Entre grupos	22,99	2	11,50	21,07
Dentro dos grupos	11,46	21	0,55	
Total	34,45	23		

Adotando $\alpha = 0,05$, temos o valor crítico $f_c = 3,47$ (Tabela 6). Como o valor calculado ($f = 21,07$) é superior ao valor crítico, então o teste rejeita H_0 , provando estatisticamente que há diferença entre os três tipos de rede, em termos do tempo médio de transmissão.⁸

Análise dos resíduos

Já comentamos que o teste F é válido se os dados provêm de distribuições normais com variância constante nos diferentes grupos. Essas suposições são mais fáceis de ser verificadas a partir dos *resíduos* (diferenças entre *valores observados e médias dos grupos*):

$$e_{ij} = y_{ij} - \bar{y}_i \quad (i = 1, 2, \dots, g; j = 1, 2, \dots, n) \quad (9.16)$$

A Figura 9.9 mostra os valores dos resíduos para cada tipo de rede e o gráfico de probabilidade normal dos resíduos. O primeiro gráfico indica que é ra-

⁸ Se a análise for feita com um *software* estatístico, ou mesmo com uma planilha eletrônica, o sistema apresenta a tabela da ANOVA com o valor p associado à estatística F , bastando, então, comparar p com o nível de significância estabelecido para decidir entre H_0 e H_1 .

zoável a suposição de variância constante nos três grupos; e o segundo mostra os pontos em torno de uma linha, validando a suposição de normalidade dos dados (ver seção 6.4).

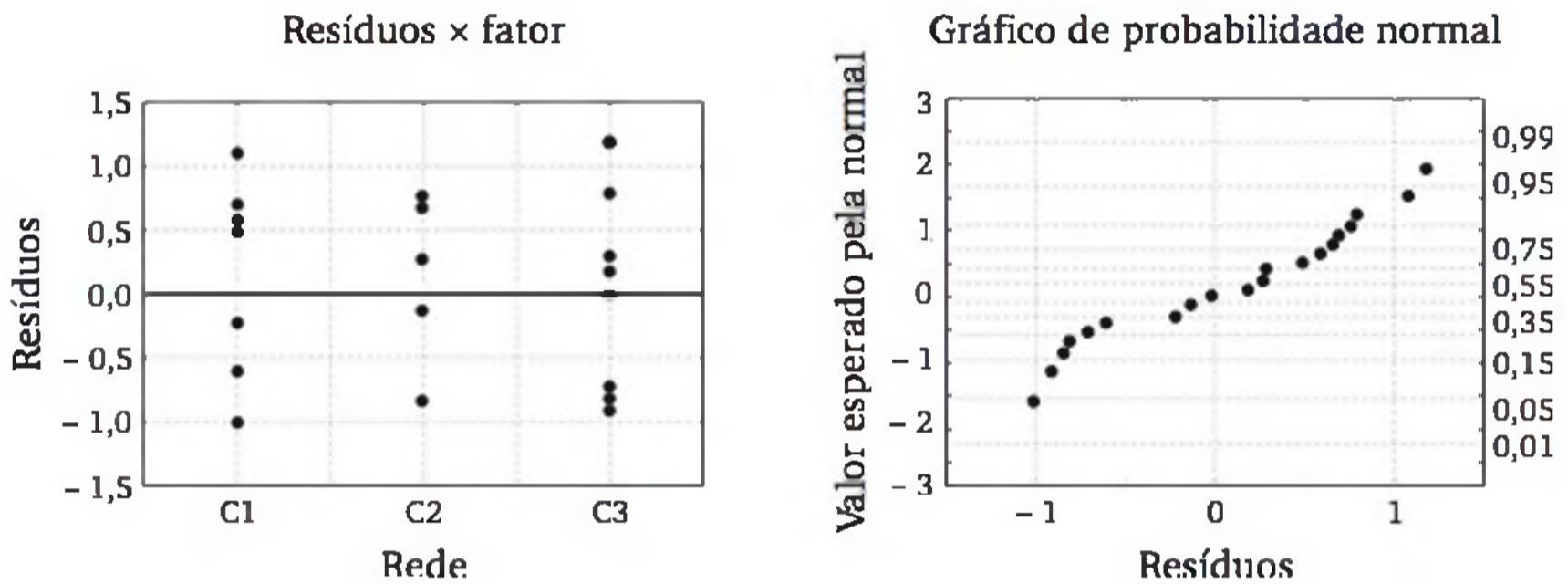


Figura 9.9 Avaliação das suposições da ANOVA através de gráficos dos resíduos.

Estimação das médias

As médias aritméticas de cada grupo servem como uma estimativa pontual da resposta esperada de cada tratamento. Podemos, também, usar a abordagem de intervalos de confiança para efetuarmos uma inferência mais completa. O erro padrão pode ser estimado através da variância conjunta das g amostras. Daí, o intervalo de confiança para o valor esperado da resposta sob o i -ésimo tratamento é dado por:

$$IC(\mu_i, \gamma) = \bar{y}_i \pm t_\gamma \sqrt{\frac{QM_{erro}}{n}} \tag{9.17}$$

onde t_γ é obtido na Tabela 4, em função do nível de confiança γ estabelecido e dos graus de liberdade: $gl = g(n - 1)$.

Exemplo 9.4 (continuação) Usando nível de confiança de 95% e $gl = N - g = 24 - 3 = 21$, temos o valor tabelado $t_{95\%} = 2,08$. Então, para a rede C1, temos:

$$IC(\mu_1, 95\%) = 8,21 \pm 2,08 \sqrt{\frac{0,55}{8}} = 8,21 \pm 0,55$$

A Figura 9.10 apresenta graficamente intervalos de confiança para os três tipos de rede.

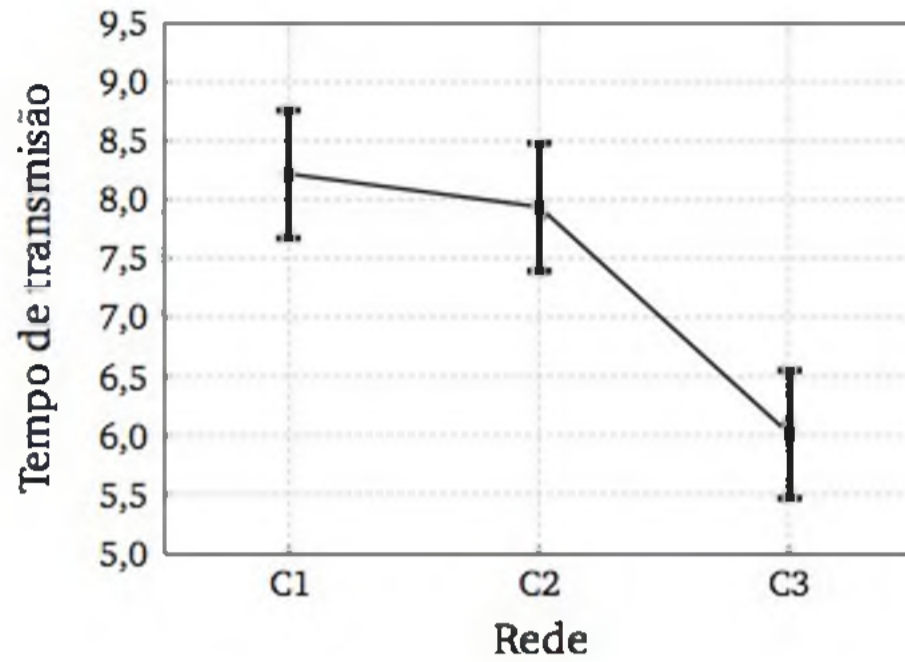


Figura 9.10 Estimativas, através de intervalos de 95% de confiança, para o tempo esperado de transmissão, em três tipos de rede.

A Figura 9.10 sugere que as redes C1 e C2 podem ser consideradas com o mesmo tempo esperado de transmissão, pois seus intervalos de confiança se sobrepõem, indicando que a diferença entre as médias aritméticas pode ser meramente casual. Já a rede C3 parece ser a causa do teste F da ANOVA ter rejeitado H_0 , pois apresenta um intervalo de confiança para a média não sobreposto com os outros intervalos de confiança, evidenciando produzir tempo esperado de transmissão menor.⁹

9.6.2 Amostras em blocos

Considere, agora, que as unidades experimentais são agrupadas em h blocos, de tal forma que todos os g tratamentos sejam realizados em cada bloco, conforme o esquema a seguir:

Bloco	Tratamento				Soma
	1	2	...	g	
1	y_{11}	y_{21}	...	y_{g1}	$y_{.1}$
2	y_{12}	y_{22}	...	y_{g2}	$y_{.2}$
...	
h	y_{1h}	y_{2h}	...	y_{gh}	$y_{.h}$
Soma	$y_{1.}$	$y_{2.}$...	$y_{g.}$	$y_{..} = \sum_i y_{i.} = \sum_j y_{.j}$

⁹ Sugerimos que a análise por intervalos de confiança seja feita somente após a ANOVA rejeitar a hipótese de igualdade entre todas as médias, porque somente o teste F é construído especificamente para a comparação de vários tratamentos.

Nesse projeto, cada observação Y_{ij} pode ser influenciada pelo efeito do tratamento, pelo efeito do bloco e pelo efeito aleatório. Assim, consideraremos:

$$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij} \tag{9.18}$$

onde: μ é a média global da resposta;

τ_i é o efeito do i -ésimo tratamento;

β_j é o efeito do j -ésimo bloco; e

ε_{ij} é o efeito aleatório ($i = 1, 2, \dots, g; j = 1, 2, \dots, h$).

Ao fazer a análise de variância, devemos excluir a variação devida aos blocos da variação do erro experimental. O Quadro 9.2 mostra os cálculos de uma ANOVA de um projeto em blocos, considerando em cada bloco *uma* observação de cada tratamento (*blocos completos não replicados*), donde o número total de observações é dado por $N = gh$.

Quadro 9.2 Cálculos da ANOVA num experimento em blocos completos não replicado.

Fonte de variação	Somas de quadrados	gl	Quadrados médios	Razão f
Entre tratamentos	$SQ_{Trat} = \sum_{i=1}^g \frac{y_i^2}{h} - \frac{y_{..}^2}{N}$	$g - 1$	$QM_{Trat} = \frac{SQ_{Trat}}{gl_{Trat}}$	$f = \frac{QM_{Trat}}{QM_{Erro}}$
Entre blocos	$SQ_{Blocos} = \sum_{j=1}^h \frac{y_{.j}^2}{g} - \frac{y_{..}^2}{N}$	$h - 1$	$QM_{Bloco} = \frac{SQ_{Bloco}}{gl_{Bloco}}$	
Dentro (Erro)	$SQ_{Erro} = SQ_{Tot} - SQ_{Trat} - SQ_{Bloco}$	$(g-1)(h-1)$	$QM_{Erro} = \frac{SQ_{Erro}}{gl_{Erro}}$	
Total	$SQ_{Tot} = \sum_{i=1}^g \sum_{j=1}^h y_{ij}^2 - \frac{y_{..}^2}{N}$	$N - 1$		

Exemplo 9.5 Seja o problema de comparar três algoritmos de busca em um banco de dados. Realiza-se um experimento com seis buscas experimentais, sendo que em cada uma é sorteado um número aleatório que indica o registro do banco de dados a ser localizado. Em cada um dos seis processos de busca, são usados separadamente os três algoritmos em estudo, mas sob as mesmas condições, em termos dos fatores controláveis. São anotados os tempos de resposta ao usuário.

Inicialmente, queremos testar as hipóteses:

H_0 : em média, os três algoritmos *são igualmente* rápidos; e

H_1 : em média, os três algoritmos *não são igualmente* rápidos.

Resultados do experimento:

Ensaio (bloco)	Algoritmo de busca		
	A1	A2	A3
1	8,3	8,1	9,2
2	9,4	8,9	9,8
3	9,1	9,3	9,9
4	9,9	9,6	10,3
5	8,2	8,1	8,9
6	10,9	11,2	13,1
Soma	55,8	55,2	61,2
Média	9,43	9,2	10,2

ANOVA (usando as expressões do Quadro 9.2):

Fonte da variação	SQ	gl	QM	f
Algoritmos	3,64	2	1,82	14,29
Blocos	21,95	5	4,39	
Erro	1,27	10	0,13	
Total	26,86	17		

Adotando $\alpha = 0,05$, temos o valor crítico $f_c = 4,10$ (Tabela 4 do apêndice, com $gl = 2$ no numerador e $gl = 10$ no denominador). Como o valor calculado ($f = 14,29$) é superior ao valor crítico, então o teste rejeita H_0 , provando estatisticamente que há diferença entre os três algoritmos de busca, em termos do tempo médio de resposta ao usuário.

Usando a mesma abordagem da seção anterior, construímos intervalos de confiança para o tempo esperado de resposta para cada algoritmo (Figura 9.11). Observamos que os algoritmos A1 e A2 são significativamente mais rápidos que A3.

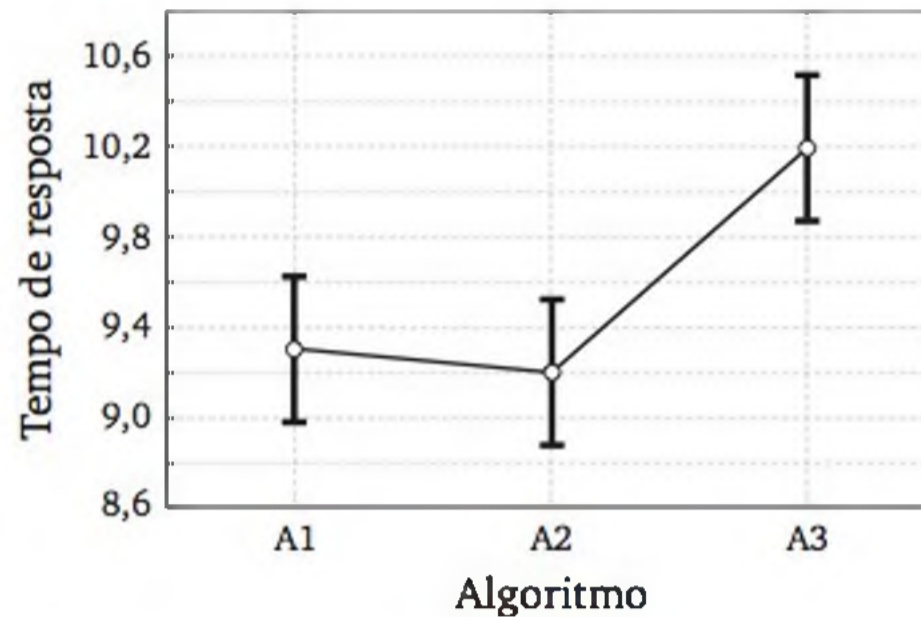


Figura 9.11 Estimativas, através de intervalos de 95% de confiança, para o tempo esperado de resposta de três algoritmos de busca em banco de dados.

9.7 ANOVA EM PROJETOS FATORIAIS

Nos estudos experimentais, em geral procuramos avaliar ou testar o efeito de mais de um fator sobre uma resposta de interesse. Por exemplo, o profissional de informática pode estar interessado em estudar o tempo de resposta para o usuário quando se varia a topologia, o protocolo e o número de nós de uma rede local. O engenheiro civil quer conhecer o quanto o tempo de hidratação, a dosagem de cimento e o uso de aditivos interferem na resistência à compressão de um concreto. O engenheiro químico quer saber a influência da temperatura e do tempo de reação sobre o rendimento de uma reação química.

Um projeto experimental é dito *fatorial* quando cada nível de um fator é ensaiado com todos os níveis dos outros fatores, sem restrições. E o primeiro interesse, normalmente, é testar se existe diferença no valor esperado da resposta entre os níveis de cada fator, além de testar eventuais *interações* entre os fatores.¹⁰ Nesta seção, restringiremos o estudo a um projeto fatorial com dois fatores.

Sejam dois fatores, A e B, com g e h níveis, respectivamente. Considere que em cada cruzamento dos níveis desses fatores sejam realizadas n replicações. As $N = ghn$ observações podem ser descritas pelo seguinte modelo estatístico:

$$Y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ijk} \quad (9.19)$$

onde: μ é a média global da resposta (independentemente dos efeitos dos fatores);

τ_i é o efeito do i -ésimo nível do fator A;

¹⁰ Dizemos que existe *interação* entre dois fatores quando a diferença na resposta entre os níveis de um fator não é a mesma para todos os níveis do outro fator (ver Capítulo 2).

β_j é o efeito do j -ésimo nível do fator B;

$(\tau\beta)_{ij}$ é o efeito da interação entre τ_i e β_j ; e

ε_{ijk} é o efeito aleatório ou erro experimental ($i = 1, 2, \dots, g; j = 1, 2, \dots, h; k = 1, 2, \dots, n$).

Uma única análise de variância permite efetuar três testes estatísticos, associados às seguintes hipóteses nulas:

$H_0^{(A)}: \tau_1 = \tau_2 = \dots = \tau_g = 0$ (não há diferença no valor esperado da resposta nos g níveis do fator A);

$H_0^{(B)}: \beta_1 = \beta_2 = \dots = \beta_h = 0$ (não há diferença no valor esperado da resposta nos h níveis do fator B);

$H_0^{(AB)}: (\tau\beta)_{ij} = 0, \forall i, j$ (não há interação entre os fatores A e B).

Sejam as $N = ghn$ observações e as seguintes notações de somas:

Fator B	Fator A				Soma
	1	2	...	g	
1	y_{111}, \dots, y_{11n}	y_{211}, \dots, y_{21n}	...	y_{g11}, \dots, y_{g1n}	$y_{.1.}$
2	y_{121}, \dots, y_{12n}	y_{221}, \dots, y_{22n}	...	y_{g21}, \dots, y_{g2n}	$y_{.2.}$
...	
h	y_{1h1}, \dots, y_{1hn}	y_{2h1}, \dots, y_{2hn}	...	y_{gh1}, \dots, y_{ghn}	$y_{.h.}$
Soma	$y_{1..}$	$y_{2..}$...	$y_{g..}$	$y_{...} = \sum_i y_{i.} = \sum_j y_{.j.}$

A soma das observações em cada casela é representada por:

$$y_{ij.} = \sum_{k=1}^n y_{ijk} \tag{9.20}$$

E a soma de quadrados entre as caselas é dada por:

$$SQ_{Subtot} = \sum_{i=1}^g \sum_{j=1}^h \frac{y_{ij.}^2}{n} - \frac{y_{...}^2}{N} \tag{9.21}$$

O Quadro 9.3 apresenta os cálculos para a realização de uma ANOVA para esse tipo de projeto experimental.

Quadro 9.3 Cálculos da ANOVA num projeto fatorial com dois fatores.

Fonte de variação	Somas de quadrados	gl	Quadrados médios	Razão f
Fator A	$SQ_A = \sum_{i=1}^g \frac{y_{i.}^2}{hn} - \frac{y_{...}^2}{N}$	$g - 1$	$QM_A = \frac{SQ_A}{gl_A}$	$f = \frac{QM_A}{QM_{Erro}}$
Fator B	$SQ_B = \sum_{j=1}^h \frac{y_{.j}^2}{gn} - \frac{y_{...}^2}{N}$	$h - 1$	$QM_B = \frac{SQ_B}{gl_B}$	$f = \frac{QM_B}{QM_{Erro}}$
Interação A*B	$SQ_{AB} = SQ_{Subtot} - SQ_A - SQ_B$	$(g-1)(h-1)$	$QM_{AB} = \frac{SQ_{AB}}{gl_{AB}}$	$f = \frac{QM_{AB}}{QM_{Erro}}$
Erro	$SQ_{Erro} = SQ_{Tot} - SQ_{Subtot}$	$hg(n-1)$	$QM_{Erro} = \frac{SQ_{Erro}}{gl_{Erro}}$	
Total	$SQ_{Tot} = \sum_{i=1}^g \sum_{j=1}^h \sum_{k=1}^n y_{ijk}^2 - \frac{y_{...}^2}{N}$	$N - 1$		

Exemplo 9.6 Considere o problema de comparar três topologias de rede de computadores (C1, C2 e C3) e dois protocolos (L1 e L2), em termos do tempo de resposta ao usuário. Realizou-se um experimento com quatro replicações em cada combinação de topologia e protocolo. Deseja-se verificar se há diferenças entre as topologias, entre os protocolos e eventual interação entre topologia e protocolo. Então, deseja-se testar as seguintes hipóteses nulas:

$H_0^{(A)}$: os tempos esperados de resposta são iguais para as três topologias;

$H_0^{(B)}$: os tempos esperados de resposta são iguais para os dois protocolos;

$H_0^{(AB)}$: a mudança de protocolo não altera as diferenças médias do tempo de resposta nas três topologias (ausência de interação).

Resultados do experimento:

Protocolo	Topologia			Soma	Média
	C1	C2	C3		
L1	6,2 7,6 7,2 8,8	5,9 8,4 7,1 7,1	5,9 6,2 5,2 7,2	$y_{1.} = 82,8$	6,90
L2	9,0 8,9 9,4 8,0	7,1 8,6 9,1 7,8	6,2 6,1 8,9 6,8	$y_{2.} = 95,9$	7,99
Soma Média	$y_{1..} = 65,1$ 8,21	$y_{2..} = 61,1$ 7,94	$y_{3..} = 52,5$ 6,01	$y_{...} = 178,7$	7,45

Soma das observações em cada casela (y_{ij}):

Protocolo	Topologia		
	C1	C2	C3
L1	29,8	28,5	24,5
L2	35,3	32,6	28,0

Somas de quadrados:

$$SQ_{Subtot} = \sum_{i=1}^g \sum_{j=1}^h \frac{y_{ij}^2}{n} - \frac{y_{...}^2}{N} = \frac{5393,39}{4} - \frac{31933,69}{24} = 17,77$$

$$SQ_{Tot} = \sum_{i=1}^k \sum_{j=1}^h \sum_{k=1}^n y_{ijk}^2 - \frac{y_{...}^2}{N} = 1365,49 - 1330,57 = 34,92$$

$$SQ_A = \sum_{i=1}^g \frac{y_{i.}^2}{hn} - \frac{y_{...}^2}{N} = \frac{10727,47}{8} - 1330,57 = 10,36$$

$$SQ_B = \sum_{j=1}^h \frac{y_{.j}^2}{gn} - \frac{y_{...}^2}{N} = \frac{16052,65}{12} - 1330,57 = 7,15$$

A ANOVA e os valores críticos, f_c , ao nível de significância de 5%, são descritos a seguir:

Fonte da variação	SQ	gl	QM	f	f _c
Topologia	10,36	2	5,18	5,44	3,55
Protocolo	7,15	1	7,15	7,51	4,41
Interação	0,26	2	0,13	0,14	3,55
Erro	17,14	18	0,95		
Total	34,92	23			

Concluimos que tanto as diferentes topologias, quanto os diferentes protocolos utilizados alteram significativamente a média de tempo ao usuário, mas não há interação entre esses dois fatores.

No projeto fatorial, os valores *preditos* são as médias dos subgrupos:

$$\bar{y}_{ij.} = \frac{1}{n} \sum_{k=1}^n y_{ijk} \tag{9.22}$$

e os *resíduos* são as diferenças entre os valores observados e a média do respectivo subgrupo:

$$e_{ijk} = y_{ijk} - \bar{y}_{ij.} \tag{9.23}$$

A Figura 9.12 (a) ilustra as médias do tempo de resposta para cada topologia e protocolo; e a Figura 9.12 (b) apresenta o gráfico dos valores preditos com os resíduos.

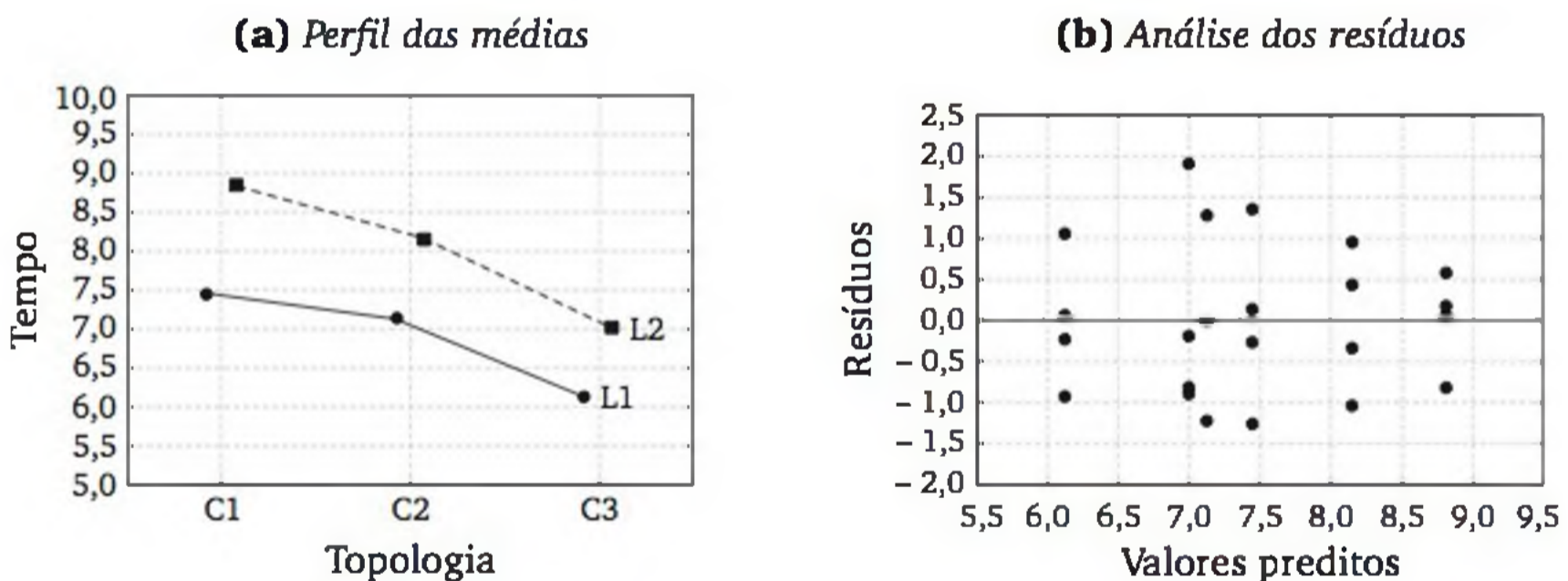


Figura 9.12 (a) Média do tempo de resposta ao usuário, por topologia e protocolo, e (b) gráfico dos valores preditos com os resíduos.

Na Figura 9.12 (a), podemos observar diferenças entre os níveis dos dois fatores, mas os perfis são quase paralelos, o que levou o teste F a aceitar que os dois fatores agem aditivamente (ausência de interação).

Na Figura 9.12 (b), observamos que os pontos se distribuem aleatoriamente em torno da linha horizontal, associada ao resíduo nulo. Além disso, os pontos estão em maior intensidade perto da linha horizontal. Isso sugere que as suposições de normalidade e variância constante estão razoavelmente satisfeitas, validando os resultados da ANOVA.

9.8 ANOVA EM PROJETOS DO TIPO 2^k

Como discutimos no Capítulo 2, quando se quer analisar muitos fatores, é comum iniciar com um experimento em que todos os fatores são ensaiados em apenas dois níveis. Havendo replicações, podemos verificar formalmente a significância de cada efeito através do teste F e, nesses projetos, os cálculos da ANOVA são relativamente simples.¹¹

Sejam:

k o número de fatores;

n o número de observações em cada condição experimental (número de replicações); e

$$c_{ij} = \begin{cases} -1, & \text{se o } j\text{-ésimo fator for ensaiado do nível inferior, no} \\ & i\text{-ésimo ensaio} \\ +1, & \text{se o } j\text{-ésimo fator for ensaiado no nível superior,} \\ & \text{no } i\text{-ésimo ensaio}^{12} \end{cases} \quad \begin{matrix} (i = 1, 2, \dots, N; j = \\ = 1, 2, \dots, k). \end{matrix}$$

Em um projeto com $k = 3$ fatores e $n = 2$ replicações, temos um total de $N = 2^k n = 16$ observações. A Tabela 9.4 mostra os c_{ij} associados aos fatores A, B e C, além dos c_{ij} associados às interações AB, AC, BC e ABC, que são obtidos pela multiplicação elemento a elemento das colunas dos fatores envolvidos. Em geral, num projeto 2^k , o número total de fatores e interações é igual a $2^k - 1$.

A soma de quadrados total é obtida por:

$$SQ_{Tot} = \sum_{i=1}^N y_i^2 - \frac{1}{N} \left(\sum_{i=1}^N y_i \right)^2 \quad (9.24)$$

11 Para projetos do tipo fracionado 2^{k-p} , pode-se usar o mesmo processo desta seção, substituindo nas fórmulas k por $k - p$.

12 Para fatores qualitativos, alocamos, arbitrariamente, o código -1 para uma categoria; e $+1$ para a outra categoria.

A soma de quadrados do efeito j ($j = 1, 2, \dots, 2^k - 1$), associada a um fator ou a uma interação entre fatores, pode ser calculada por:

$$SQ_j = \frac{1}{N} \left(\sum_{i=1}^N c_{ij} y_i \right)^2 \tag{9.25}$$

A soma de quadrados do erro pode ser obtida pela diferença:

$$SQ_{\text{Erro}} = SQ_{\text{Tot}} - \sum_{j=1}^{2^k-1} SQ_j \tag{9.26}$$

Todas as SQ_j ($j = 1, 2, \dots, 2^k - 1$) têm apenas *um* grau de liberdade cada uma, enquanto a SQ_{Erro} tem $N - 2^k$ graus de liberdade.

Exemplo 9.7 Usaremos o problema do Exemplo 2.6, em que descrevemos um estudo para verificar os fatores que influenciam a qualidade da transmissão de dados através da porta serial de microcomputadores com cabos longos. Observou-se a taxa de falhas de transmissão (y) em função dos fatores: (A) velocidade da transmissão (2400/9600 *bauds*), (B) tamanho do arquivo (100/200 *bytes*) e (C) comprimento do cabo serial (15/20 m). O experimento foi realizado com duas replicações. Os sinais para o cálculo dos efeitos e os resultados do experimento são apresentados na Tabela 9.4.

Tabela 9.4 Projeto experimental e resultados do experimento do Exemplo 9.7.

Ensaio*	Replicação	A	B	C	AB	AC	BC	ABC	y
1	1	-1	-1	-1	1	1	1	-1	32,5
2	2	-1	-1	-1	1	1	1	-1	32,3
3	1	-1	-1	1	1	-1	-1	1	35,7
4	2	-1	-1	1	1	-1	-1	1	35,9
5	1	-1	1	-1	-1	1	-1	1	33,1
6	2	-1	1	-1	-1	1	-1	1	33,4
7	1	-1	1	1	-1	-1	1	-1	35,9
8	2	-1	1	1	-1	-1	1	-1	36,1
9	1	1	-1	-1	-1	-1	1	1	34,1
10	2	1	-1	-1	-1	-1	1	1	34,4
11	1	1	-1	1	-1	1	-1	-1	36,6
12	2	1	-1	1	-1	1	-1	-1	36,9
13	1	1	1	-1	1	-1	-1	-1	34,2
14	2	1	1	-1	1	-1	-1	-1	34,2
15	1	1	1	1	1	1	1	1	37,1
16	2	1	1	1	1	1	1	1	36,9

* A ordem de realização dos ensaios foi aleatória.

Aplicando (9.24) e (9.25), temos:

$$SQ_{Tot} = \sum_{i=1}^N y_i^2 - \frac{1}{N} \left(\sum_{i=1}^N y_i \right)^2 = 19590,67 - \frac{(559,3)^2}{16} = 39,639$$

$$SQ_A = \frac{1}{N} \left(\sum_i c_{ij} y_i \right)^2 = \frac{1}{16} (-32,5 - \dots - 36,1 + 34,1 + \dots + 36,9)^2 = 5,641$$

As somas de quadrados dos demais efeitos são calculadas de forma análoga, resultando na seguinte ANOVA:

Fonte da variação	SQ	gl	QM	f
A	5,641	1	5,641	208,9
B	0,391	1	0,391	14,5
C	32,776	1	32,776	1.213,9
A*B	0,181	1	0,181	6,7
A*C	0,181	1	0,181	6,7
B*C	0,031	1	0,031	1,1
A*B*C	0,226	1	0,226	8,4
Erro	0,215	8	0,027	
Total	39,639	15		

Para testar a significância de cada efeito, comparamos os valores calculados de f com o valor crítico da distribuição F , com $gl = 1$ no numerador e $gl = 8$ no denominador. Pela Tabela 4 do apêndice, o valor crítico é $f_c = 5,32$. Assim, com exceção da interação $B*C$, todos os outros efeitos são significativos.

EXERCÍCIOS

10. Com o objetivo de comparar três tipos de cimento em termos da resistência à compressão do concreto, foi realizado um experimento completamente aleatorizado, com cinco corpos de prova de cada tipo de cimento. Os resultados foram os seguintes:

Cimento	Resistência à compressão				
1	9	12	10	8	15
2	20	21	23	17	30
3	10	9	12	20	11

- a) Faça uma ANOVA e verifique se há evidência de que existe diferença real entre as resistências médias dos três tipos de cimento. Use $\alpha = 0,05$.
- b) Construa intervalos de 95% de confiança para as médias.
- c) Através da análise dos itens anteriores, pode-se dizer que existe um cimento melhor (maior valor esperado de resistência à compressão) que os outros. E pior?
11. Considerando os dois primeiros tipos de cimento do Exercício 10, verifique se existe diferença significativa entre as variâncias dos tempos de resposta. Use $\alpha = 0,05$.
12. Para comparar a absorção de água de quatro tipos de massa cerâmica, analisaram-se corpos de prova de três fornadas. Em cada fornada (bloco), era analisado um corpo de prova de cada tipo de massa cerâmica. Os resultados (porcentagem de absorção de água) foram:

Fornada	Massa cerâmica			
	C1	C2	C3	C4
1	1,2	1,5	1,1	2,1
2	2,1	2,1	1,3	2,7
3	1,5	1,9	1,3	2,4

Os dados mostram evidência suficiente para garantir diferença na porcentagem esperada de absorção de água nos quatro tipos de massa cerâmica? Use $\alpha = 0,05$.

13. Considere o problema de estudar os efeitos do tamanho da memória principal (fator A) e tamanho da memória *cache* (fator B) no desempenho de um sistema de arquivos de uma rede local de computadores (LAN). O fator A foi ensaiado nos níveis 128 e 256 Mbytes e o fator B nos níveis 256 e 512 kbytes, segundo um projeto fatorial com 3 replicações. A medida de desempenho do sistema foi o número de operações de transferência de arquivos por segundo. Os resultados foram os seguintes:

Memória <i>cache</i>	Memória principal	
	128 Mbytes	256 Mbytes
256 kbytes	500	1.100
	550	1.140
	540	1.090
512 kbytes	900	1.900
	850	1.950
	920	1.910

Baseado em exemplo de FREITAS, P. J. *Introdução à modelagem e simulação de sistemas*. Visual Books, 2001, p. 277. Com permissão do Autor.

ANOVA

Fonte da variação	SQ	gl	QM	f
Memória <i>cache</i>	1.026.675			
Memória principal	1.944.075			
Interação	151.875			
Erro	6.800			
Total	3.129.425			

- a) Complete a tabela da ANOVA acima.
- b) Quais efeitos são significativos?
- c) Os efeitos da memória *cache* e da memória principal podem ser analisados isoladamente? Explique.
14. Para estudar o desempenho, em termos do tempo de resposta (em segundos), de 3 processadores (pro 1, pro 2 e pro 3), sob quatro tipos diferentes de carga de trabalho (cargas 1, 2, 3 e 4), foi realizado um projeto fatorial com 2 replicações. Os resultados foram os seguintes:

Tempo de resposta (s)

Tipo de carga de trabalho	Processador		
	Pro 1	Pro 2	Pro 3
Carga 1	17	8	28
	20	10	27
Carga 2	10	18	27
	7	13	31
Carga 3	13	18	30
	15	21	23
Carga 4	21	7	32
	17	12	26

- a) Complete a tabela da ANOVA e descreva as conclusões.

Fonte da variação	SQ	gl	QM	f
Processador	1.028			
Tipo de carga	18			
Interação	286			
Erro	102			
Total	1.934			

- b)** Apresente um gráfico apropriado para observar uma possível interação entre os dois fatores.
- c)** Existe um processador que pode ser considerado superior aos outros? Explique.

EXERCÍCIOS COMPLEMENTARES

15. No desenvolvimento de um sistema de reconhecimento de fala, fez-se um experimento para avaliar dois tipos de parâmetros acústicos: MFCC (Componentes Mel Cepstrais) e NMF (Componentes Mel Cepstrais Normalizados). Foram observadas duas amostras independentes com cada tipo de parâmetro e anotadas as taxas de acerto (em %):¹³

MFCC:	78,67	81,00	84,67	80,97	81,46
	85,12	80,32	80,95	84,76	
NMF:	86,67	88,33	92,67	88,05	89,76
	93,66	86,03	87,94	91,75	

- a)** Há evidência de diferença entre os dois parâmetros acústicos, em termos da taxa média de acertos? Use $\alpha = 0,01$.
- b)** E em termos de variabilidade? Use $\alpha = 0,05$.
- c)** Considere que o experimento tenha sido feito de forma pareada, isto é, em cada conjunto de ensaios, para reconhecimento de fala, o sistema rodava com os dois parâmetros (MFCC e NMF) paralelamente. Suponha que os resultados apresentados estão na ordem correta dos pares. Refaça o item (a).
- 16.** Com respeito ao exercício anterior, também foi avaliado o parâmetro RNMF (Componentes Mel Cepstrais Normalizados em Tempo Real), com os seguintes resultados:

RNMF:	77,67	81,67	89,69	80,73	82,43
	90,97	77,14	81,27	88,89	

¹³ Os exercícios 15 a 17 foram baseados em trabalho de disciplina dos acadêmicos Ruy Seara Jr. e Izabel Seara, no Curso de Pós-Graduação em Ciências da Computação/UFSC, 2003.

- a)** Analisando como um projeto completamente aleatorizado com um fator (parâmetro), verifique se há diferenças significativas entre os três parâmetros. Use $\alpha = 0,01$.
- b)** Em cada conjunto de ensaios para reconhecimento de fala, o sistema rodava com os três parâmetros (MFCC, NMF e RNMF) paralelamente. Analisando como um projeto em blocos (nove blocos e três tratamentos), verifique se há diferenças significativas entre os três parâmetros. Use $\alpha = 0,01$.

17. Ainda com respeito ao Exercício 15, foram avaliados alguns ruídos de preenchimento (nenhum, uniforme e R1), que podem melhorar o reconhecimento de fala. O projeto foi o fatorial com dois fatores (parâmetro e ruído), tendo três replicações, e produzindo os resultados:

Parâmetro	Ruído de preenchimento								
	Nenhum			Uniforme			R1		
MFCC	78,67	81,00	84,67	80,97	81,46	85,12	80,32	80,95	84,76
NMF	86,67	88,33	92,67	88,05	89,76	93,66	86,03	87,94	91,75
RNMF	77,67	81,67	89,69	80,73	82,43	90,97	77,14	81,27	88,89

Realize uma ANOVA para verificar se existe efeito significativo de Parâmetro, Ruído e interação. Use $\alpha = 0,01$.

18. Um produto usado como piso na “maternidade” da criação de suínos é fabricado por injeção de plástico. Na tentativa de melhorar a qualidade do produto, realizou-se um experimento, variando os fatores: (A) tempo de resfriamento, (B) temperatura do fluido, (C) percentual de elastrômetro e (D) percentual de talco, de acordo com um projeto fatorial 2^4 com duas replicações. A variável resposta foi a dureza (HRc) do material produzido. Os resultados foram:¹⁴

¹⁴ Parte dos dados experimentais do trabalho de dissertação (Engenharia de Produção) de Morgana Pizzolato. UFRGS, 2002.

A	B	C	D	y ₁	y ₂
-1	-1	-1	-1	51,7	49,4
-1	-1	-1	1	75,2	76,0
-1	-1	1	-1	65,9	60,0
-1	-1	1	1	63,7	69,6
-1	1	-1	-1	71,1	72,9
-1	1	-1	1	67,1	69,1
-1	1	1	-1	71,9	75,1
-1	1	1	1	52,8	57,6
1	-1	-1	-1	74,5	67,0
1	-1	-1	1	54,5	70,3
1	-1	1	-1	71,3	70,5
1	-1	1	1	73,4	74,3
1	1	-1	-1	58,5	58,5
1	1	-1	1	49,2	50,2
1	1	1	-1	71,8	71,5
1	1	1	1	72,4	66,6

Calcule os efeitos de cada fator isoladamente (efeitos principais). Teste a significância de cada efeito principal ao nível de significância de 10%.

- 19.** Estudo do desempenho de uma central de comutação telefônica para serviços especializados, tais como: auxílio à lista, hora certa etc. Fatores em estudo: (A) número de linhas ligadas à rede pública (250 ou 400), (B) número de filtros digitais para o reconhecimento de cifras do cliente que disca (80 ou 120), (C) número de canais de máquinas de mensagem (180 ou 270) e (D) número de atendentes conectadas ao sistema (200 ou 300). A variável resposta foi o tempo de resposta ao usuário, considerando o momento do atendimento até o cliente receber a resposta de sua solicitação. Foram simuladas respostas, segundo um projeto 2^4 com quatro replicações. Os resultados estão apresentados a seguir, onde -1 refere-se ao nível inferior do fator e +1 ao nível superior.¹⁵

¹⁵ Parte da dissertação de mestrado de Marcus Vinícius Silva, Ciência da Computação, UFSC, 2002.

A	B	C	D	Tempo (s)			
-1	-1	-1	-1	51,58	51,33	51,44	51,41
-1	-1	-1	1	42,31	42,22	42,16	42,22
-1	-1	1	-1	51,58	51,33	51,44	51,41
-1	-1	1	1	42,31	42,22	42,16	42,22
-1	1	-1	-1	51,58	51,33	51,44	51,41
-1	1	-1	1	42,19	42,21	42,04	42,24
-1	1	1	-1	51,58	66,59	51,44	51,41
-1	1	1	1	42,19	49,95	42,04	42,24
1	-1	-1	-1	66,87	66,59	66,81	66,58
1	-1	-1	1	50,23	49,95	49,93	49,62
1	-1	1	-1	66,87	66,59	66,81	66,58
1	-1	1	1	50,23	49,95	49,93	49,62
1	1	-1	-1	66,87	66,88	66,54	66,81
1	1	-1	1	49,42	49,41	49,40	49,05
1	1	1	-1	66,87	66,88	66,54	66,81
1	1	1	1	49,42	49,41	49,40	49,05

- a) Calcule os efeitos de cada fator isoladamente (efeitos principais) e as interações entre dois fatores.
- b) Teste a significância ($\alpha = 0,05$) de cada efeito calculado no item (a).
20. Um processo químico é influenciado pelo tempo e pela temperatura de reação. Um experimento é realizado para diferentes níveis do tempo de reação (20, 25 e 30 minutos) e da temperatura de reação (60, 70 e 80^o C), segundo um projeto fatorial com seis replicações. A resposta é o rendimento (em %). Usando a planilha eletrônica Excel, obtiveram-se os seguintes resultados:¹⁶

Médias por tratamento			
Temperatura (^o C)	Tempo (minutos)		
	20	25	30
60	30,5	31,4	34,0
70	35,0	38,2	38,2
80	35,6	35,6	35,3

¹⁶ No Excel, acionar “Ferramentas”, “Análise de dados” e “Anova: duplo fator com repetição”.

ANOVA

Fonte da variação	SQ	gl	QM	f	Valor-P	F crítico
Temperatura	251,25	2	125,62	18,64	1,3E-06	3,20
Tempo	41,84	2	20,92	3,10	0,0546	3,20
Interação	38,77	4	9,69	1,44	0,2370	2,58
Erro	303,27	45	6,74			
Total	635,12	53				

Interprete os resultados.

Testes Não Paramétricos

Os testes descritos no Capítulo 9 são ditos *paramétricos*, porque supõem que os dados seguem determinada distribuição de probabilidades – no caso, a *distribuição normal*. Imagine que as suposições necessárias para a aplicação dos testes paramétricos não sejam satisfeitas. Suponha que ocorram alguns dos casos abaixo:

1. Os dados sob análise têm um nível de mensuração *qualitativo*: ordinal ou nominal.
2. Os dados sob análise têm nível de mensuração quantitativo, mas há indícios de que a distribuição populacional *não é normal*.
3. Há interesse em realizar inferência sobre outras características da população, além dos parâmetros de sua distribuição, como a própria forma da distribuição.

Uma alternativa para essas situações é a utilização dos *testes não paramétricos*, ou *testes livres de distribuição*. As suposições necessárias para a aplicação desses testes são menos rígidas que as dos paramétricos, possibilitando uma aplicação mais generalizada. Obviamente, suposições mais relaxadas significam que o *poder estatístico* de um teste não paramétrico é inferior ao teste paramétrico equivalente, mas, em muitos casos, é a única alternativa disponível para análise dos dados e realização da inferência estatística.

Este texto abordará basicamente três tipos de testes não paramétricos: testes de aderência, testes de independência e testes de comparação de populações. A Figura 8.1 permite visualizar alguns testes não paramétricos.

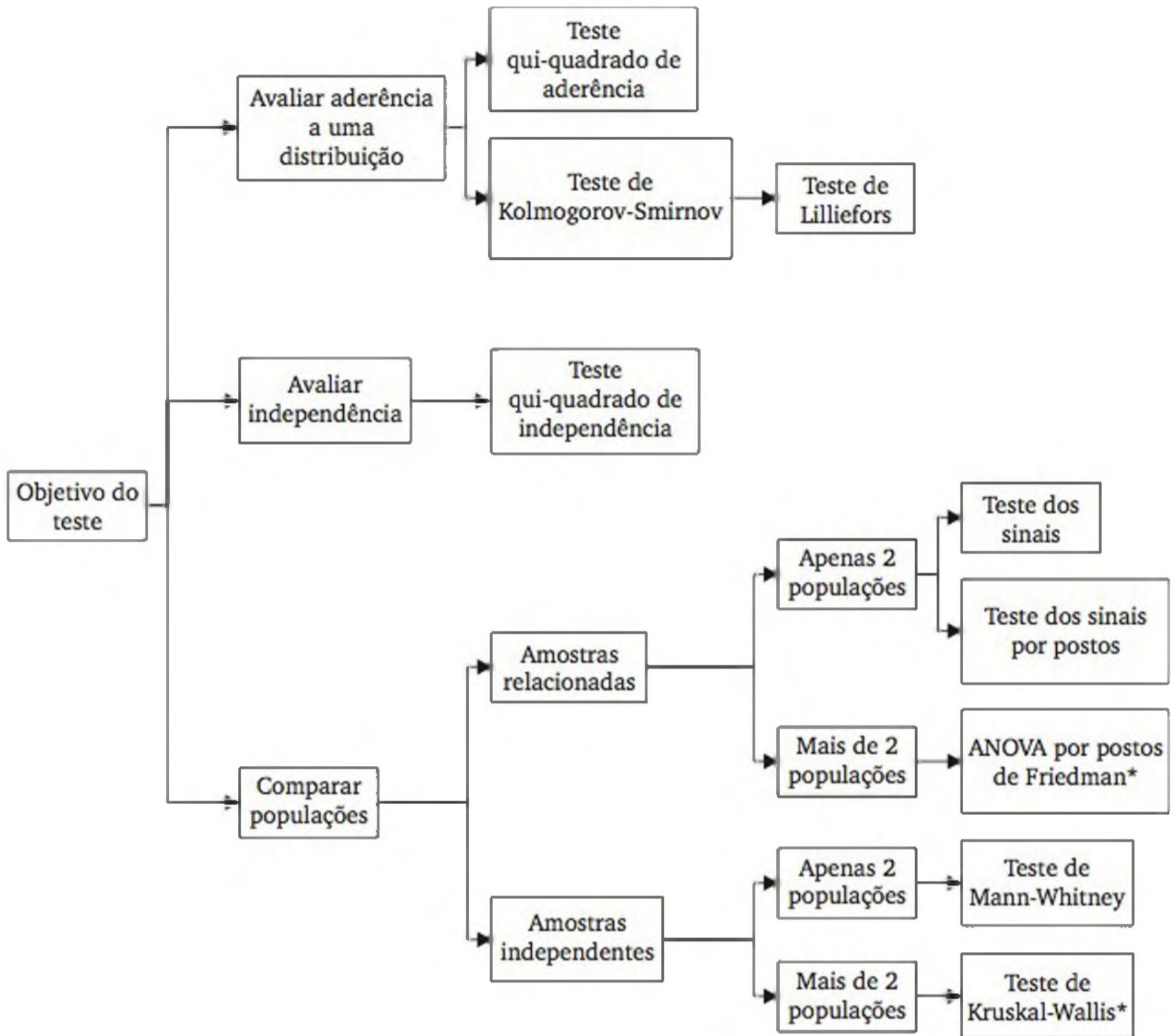


Figura 10.1 Alguns testes não paramétricos. Os testes marcados com asterisco não serão abordados neste texto.

As próximas seções irão detalhar alguns dos testes não paramétricos mais importantes, mostrando as condições para sua aplicação e apresentando exemplos.

10.1 TESTES DE ADERÊNCIA

O objetivo de um teste de aderência é verificar se os dados de uma amostra comportam-se de acordo com uma distribuição teórica. Essa distribuição teórica pode ser uma distribuição de probabilidades clássica (como normal, exponencial etc.), ou proporções definidas especificamente para o problema (por exemplo, a Lei de Mendel para ervilhas lisas e rugosas).

10.1.1 Teste qui-quadrado de aderência

O teste *qui-quadrado de aderência* pode ser aplicado quando estamos estudando dados distribuídos em categorias e há interesse em verificar se as frequências *observadas* nas K diferentes categorias ($O_i, i = 1, 2, \dots, K$) são significativamente distintas de um conjunto de K frequências esperadas ($E_i, i = 1, 2, \dots, K$).¹ As hipóteses são:

$$H_0: O_i = E_i \text{ para todo } i = 1, 2, \dots, K;$$

$$H_1: O_i \neq E_i \text{ para algum } i = 1, 2, \dots, K.$$

A estatística desse teste, chamada de Q^2 , é uma espécie de medida de distância entre as frequências observadas e as frequências esperadas de cada categoria. Sua expressão é dada por:

$$Q^2 = \sum_{i=1}^K \left[\frac{(O_i - E_i)^2}{E_i} \right] \quad (10.1)$$

Havendo aderência (H_0 verdadeira), as frequências observadas devem ficar próximas das esperadas, acarretando um valor pequeno para Q^2 : *as variações encontradas seriam apenas casuais*. Contudo, se não houver aderência (H_1 verdadeira), diferenças entre frequências observadas e esperadas poderão ser grandes, resultando em um valor grande para Q^2 : *é pouco provável que as variações tenham sido casuais*.

Supondo amostra grande (digamos, $E_i > 5$ para todo $i = 1, 2, \dots, K$) e supondo a hipótese de aderência (H_0 verdadeira), a estatística Q^2 segue aproximadamente uma distribuição *qui-quadrado com $K - 1$ graus de liberdade*. Observada uma amostra, podemos obter o valor da estatística Q^2 , o qual representaremos por q^2 . Usando a abordagem clássica, obtemos o valor crítico χ_c^2 , em função do nível de significância α adotado, formando a seguinte regra de decisão:

$$q^2 < \chi_c^2 \Rightarrow \text{aceita } H_0 \text{ (há aderência à distribuição especificada)}$$

$$q^2 \geq \chi_c^2 \Rightarrow \text{rejeita } H_0 \text{ (não há aderência à distribuição especificada)}$$

Exemplo 10.1 Determinado veículo utilitário está sofrendo pesadas críticas de seus proprietários, com relação à grande frequência de defeitos no pneu traseiro esquerdo. Preocupado com sua imagem, e procurando defender-se de

¹ Os dados em categorias podem provir de uma *variável qualitativa* ou mesmo de uma *variável quantitativa* (discreta ou contínua). No caso de dados de variáveis contínuas, torna-se necessário categorizá-los, formando uma distribuição de frequências com dados grupados (Capítulo 3).

eventuais pedidos de indenização, o fabricante do veículo resolveu coletar informações sobre 152 ocorrências de defeitos, classificando-as por posição do pneu. Os resultados estão na Tabela 10.1. Usando nível de significância de 5%, há razão para acreditar que a probabilidade de defeito é diferente para alguma das posições?

Tabela 10.1 Ocorrências de defeitos por posição do pneu de um veículo utilitário.

Posição do pneu	Dianteiro esquerdo	Dianteiro direito	Traseiro esquerdo	Traseiro direito	Total
Frequência	35	32	57	28	152

Para responder à questão colocada, temos as seguintes hipóteses:

H_0 : as frequências de defeitos nos quatro pneus são iguais;

H_1 : pelo menos uma das frequências de defeitos nos pneus é diferente.

Se H_0 for verdadeira, as 152 ocorrências devem distribuir-se igualmente pelas quatro categorias, resultando na frequência esperada de cada categoria igual a $152/4 = 38$ ocorrências. Calculando q^2 :

Dianteiro esquerdo	Dianteiro direito	Traseiro esquerdo	Traseiro direito
$(35 - 38)^2/38 = 0,237$	$(32 - 38)^2/38 = 0,947$	$(57 - 38)^2/38 = 9,5$	$(28 - 38)^2/38 = 2,632$

E, portanto:

$$q^2 = 0,237 + 0,947 + 9,5 + 2,632 = 13,316$$

Abordagem clássica.² Para construir a regra de decisão, obtemos o valor crítico χ_c^2 da distribuição qui-quadrado, com $gl = K - 1 = 4 - 1 = 3$ graus de liberdade, o qual leva à $P(Q^2 > \chi_c^2) = \alpha = 0,05$. Usando a Tabela 5 do apêndice, obtém-se $\chi_c^2 = 7,815$, tal como mostrado na Figura 10.2.

² Neste capítulo, adotaremos apenas a abordagem clássica na resolução dos exemplos, embora o uso da abordagem do valor p possa ser realizado de forma análoga aos capítulos anteriores.

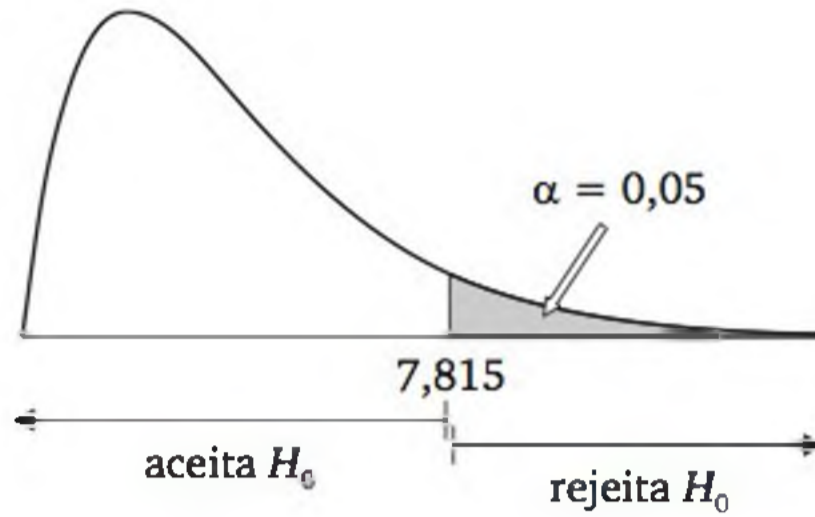


Figura 10.2 Regra de decisão de um teste de aderência usando o modelo qui-quadrado com 3 graus de liberdade.

Como $q^2 = 13,316$ cai na região de rejeição, o teste rejeita H_0 em favor de H_1 . Assim, há evidência de que as frequências de ocorrência dos defeitos dependem da posição do pneu.

10.1.2 Teste de Kolmogorov-Smirnov

Considere uma situação em que desejamos verificar a aderência de um conjunto de valores em relação a uma *distribuição de probabilidades especificada* (discreta ou contínua). Embora seja possível aplicar o teste qui-quadrado de aderência, geralmente é melhor aplicar o chamado *teste de aderência de Kolmogorov-Smirnov*, que é uma alternativa mais poderosa do que o teste qui-quadrado, nestas situações.

Seja $F(x)$ a *função de distribuição acumulada*, com parâmetros *especificados*, para a qual se quer verificar a aderência dos dados. As hipóteses são:

H_0 : os dados provêm de $F(x)$ (há aderência);

H_1 : os dados *não* provêm de $F(x)$ (*não* há aderência).

Sejam as distribuições de frequências acumuladas: a empírica, $S(x)$, e a teórica, $F(x)$. Para cada elemento da amostra, obtém-se a diferença absoluta entre essas duas distribuições. A estatística do teste é a diferença absoluta máxima, D . O procedimento é descrito a seguir.

1. Definimos $S(x)$ para cada valor x_i ($i = 1, 2, \dots, n$) como:

$$S(x_i) = \frac{\text{número de valores} \leq x_i}{n} \quad (10.2)$$

onde n é o tamanho da amostra e x_i é um valor qualquer da amostra.

2. Obtemos, para cada valor x_i ($i = 1, 2, \dots, n$), os valores teóricos $F(x_i)$, calculados pela função de distribuição acumulada $F(x)$, especificada em H_0 . Observe que, se a amostra realmente provém de uma população que segue a distribuição teórica $F(x)$, a distribuição observada $S(x)$ não deve afastar-se muito de $F(x)$.
3. Verificamos a discrepância entre $S(x)$ e $F(x)$ através das diferenças absolutas entre $F(x_i)$ e $S(x_i)$, e entre $F(x_i)$ e $S(x_{i-1})$, para $i = 1, 2, \dots, n$.
4. Calculamos a estatística de teste, D , em termos da amostra em análise:

$$d = \max_i \{ |F(x_i) - S(x_i)|, |F(x_i) - S(x_{i-1})| \} \quad (10.3)$$

5. Uma vez identificada a distância máxima d (valor de D para a particular amostra em análise), comparamos seu valor com um valor tabelado, d_c , de acordo com o nível de significância α e do tamanho n da amostra (ver Tabela 7 do apêndice). Regra de decisão:
 - $d < d_c \Rightarrow$ aceita H_0 (há aderência à distribuição especificada);
 - $d \geq d_c \Rightarrow$ rejeita H_0 (não há aderência à distribuição especificada).

Exemplo 10.2 Um fabricante de autopeças está próximo de fechar um grande contrato com uma montadora. O ponto-chave é a garantia da qualidade de seus produtos, especialmente do diâmetro (em mm) dos eixos produzidos, que ele supõe seguir uma distribuição normal com média 100 e desvio padrão 2. A montadora selecionou uma amostra aleatória de 15 eixos, para testar as especificações a 5% de significância. Os valores estão descritos a seguir:

93,45 94,46 94,93 96,17 96,74 97,07 97,68 97,93 99,10
 99,30 100,73 103,29 103,60 103,83 105,20

O diâmetro é uma variável contínua e desejamos verificar a aderência a uma distribuição normal com média e desvio padrão fornecidos. Assim, o teste de aderência de Kolmogorov-Smirnov pode ser empregado. As hipóteses são:

H_0 : a amostra provém de uma população que segue uma distribuição normal com $\mu = 100$ e $\sigma = 2$ mm;

H_1 : a amostra *não* provém de uma população que segue distribuição normal com $\mu = 100$ e $\sigma = 2$ mm.

1. Construção da distribuição acumulada da amostra, $S(x)$:

Valores ordenados	Frequência relativa	$S(x_i)$
93,45	1/15	1/15 = 0,067
94,46	1/15	2/15 = 0,133
94,93	1/15	3/15 = 0,200
96,17	1/15	4/15 = 0,267
96,74	1/15	5/15 = 0,333
97,07	1/15	6/15 = 0,400
97,68	1/15	7/15 = 0,467
97,93	1/15	8/15 = 0,533
99,10	1/15	9/15 = 0,600
99,30	1/15	10/15 = 0,667
100,73	1/15	11/15 = 0,733
103,29	1/15	12/15 = 0,800
103,60	1/15	13/15 = 0,867
103,83	1/15	14/15 = 0,933
105,20	1/15	15/15 = 1,000

2. Construção da função de distribuição acumulada $F(x)$, para cada valor x_i ($i = 1, 2, \dots, n$). Cada valor de diâmetro x pode ser convertido em escore padronizado z . Por exemplo, para $x_1 = 93,45$, temos:

$$z_1 = \frac{x_1 - \mu}{\sigma} = \frac{93,45 - 100}{2} = -3,28$$

A probabilidade acumulada até cada escore z é obtida da Tabela 3, resultando nos seguintes valores $F(x_i)$:

x_i	z_i	$F(x_i)$
		0
93,45	- 3,28	0,001
94,46	- 2,77	0,003
94,93	- 2,54	0,006
96,17	- 1,92	0,028
96,74	- 1,63	0,052
97,07	- 1,47	0,071
97,68	- 1,16	0,123
97,93	- 1,04	0,150
99,10	- 0,45	0,326
99,30	- 0,35	0,363
100,73	0,37	0,643
103,29	1,64	0,950
103,60	1,80	0,964
103,83	1,91	0,972
105,20	2,60	0,995

3. Obtenção das diferenças absolutas entre as distribuições acumuladas esperadas e observadas, $|F(x_i) - S(x_i)|$ e $|F(x_i) - S(x_{i-1})|$:

x_i	$S(x_i)$	$F(x_i)$	$ F(x_i) - S(x_{i-1}) $	$ F(x_i) - S(x_i) $
	0	0		
93,45	0,067	0,001	0,001	0,066
94,46	0,133	0,003	0,064	0,131
94,93	0,200	0,006	0,128	0,194
96,17	0,267	0,028	0,172	0,239
96,74	0,333	0,052	0,215	0,282
97,07	0,400	0,071	0,262	0,329
97,68	0,467	0,123	0,277	0,344
97,93	0,533	0,150	0,316	0,383
99,10	0,600	0,326	0,207	0,274
99,30	0,667	0,363	0,237	0,304
100,73	0,733	0,643	0,024	0,090
103,29	0,800	0,950	0,216	0,150
103,60	0,867	0,964	0,164	0,097
103,83	0,933	0,972	0,106	0,039
105,20	1,000	0,995	0,062	0,005

Distância máxima

4. A maior diferença absoluta foi igual a 0,383; logo, $d = 0,383$.
5. Pela Tabela 7, para $n = 15$ e $\alpha = 0,05$, obtemos a distância máxima admissível, $d_c = 0,338$. Como $d > d_c$, o teste rejeita H_0 , concluindo que não há aderência dos dados à distribuição normal com $\mu = 100$ e $\sigma = 2$. Observe o gráfico das duas distribuições acumuladas na Figura 10.3.

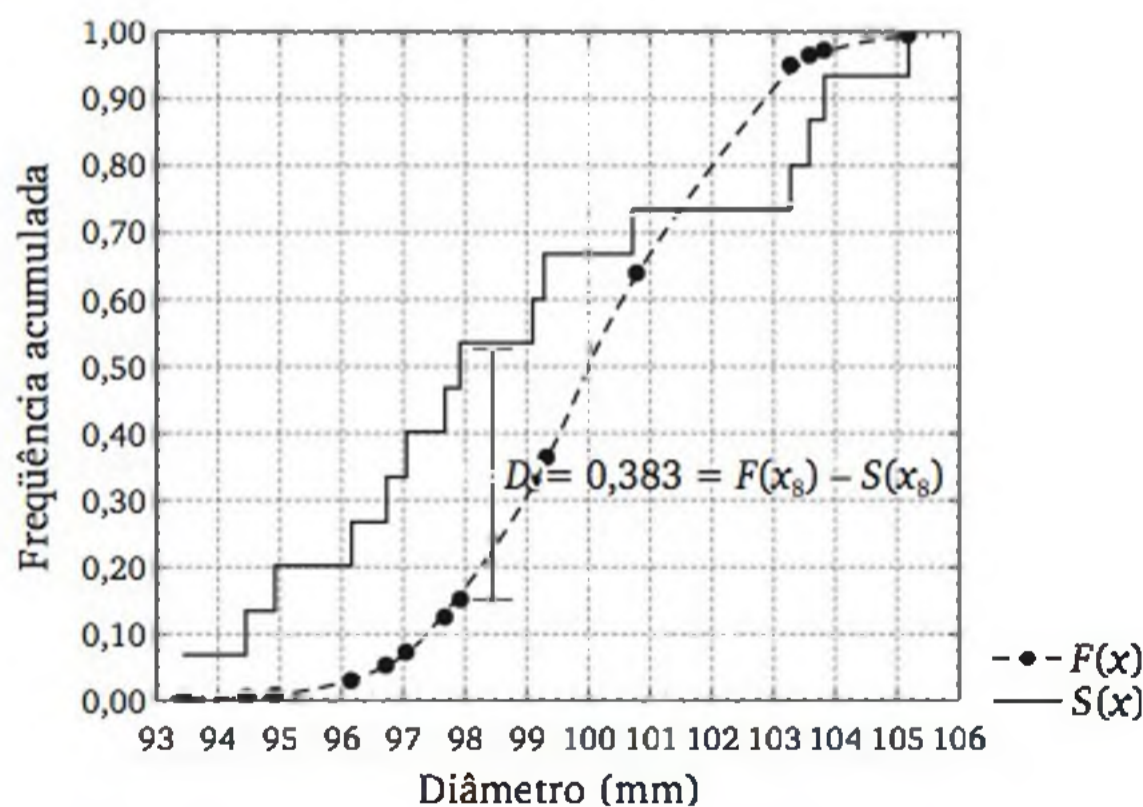


Figura 10.3 Distribuições acumuladas $F(x)$ e $S(x)$ para o Exemplo 10.2.

Exemplo 10.3 A metodologia usada para calcular os índices de confiabilidade de um sistema de transmissão de energia elétrica exige que os tempos para a falha dos componentes sigam distribuições exponenciais. Observações anteriores indicaram a validade de tal suposição, mas um engenheiro decidiu verificar se o tempo para a falha (em horas) de um componente, especialmente crítico, pode ser admitido com distribuição exponencial de média de 500 horas. Para testar essa hipótese, utilizando nível de significância de 1%, coletou-se uma amostra de 20 observações do tempo de falha desse componente:

7,55 25,20 41,00 133,59 146,77 157,55 158,07 206,08 385,09 426,89
555,86 639,43 816,11 847,57 924,63 945,66 968,66 1.130,39 1.143,93 1.365,69

As hipóteses são:

H_0 : a amostra provém de uma população que segue uma distribuição exponencial com média 500;

H_1 : a amostra *não* provém de uma população que segue uma distribuição exponencial com média 500.

A distribuição acumulada da amostra, $S(x)$, é construída da mesma forma que no exemplo anterior. Já para os valores teóricos da distribuição exponencial de média 500, lembramos que o parâmetro λ de uma distribuição exponencial é o inverso de sua média (ver Expressão 6.17, Capítulo 6). Assim, pela hipótese nula:

$$\lambda = \frac{1}{E(X)} = \frac{1}{500} = 0,002$$

Logo, os valores teóricos associados a x_i ($i = 1, 2, \dots, n$) são obtidos por (ver Expressão 6.14, Capítulo 6):

$$F(x_i) = P(X \leq x_i) = 1 - e^{-\lambda x_i} = 1 - e^{-0,002x_i}$$

Os resultados dessa expressão e os demais cálculos para o teste são apresentados a seguir:

x_i	$S(x_i)$	$F(x_i)$	$ F(x_i) - S(x_{i-1}) $	$ F(x_i) - S(x_i) $
	0	0		
7,55	0,05	0,015	0,015	0,035
25,20	0,10	0,049	0,001	0,051
41,00	0,15	0,079	0,021	0,071
133,59	0,20	0,234	0,084	0,034
146,77	0,25	0,254	0,054	0,004
157,55	0,30	0,270	0,020	0,030
158,07	0,35	0,271	0,029	0,079
206,08	0,40	0,338	0,012	0,062
385,09	0,45	0,537	0,137	0,087
426,89	0,50	0,574	0,124	0,074
555,86	0,55	0,671	0,171	0,121
639,43	0,60	0,722	0,172	0,122
816,11	0,65	0,805	0,205	0,155
847,57	0,70	0,816	0,166	0,116
924,63	0,75	0,843	0,143	0,093
945,66	0,80	0,849	0,099	0,049
968,66	0,85	0,856	0,056	0,006
1.130,39	0,90	0,896	0,046	0,004
1.143,93	0,95	0,899	0,001	0,051
1.365,69	1,00	0,935	0,015	0,065

Distância máxima

A maior diferença absoluta foi igual a $d = 0,205$. Procurando na Tabela 7, para $n = 20$ e $\alpha = 0,01$, obtemos uma distância máxima admissível, $d_c = 0,352$. Como $d < d_c$, o teste aceita H_0 ao nível de significância de 1%, indicando aderência dos dados à distribuição exponencial de média 500. Observe o gráfico das duas distribuições acumuladas na Figura 10.4.

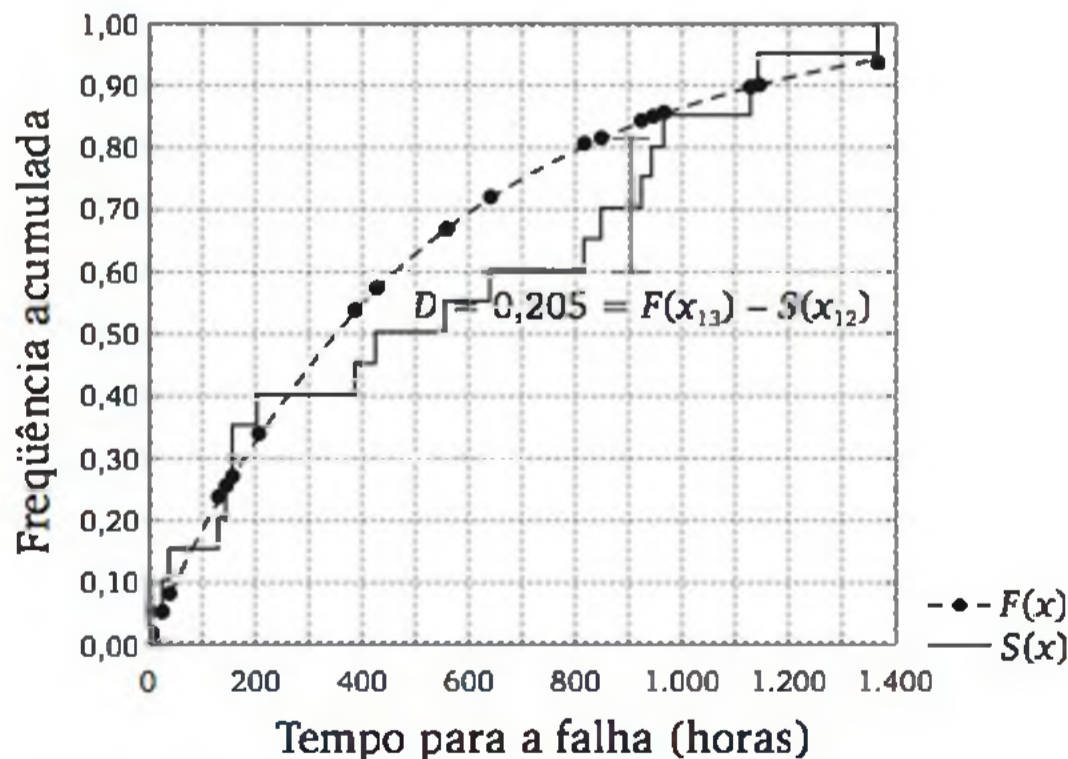


Figura 10.4 Distribuições acumuladas $F(x)$ e $S(x)$ para o Exemplo 10.3.

O teste de Kolmogorov-Smirnov pode ser aplicado para avaliar a aderência a *qualquer* distribuição, desde que seus parâmetros sejam especificados.

10.1.3 Teste de Lilliefors

O teste de Lilliefors é usado para verificar a aderência dos dados a uma distribuição normal qualquer, isto é, sem a especificação de seus parâmetros. É bastante parecido com o teste de aderência de Kolmogorov-Smirnov, pois também avaliamos as distribuições acumuladas $S(x)$ e $F(x)$; obtemos a distância máxima D entre elas; e a comparamos com um valor tabelado, em função do nível de significância e do tamanho da amostra. As diferenças residem na forma de obtenção de $F(x)$ (pois a média e o desvio padrão são calculados com base na amostra) e na tabela utilizada para a decisão do teste (ver lado direito da Tabela 7, no apêndice).

Exemplo 10.4 No Exemplo 10.2, não houve aderência dos dados à distribuição normal com média 100 e desvio padrão 2. Mas a montadora quer saber se é possível considerar que os diâmetros dos eixos distribuem-se segundo uma distribuição normal (qualquer), ao nível de significância de 5%.

As hipóteses são:

H_0 : a amostra provém de uma população que segue uma distribuição normal;

H_1 : a amostra *não* provém de uma população que segue uma distribuição normal.

Calculando a média e o desvio padrão da amostra, que servirão como estimativas dos parâmetros da distribuição normal, na população:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = 98,90$$

$$s = \sqrt{\frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right)} = 3,70$$

Para cada valor x_i ($i = 1, 2, \dots, n$), calculamos o correspondente escore z_i , usando \bar{x} e s no lugar de μ e σ , respectivamente. Por exemplo, para $x_1 = 93,45$, temos:

$$z_1 = \frac{x_1 - \bar{x}}{s} = \frac{93,45 - 98,90}{3,70} = -1,71$$

E os resultados para o cálculo da estatística do teste:

x_i	$S(x_i)$	z_i	$F(x_i)$	$ F(x_i) - S(x_{i-1}) $	$ F(x_i) - S(x_i) $
	0		0		
93,45	0,067	- 1,47	0,071	0,071	0,004
94,46	0,133	- 1,20	0,115	0,049	0,018
94,93	0,200	- 1,07	0,142	0,009	0,058
96,17	0,267	- 0,74	0,230	0,030	0,036
96,74	0,333	- 0,58	0,280	0,013	0,053
97,07	0,400	- 0,49	0,311	0,023	0,089
97,68	0,467	- 0,33	0,371	0,029	0,096
97,93	0,533	- 0,26	0,397	0,070	0,136
99,10	0,600	0,05	0,522	0,012	0,078
99,30	0,667	0,11	0,543	0,057	0,124
100,73	0,733	0,50	0,690	0,023	0,043
103,29	0,800	1,19	0,882	0,149	0,082
103,60	0,867	1,27	0,898	0,098	0,031
103,83	0,933	1,33	0,909	0,042	0,025
105,20	1,000	1,70	0,956	0,022	0,044

Distância máxima

A estatística do teste para esta amostra é $d = 0,149$. No lado direito da Tabela 7, para $\alpha = 0,05$ e $n = 15$ obtemos a distância máxima admissível $d_c = 0,220$. Como $d < d_c$, o teste *aceita* H_0 ao nível de significância de 5%, concluindo que há aderência dos dados a uma distribuição normal, embora com parâmetros diferentes dos especificados pela montadora no Exemplo 10.2. Veja o gráfico das duas distribuições, exposto na Figura 10.5.

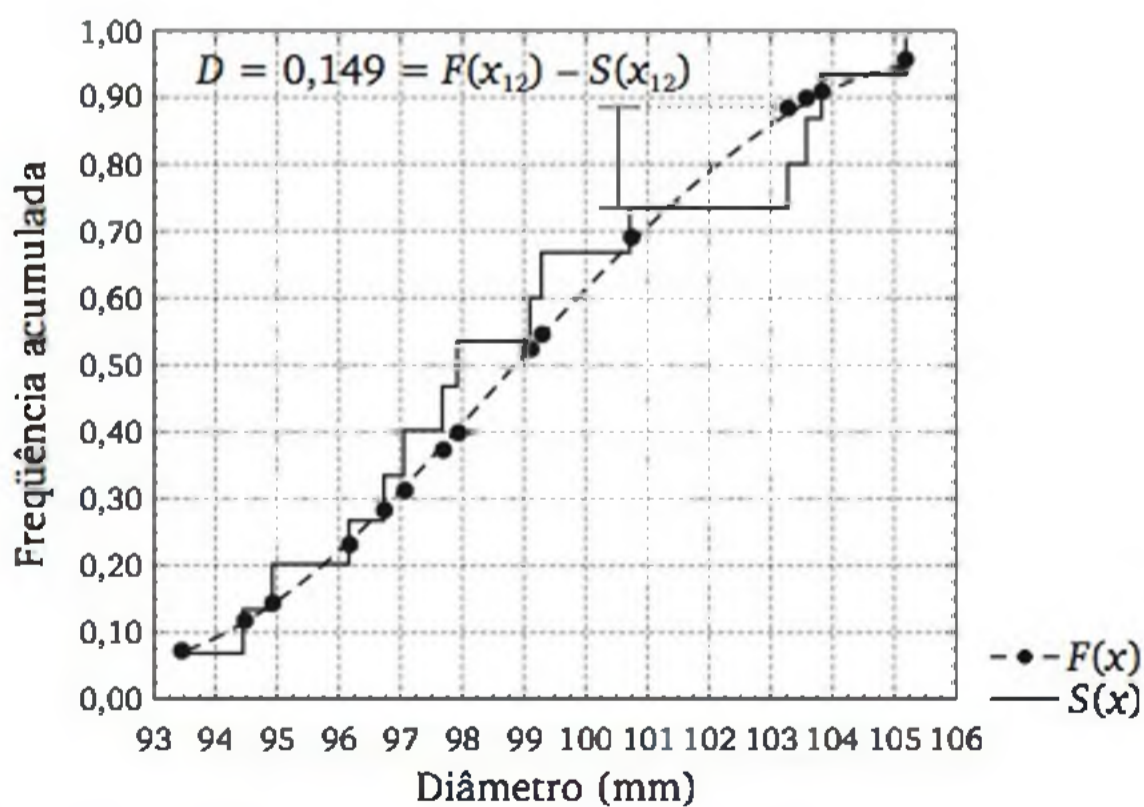


Figura 10.5 Distribuições acumuladas $F(x)$ e $S(x)$ para o Exemplo 10.4.

A Figura 10.5 mostra grande proximidade entre as curvas das distribuições $S(x)$ e $F(x)$, razão pela qual o teste de Lilliefors aceitou a hipótese de aderência à distribuição normal.

Rotineiramente, quando estamos estudando uma variável quantitativa, o teste de Lilliefors é utilizado para avaliar se é possível aplicar um teste paramétrico que supõe distribuição normal. Quando não há aderência, pode ser necessário usar uma técnica não paramétrica alternativa, sendo que algumas delas serão discutidas nas próximas seções.

EXERCÍCIOS

1. Uma empresa possui três laboratórios de pesquisa (A, B, C), cujos computadores estão conectados a um servidor, para onde enviam pacotes de dados para serem analisados em um programa estatístico (disponível apenas no servidor). Os usuários do laboratório A pediram prioridade ao gerente de rede, pois costumam enviar mais pacotes ao servidor. O gerente observou 500 pacotes de dados enviados e classificou-os de acordo com a origem, conforme a tabela a seguir:

Laboratório	A	B	C	Total
Número de pacotes	165	179	156	500

Os dados constituem evidência suficiente para corroborar o pedido do laboratório A? Utilize nível de significância de 1%.

2. Quatro fábricas de um mesmo grupo produzem peças automotivas. Historicamente, cerca de 20% do total de peças defeituosas vêm da fábrica 1, 30% da fábrica 2, 25% da fábrica 3 e 25% da fábrica 4. Os engenheiros do setor da qualidade estão suspeitando que essas proporções não são mais válidas, e observaram o número de peças defeituosas no último lote de produção. Os dados são apresentados a seguir:

Fábrica	1	2	3	4	Total
Número de peças	120	200	175	134	629

Utilizando nível de significância de 5%, a suspeita dos engenheiros do setor da qualidade tem fundamento?

3. Você e seu sócio têm um provedor de acesso à Internet. Para auxiliar na configuração dos equipamentos, o conhecimento do comportamento dos

tempos de acesso de seus clientes é importante. Mas vocês têm divergência a esse respeito: você crê que o tempo de acesso segue uma distribuição *normal* com média de 16 minutos e desvio padrão de 4 minutos, enquanto seu sócio suspeita que o tempo de acesso segue uma distribuição *exponencial*, com média de 16 minutos. Para dirimir a dúvida, vocês coletaram uma amostra de tempos de acesso (em minutos):

6 10 1 11 1 6 27 9 17 1 18 10 5 35 17 41 2 20 16 5

Utilizando nível de significância de 5%, qual de vocês dois está certo quanto à distribuição dos tempos de acesso?

4. No controle estatístico de processos, uma suposição crucial para a utilização de gráficos de controle de média de Shewhart é de que a distribuição das médias possa ser considerada normal. Um engenheiro quer saber se é possível aplicar gráficos de controle de médias a um processo produtivo. Para tanto, quer avaliar a aderência das médias de 25 amostras à distribuição normal. Os valores estão expressos a seguir:

0,19 0,57 0,66 1,41 0,28 0,05 0,63 0,75 0,85 0,99

1,68 3,01 0,31 5,48 0,66 0,76 5,94 0,85 0,03 9,49

2,18 1,23 4,89 0,71 3,52

Com base nos dados apresentados, e supondo nível de significância de 1%, é possível usar gráficos de controle de média de Shewhart para monitorar o processo?

5. A empresa “Faça Certo S.A.” produz peças automotivas da melhor qualidade, e uma delas tem aspecto crítico na característica dimensão, a qual precisa seguir uma distribuição normal (aspecto mais importante), com média de 15 mm e desvio padrão de 0,4 mm. Após receber muitas reclamações, o departamento de vendas cobrou do departamento de produção um teste desta suposição. Uma amostra de 40 peças foi aleatoriamente coletada e suas dimensões críticas medidas, resultando:

11,37 11,38 11,51 11,57 11,58 11,58 11,58 11,65 11,72 11,75

11,78 11,78 11,83 11,90 11,90 11,92 11,93 11,93 11,97 12,00

12,01 12,06 12,06 12,06 12,07 12,09 12,11 12,11 12,14 12,19

12,20 12,21 12,23 12,23 12,25 12,46 12,48 12,52 12,65 12,81

a) A dimensão crítica segue uma distribuição normal com média de 15 mm e desvio padrão de 0,4 mm? Use $\alpha = 0,05$.

b) A dimensão crítica segue uma distribuição normal? Use $\alpha = 0,05$.

10.2 ANÁLISE DE ASSOCIAÇÃO

Dizemos que existe *associação* entre duas variáveis qualitativas quando as probabilidades de eventos de uma delas são alteradas conforme a categoria da outra. Por exemplo, a probabilidade de passar na disciplina de Estatística deve ser maior se você dedicar mais tempo à disciplina. Então, a dedicação (*alta* ou *baixa*) e aprovação (*sim* ou *não*) têm associação. O chamado *teste qui-quadrado de independência* serve para avaliar a significância de uma associação. O mesmo procedimento que discutiremos para verificar uma possível associação entre duas variáveis qualitativas também pode ser usado para verificar se diferentes populações apresentam as mesmas proporções com respeito a uma variável qualitativa. Nessa situação, o teste é chamado de *teste qui-quadrado de homogeneidade*, mas em termos matemáticos, o teste de homogeneidade é igual ao teste de independência e, por isso, não faremos distinção entre esses dois testes.

10.2.1 Tabelas de contingência

Antes de estudarmos o teste propriamente dito, precisamos conhecer as *tabelas de contingência*, que são a forma usual de apresentar uma distribuição de frequências conjunta de duas variáveis qualitativas.

Exemplo 10.5 Determinado posto de qualidade de um laticínio retira uma amostra dos pesos dos litros de leite produzidos em um dia, classificando-os de acordo com seu *tipo* (B, C, UHT), e *condições de peso* (dentro ou fora das especificações). A Tabela 10.2 mostra a distribuição de frequências conjunta de 6.850 unidades de leite, disposta numa tabela de contingência.

Tabela 10.2 *Distribuição de frequências conjunta do tipo de leite e condição do peso.*³

Condição do peso	Tipo do leite			Total
	B	C	UHT	
Dentro das especificações	500	4.500	1.500	6.500
Fora das especificações	30	270	50	350
Total	530	4.770	1.550	6.850

³ Dados extraídos da dissertação de mestrado de Luciana S. C. V. da Silva (Programa de Pós-Graduação em Engenharia de Produção/UFSC, 2001).

Podemos calcular percentuais para cada célula da tabela, seja em relação ao total das linhas (condições de peso) ou total das colunas (tipo de leite), o que possibilita uma avaliação da qualidade dos processos. Calculando percentuais em relação aos totais das colunas, obtemos:

Condição do peso	Tipo do leite			Total
	B	C	UHT	
Dentro das especificações	94,34%	94,34%	96,77%	94,89%
Fora das especificações	5,66%	5,66%	3,23%	5,11%
Total	100%	100%	100%	100%

Não obstante as diferentes quantidades produzidas, os tipos de leite têm proporções semelhantes para produtos dentro das especificações de peso, com exceção do tipo UHT, que parece estar mais de acordo com as especificações. A seção seguinte mostra como podemos *testar estatisticamente* essa possível associação entre o tipo de leite e a condição de peso.

10.2.2 Teste qui-quadrado de independência

Usado para verificar se existe associação entre duas variáveis qualitativas (categóricas), X e Y , com base em uma amostra de observações disposta numa tabela de contingência com L linhas e C colunas ($L, C \geq 2$), correspondentes às categorias de X e Y , respectivamente. A hipótese nula afirma *independência* entre X e Y , enquanto a hipótese alternativa aponta para *associação* entre X e Y .

Vimos, no Capítulo 4, que dois eventos são *independentes* se e somente se a probabilidade conjunta (da interseção) for igual ao produto das probabilidades de cada evento. Denominam-se:

p_{ij} a probabilidade de ocorrência da linha i e coluna j ($i = 1, 2, \dots, L; j = 1, 2, \dots, C$);

p_i a probabilidade de ocorrência da linha i ; e

p_j a probabilidade de ocorrência da coluna j .

Podemos escrever as hipóteses de independência e associação por (ver seção 4.3.2):

$H_0: p_{ij} = p_i p_j$ para todo $i = 1, 2, \dots, L$ e $j = 1, 2, \dots, C$;

$H_1: p_{ij} \neq p_i p_j$ para algum $i = 1, 2, \dots, L$ e $j = 1, 2, \dots, C$.

Sejam:

O_{ij} a frequência observada na célula (i, j) da tabela de contingências; e E_{ij} a frequência esperada na célula (i, j) , supondo H_0 verdadeira.

Sob H_0 , as frequências esperadas podem ser calculadas por:

$$E_{ij} = \frac{(\text{total da linha } i) \times (\text{total da coluna } j)}{\text{total geral}} \quad (10.4)$$

$(i = 1, 2, \dots, L; j = 1, 2, \dots, C).$ ⁴

A estatística do teste é um tipo de distância entre as frequências observadas e as frequências esperadas por H_0 . Ela é dada por:

$$Q^2 = \sum_{i=1}^L \sum_{j=1}^C \left[\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \right] \quad (10.5)$$

Sob H_0 , a estatística Q^2 segue uma distribuição qui-quadrado com graus de liberdade igual a:

$$gl = (L - 1) (C - 1) \quad (10.6)$$

Adotando nível de significância α , podemos obter χ_c^2 na Tabela 5 do apêndice e construir a regra de decisão para o valor calculado q^2 , como segue:

$q^2 < \chi_c^2 \Rightarrow$ aceita H_0 (as duas variáveis são independentes);

$q^2 \geq \chi_c^2 \Rightarrow$ rejeita H_0 (há associação entre as duas variáveis).

Exemplo 10.5 (continuação) Com os dados do laticínio, vamos testar se há associação entre o *tipo do leite* (B, C, UHT) e *condições de peso* (dentro ou fora das especificações). Temos as hipóteses:

H_0 : as condições de peso *independem* do tipo do leite;

H_1 : há *associação* entre condições de peso e tipos de leite.

Os dados amostrais estão na Tabela 10.2, onde observamos que há 94,89% de produtos dentro das especificações e 5,11% fora. Se as condições de peso forem *independentes* dos tipos de leite (B, C, UHT), devemos esperar que

⁴ Representando por n o total geral, observe que $E_{ij} = n\hat{p}_i\hat{p}_j$, onde \hat{p}_i é a proporção observada na linha i ($i = 1, 2, \dots, L$) e \hat{p}_j é a proporção observada na coluna j ($j = 1, 2, \dots, C$). Sob H_0 , esta quantidade é uma estimativa de np_{ij} .

as porcentagens se mantenham para *todos* os tipos de leite. Assim, como foram observados 530 produtos leite tipo B, devemos esperar uma frequência em torno de 502,917 (94,89% de 530) dentro das especificações. Note que chegaremos ao mesmo resultado se aplicarmos (10.4):

$$E_{11} = \frac{(\text{total da linha 1}) \times (\text{total da coluna 1})}{\text{total geral}} = \frac{(6500)(530)}{6850} = 502,917$$

Aplicando (10.4) para todos (i, j) da tabela (i = 1, 2; j = 1, 2, 3), temos:

	j = 1	j = 2	j = 3	Total
i = 1	$\frac{(6500)(530)}{6850} = 502,917$	$\frac{(6500)(4770)}{6850} = 4526,253$	$\frac{(6500)(1550)}{6850} = 1470,795$	6.500
i = 2	$\frac{(350)(530)}{6850} = 27,083$	$\frac{(350)(4770)}{6850} = 243,747$	$\frac{(350)(1550)}{6850} = 79,205$	350
Total	530	4.770	1.550	6.850

Para aplicar (10.5), podemos calcular a contribuição de cada célula (i, j):

	j = 1	j = 2	j = 3	Total
i = 1	$\frac{(500 - 502,917)^2}{502,917} = 0,017$	$\frac{(4500 - 4526,253)^2}{4526,253} = 0,152$	$\frac{(1500 - 1470,795)^2}{1470,795} = 0,580$	6.500
i = 2	$\frac{(30 - 27,083)^2}{27,083} = 0,314$	$\frac{(270 - 243,747)^2}{243,747} = 2,828$	$\frac{(50 - 79,205)^2}{79,205} = 10,69$	350
Total	530	4.770	1.550	6.850

Donde, para a presente amostra, temos o seguinte valor para a estatística Q^2 :

$$q^2 = 0,017 + 0,152 + 0,58 + 0,314 + 2,828 + 10,769 = 14,66$$

com

$$gl = (L - 1)(C - 1) = (2 - 1)(3 - 1) = 2$$

Adotando nível de significância de 5%, obtemos, na Tabela 5 do apêndice, $\chi_c^2 = 5,99$ (ver Figura 10.6). Como $q^2 > \chi_c^2$, o teste rejeita a hipótese nula, evidenciando *associação entre a condição de peso e o tipo de leite*. Tal resultado deve-se, provavelmente, ao leite tipo UHT, para o qual foram observadas as maiores diferenças e, portanto, as maiores contribuições no cálculo da estatística Q^2 .⁵

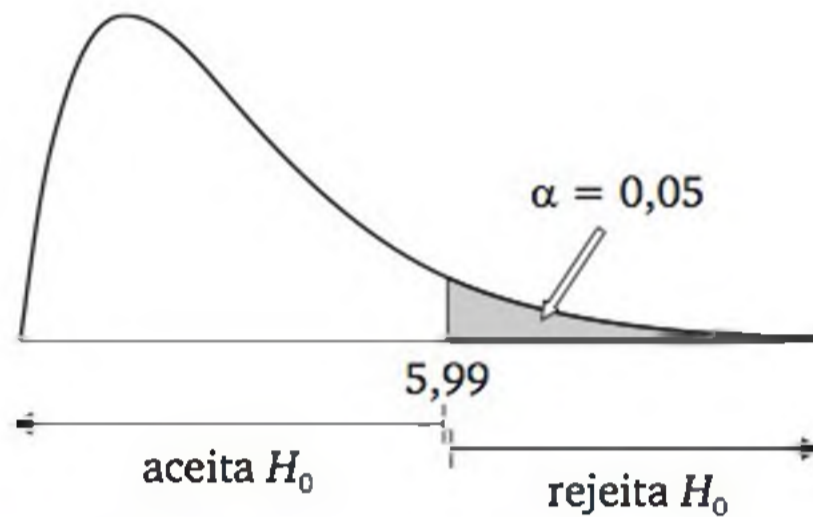


Figura 10.6 Regra de decisão de um teste de independência usando o modelo qui-quadrado com dois graus de liberdade.

O teste qui-quadrado de independência pode ser usado para o estudo de associação entre duas variáveis qualitativas nominais, mas também é possível usá-lo com variáveis ordinais ou quantitativas (discretas ou contínuas), desde que seus resultados sejam classificados em *categorias*. Contudo, tal como no teste qui-quadrado de aderência, há uma restrição importante: todos os totais da tabela de contingência devem ser razoavelmente grandes, de tal forma a garantir $E_{ij} \geq 5$, para todo $i = 1, 2, \dots, L$ e $j = 1, 2, \dots, C$.

Para tabelas 2×2 , isto é, quando as duas variáveis têm apenas 2 categorias ($L = 2$ e $C = 2$), é recomendável fazer uma pequena alteração na Expressão (10.5), para que a distribuição da estatística Q^2 fique mais próxima do modelo teórico (distribuição qui-quadrado com $gl = 1$). A alteração consiste em subtrair 0,5 unidade da diferença absoluta entre cada par O_{ij} e E_{ij} , ou seja:

$$Q^2 = \sum_{i=1}^L \sum_{j=1}^C \left[\frac{(|O_{ij} - E_{ij}| - 0,5)^2}{E_{ij}} \right] \quad (10.7)$$

A alteração proposta em (10.7) corresponde à correção de continuidade discutida na seção 6.3, pois a distribuição qui-quadrado com $gl = 1$ vem de uma variável aleatória com distribuição normal padrão ao quadrado.

⁵ Como o tamanho da amostra deste exemplo é bastante grande, é natural o teste rejeitar H_0 , porque as diferenças, mesmo que pequenas, têm pouca chance de ser explicadas pelo acaso.

EXERCÍCIOS

6. Há dúvidas sobre os desempenhos dos alunos, na disciplina de Estatística, de alguns cursos de Engenharia. Alguns argumentam que, dependendo do curso, o percentual de aprovação pode ser diferente, mesmo que a disciplina tenha o mesmo programa. Um estudo foi realizado, selecionando aleatoriamente alunos de três cursos, registrando os aprovados e reprovados na disciplina. Os resultados estão na tabela a seguir:

Situação	Curso			Total
	Eng. Civil	Eng. Química	Eng. Mecânica	
Aprovados	44	26	35	105
Reprovados	11	26	15	52
Total	55	52	50	157

Considerando nível de significância de 5%, os percentuais de aprovação podem ser considerados iguais?

7. Uma metalúrgica produz grandes quantidades de parafusos, trabalhando em três turnos. O setor da qualidade deseja verificar se o desempenho dos turnos é semelhante, o que poderia ser avaliado através das proporções de peças aprovadas, direcionadas a retrabalho ou rejeitadas. Como parte do Controle Estatístico de Processos, amostras aleatórias de parafusos são coletadas de cada turno. Uma dessas amostras, com a classificação das peças, está mostrada na tabela a seguir:

Situação das peças	Turno			Total
	Matutino	Vespertino	Noturno	
Aprovadas	432	456	424	1.312
Retrabalho	185	190	180	555
Rejeitadas	45	48	39	132
Total	662	694	643	1.999

É possível considerar semelhante o desempenho dos turnos? Use nível de significância de 1%.

8. Uma rede local de computadores tem cinco clientes que enviam pacotes de dados (gerados por um aplicativo) ao servidor. Os pacotes podem ser con-

siderados completos, incompletos mas aproveitáveis e inaproveitáveis, dependendo de como são recebidos pelo servidor. Suspeita-se que pode haver problemas em um ou mais dos clientes, o que poderia ser evidenciado por diferentes percentuais de pacotes completos, aproveitáveis e inaproveitáveis. Um estudo foi realizado, fazendo com que cada cliente enviasse certo número de pacotes ao servidor, donde os pacotes foram observados. Os resultados foram:

Cliente	Situação dos pacotes			Total
	Completos	Aproveitáveis	Inaproveitáveis	
1	485	10	5	500
2	768	24	8	800
3	624	20	6	650
4	522	40	18	580
5	650	3	17	670
Total	3.049	97	54	3.200

A suspeita tem fundamento? Use nível de significância de 2,5%.

10.3 TESTES PARA DUAS POPULAÇÕES

Muitas vezes, não é possível aplicar os testes paramétricos para comparar diferentes tratamentos ou populações (Capítulo 9), porque alguma das suposições não é satisfeita. Um caso típico ocorre nas situações em que os dados não podem ser considerados provenientes de uma população com distribuição normal.⁶ Já vimos na seção anterior o teste qui-quadrado de independência (ou de homogeneidade), que permite comparar duas ou mais populações quando a resposta é proveniente de uma variável nominal. Nesta seção, discutiremos testes para comparar duas populações com dados ordinais. Na prática, podemos aplicá-los em dados provenientes de variável quantitativa, usando apenas a ordenação dos valores.

10.3.1 Teste dos sinais

Trata-se de um teste para casos em que a variável observada tem nível de mensuração pelo menos ordinal. O teste dos sinais é utilizado para comparar a

⁶ Podemos aplicar o teste de Lilliefors para verificar essa suposição.

posição central ou *locação* de duas distribuições populacionais, com base em amostras *pareadas*. Exemplos:

- comparar as notas de alunos de Engenharia em uma prova padrão de Estatística, antes e após a realização de um curso de reforço. Observe que são os *mesmos* alunos, avaliados em momentos diferentes, que provocam dados pareados; e o interesse é avaliar se há diferença significativa entre o desempenho *mediano* dos dois grupos;
- verificar se há diferença no tempo gasto por computadores para executar uma mesma tarefa, antes e após a atualização de seus sistemas operacionais (supondo computadores com configurações idênticas). Novamente, são os mesmos elementos, avaliados em momentos diferentes, gerando dados pareados; e o teste possibilitará verificar se há diferença significativa entre as posições centrais das distribuições dos tempos de resposta das duas situações;
- um grupo de consumidores de determinado *software* recebe uma nova versão desse produto, e deve opinar se considera o desempenho da nova versão melhor ou pior que a antiga. Há interesse em verificar se a nova versão atendeu às expectativas dos clientes, ou seja, se em geral (ou na *mediana*) os consumidores consideram o desempenho da nova versão melhor do que o desempenho da antiga.

Nos dois primeiros exemplos, supondo que a variável resposta siga aproximadamente um modelo normal, pode ser aplicado o teste *t* para dados pareados (seção 9.2). Já no último exemplo, o teste *t* não pode ser usado, porque não são observadas *medidas quantitativas* da variável resposta (desempenho do *software*). Este é um caso típico de aplicação do teste dos sinais.

Sejam X_1 e X_2 as variáveis aleatórias que indicam as duas condições de medida da variável resposta, no experimento *pareado*. Em termos de nossos exemplos, X_1 é a medida *antes* e X_2 é a medida *depois*. Essas variáveis não precisam ser observadas diretamente, mas supõe-se que sejam *contínuas*. Seja $D = X_2 - X_1$ a diferença num par de medidas e η_D a *mediana* de D . As hipóteses do teste dos sinais podem ser escritas como:⁷

$$H_0: \eta_D = 0 \text{ e } H_1: \eta_D \neq 0$$

Na abordagem unilateral, a hipótese alternativa é $H_1': \eta_D > 0$ ou $H_1'': \eta_D < 0$.

Pela definição de mediana, temos $P(D \geq \eta_D) = P(D < \eta_D) = 0,5$. Então, se H_0 for verdadeira,

⁷ Se for suposto $D = X_2 - X_1$, com distribuição simétrica, então a mediana é igual à média, e as hipóteses tornam-se as mesmas do teste *t* *pareado*.

$$P(D \geq 0) = 0,5 \Leftrightarrow P(X_2 - X_1 \geq 0) = 0,5 \Leftrightarrow P(X_2 \geq X_1) = 0,5$$

Assim, definindo $p = P(X_2 \geq X_1)$, que, em termos de nossos exemplos, representa a probabilidade de a avaliação *depois* ser maior do que a avaliação *antes*, a hipótese nula pode ser escrita como:

$$H_0: p = 0,5$$

caracterizando o teste dos sinais como um caso particular do teste de uma proporção (seção 8.7).

Seja n_+ o número de pares em que $X_2 > X_1$ (*pares com sinal positivo*). Se H_0 for verdadeira, por (5.14) temos que o valor esperado de n_+ é $n/2$. Assim, se for observado n_+ longe de $n/2$, há indícios para rejeitar H_0 (considerando a hipótese alternativa e o nível de significância do teste).

Amostra pequena

Quando o número n de pares avaliados (tamanho da amostra) for pequeno, usamos a distribuição binomial (Tabela 1) para obter o valor \mathbf{p} e realizar o teste. A Tabela 1 contempla valores de n não superiores a 15.

Exemplo 10.6 Um sistema de alarme possui um grande número de componentes. Há interesse em saber se houve ou não *aumento* no tempo de falha dos componentes após implementação de um programa de manutenção. Usualmente, o tempo de falha segue aproximadamente uma distribuição exponencial. Foi observada uma amostra de dez componentes, antes e depois do programa de manutenção, e os resultados (em horas) estão na tabela a seguir:

Componente	Antes	Depois	Sinal ¹
1	400	395	-
2	360	350	-
3	450	556	+
4	390	480	+
5	430	405	-
6	386	500	+
7	452	547	+
8	470	462	-
9	400	500	+
10	340	480	+

¹ Foi usado sinal *positivo* quando a medida *depois* foi maior do que a medida *antes* (compatível com a hipótese em estudo).

Observe que se trata de uma amostra pareada e pequena ($n = 10$). Embora a variável resposta seja contínua e seja observada nas duas condições, as populações não seguem distribuições normais; assim, não é recomendável utilizar o teste t para dados pareados. Aplicaremos o teste dos sinais. As hipóteses são:

H_0 : a mediana do tempo de falha *depois* é igual à mediana do tempo de falha *antes*;

H_1 : a mediana do tempo de falha *depois* é maior do que a mediana do tempo de falha *antes*.

Da amostra de $n = 10$ casos eram esperados (sob H_0) 5 sinais positivos, mas observamos $n_+ = 6$. O valor de n_+ aponta na direção de H_1 , mas vamos obter, na Tabela 1 do apêndice, o valor \mathbf{p} para verificar se a diferença em relação ao valor esperado não pode ser explicada pelo acaso. Como o teste é unilateral à direita, temos:

$$\begin{aligned} \mathbf{p} &= p(6) + p(7) + p(8) + p(9) + p(10) = \\ &= 0,2051 + 0,1172 + 0,0439 + 0,0098 + 0,0010 = \\ &= 0,3770 \end{aligned}$$

Adotando nível de significância $\alpha = 0,05$, temos $\mathbf{p} > \alpha$, levando à aceitação de H_0 (ver Capítulo 8). Concluimos, então, que o teste dos sinais não encontrou evidências estatísticas de que o plano de manutenção programada tenha aumentado o tempo mediano de falhas dos componentes do sistema.

Amostra grande

Para valores de n superiores aos apresentados na Tabela 1, é possível usar a aproximação da normal à binomial, calculando:⁸

$$z = \frac{2n_+^* - n}{\sqrt{n}} \quad (10.8)$$

onde $n_+^* = n_+ - 0,5$ se $n_+ > \frac{n}{2}$ (teste bilateral ou unilateral à direita);

$n_+^* = n_+ + 0,5$ se $n_+ < \frac{n}{2}$ (teste bilateral ou unilateral à esquerda).⁹

8 Observe que (10.8) corresponde à Expressão 8.7, com $p_0 = 0,5$.

9 Note que, se o teste for unilateral e a relação entre n_+ e $\frac{n}{2}$ for diferente da apresentada, não podemos rejeitar H_0 em favor de H_1 .

Exemplo 10.7 Uma empresa quer observar a viabilidade de utilizar um novo tipo de calibrador, eletrônico, ao invés do modelo empregado atualmente, mecânico, para medir dimensões de peças automotivas. Após treinamento apropriado, 26 operários foram sorteados para realizar as medições das mesmas peças, com o calibrador eletrônico e o mecânico: os tempos gastos (em segundos) foram registrados. Somente será viável a introdução dos novos calibradores se o tempo mediano de medição for menor do que o obtido com os calibradores mecânicos ora em uso. Sabe-se que os tempos de medição desses calibradores não costumam seguir distribuições normais, donde se optou por evitar o tradicional teste t de médias. Os tempos de medição e os sinais (positivos quando a medição no calibrador eletrônico foi maior) estão na tabela a seguir:

Operário	Eletr.	Mec.	Sinal	Operário	Eletr.	Mec.	Sinal
1	27,0	27,0	0	14	22,0	29,0	-
2	25,0	30,1	-	15	16,0	16,0	0
3	22,0	28,0	-	16	22,0	20,6	+
4	34,0	34,0	0	17	29,5	25,0	+
5	23,0	24,5	-	18	36,0	33,0	+
6	22,0	28,0	-	19	35,0	41,0	-
7	25,0	28,0	-	20	35,0	35,0	0
8	32,3	30,0	+	21	27,5	28,0	-
9	34,0	36,0	-	22	29,0	31,8	-
10	23,5	29,0	-	23	27,0	28,0	-
11	34,2	39,0	-	24	24,3	29,3	-
12	31,8	30,0	+	25	30,1	30,8	-
13	28,4	20,0	+	26	29,3	35,6	-

As hipóteses são:

H_0 : os tempos medianos de medição com o calibrador eletrônico e com o calibrador mecânico são iguais;

H_1 : o tempo mediano de medição com calibrador eletrônico é menor que o tempo mediano de medição com calibrador mecânico.

Faremos esse teste, unilateral à esquerda, com nível de significância $\alpha = 0,05$. Embora tenham sido realizadas 26 observações pareadas, vamos excluir as quatro observações em que o nível de precisão das medidas não detectou diferença entre os dois calibradores. Assim, temos $n = 22$.

Pela abordagem clássica, é preciso encontrar o valor z_c na distribuição normal padrão (Tabela 3 do Apêndice), que delimita área de $\alpha = 0,05$ na cauda direita da distribuição. Como o teste é unilateral à esquerda, o teste rejeita H_0 se e somente se o valor calculado z (Expressão 10.8) for menor ou igual a $-z_c$ (Figura 10.7).

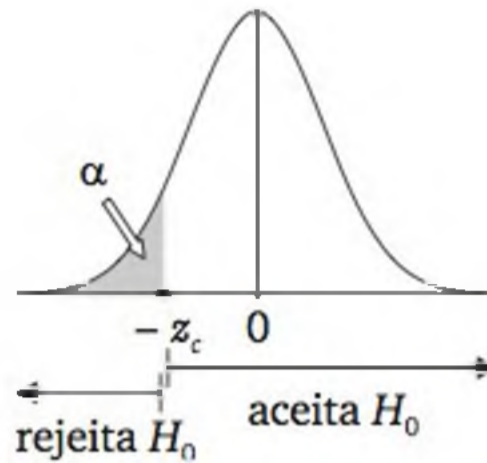


Figura 10.7 Regra de decisão para um teste unilateral à esquerda, utilizando a distribuição normal padrão.

Pela tabela da distribuição normal padrão, obtemos $z_c = 1,645$, levando à regra de decisão:

se $z \leq -1,645$, então rejeita H_0 em favor de H_1 ;

se $z > -1,645$, então aceita H_0 .

As $n = 22$ observações pareadas válidas acusaram $n_+ = 6$. Aplicando (10.8), encontramos:

$$z = \frac{2n_+ - n}{\sqrt{n}} = \frac{2(6,5) - 22}{\sqrt{22}} = -1,919$$

levando à rejeição de H_0 em favor de H_1 .

O teste dos sinais encontrou evidências estatísticas de que o tempo médio de medição efetuada pelo calibrador eletrônico foi menor do que pelo calibrador mecânico. Assim, os calibradores eletrônicos podem ser introduzidos na organização.

10.3.2 Teste dos sinais por postos

Em cada par de observações, o teste dos sinais, discutido na seção 10.3.1, avalia apenas qual tratamento foi melhor, desconsiderando a magnitude da diferença. Em consequência, o teste dos sinais não costuma detectar diferença entre os dois tratamentos se as amostras não forem grandes, a menos que a diferença real entre os tratamentos seja muito grande.

O teste dos sinais por postos, também chamado de teste de Wilcoxon, é uma boa alternativa para os casos em que é possível avaliar a magnitude das diferenças (pelo menos a nível ordinal), mas não é possível garantir as suposições para a aplicação do teste t pareado. A ordenação da magnitude das diferenças pode

ser feita por especialistas ou, quando a variável observada for quantitativa, pode ser feita objetivamente, conforme descreveremos.

Atribuimos postos às diferenças de cada par, independentemente do sinal, alocando o posto 1 à menor diferença em módulo; o posto 2 à segunda menor diferença em módulo; ...; o posto n à maior diferença em módulo. Às observações empatadas atribuímos a média dos postos correspondentes (por exemplo, considere três diferenças iguais, que, se distintas, corresponderiam aos postos 4, 5 e 6; atribuímos a cada uma delas a média desses postos, ou seja, o posto 5).

É suposto que a variável resposta seja *contínua*. Embora ela possa não ser observável diretamente, supõe-se que seja possível atribuir postos às diferenças de cada par. Adota-se a mesma formulação das hipóteses (H_0 e H_1) do teste dos sinais (em termos da mediana das diferenças da variável resposta), mas o teste dos sinais por postos é desenvolvido através da ordenação das diferenças. Sob a hipótese nula, a soma dos postos das diferenças positivas deve ser aproximadamente igual à soma dos postos das diferenças negativas.

Seja a variável aleatória S_+ definida como a *soma dos postos das diferenças positivas* de uma amostra de n pares a ser observada. Sob H_0 , é possível mostrar que seu valor esperado e sua variância são dados, respectivamente, por:

$$E(S_+) = \frac{n(n+1)}{4} \quad (10.9)$$

$$V(S_+) = \frac{n(n+1)(2n+1)}{24} \quad (10.10)$$

Observada uma amostra pareada de n observações (já descontando algum caso em que a diferença foi nula), definimos s_+ como a *soma dos postos das diferenças positivas* da amostra observada. Se s_+ estiver longe do valor esperado por H_0 (Expressão 10.9), há indícios para rejeitar H_0 (considerando a hipótese alternativa e o nível de significância do teste).

Amostra pequena

Para $n \leq 20$ (descontando algum caso em que a diferença no par foi nula), a Tabela 8 fornece o valor crítico s_c para s_+ , permitindo construir a regra de decisão, em função da hipótese alternativa e do nível de significância do teste. Veja Quadro 10.1.

Quadro 10.1 Regra de decisão do teste dos sinais por postos para amostras pequenas, usando a Tabela 8 do Apêndice.

Hipótese H_1	Regra de decisão
Unilateral à direita	<ul style="list-style-type: none"> – Encontrar s_c tal que $P(S_+ \leq s_c) \approx 1 - \alpha$ – Se $s_+ \geq s_c$, então rejeita H_0
Unilateral à esquerda	<ul style="list-style-type: none"> – Encontrar s_c tal que $P(S_+ \leq s_c) \approx \alpha$ – Se $s_+ \leq s_c$, então rejeita H_0
Bilateral	<ul style="list-style-type: none"> – Encontrar s_{c1} tal que $P(S_+ \leq s_{c1}) \approx \alpha/2$ – Encontrar s_{c2} tal que $P(S_+ \leq s_{c2}) \approx 1 - \alpha/2$ – Se $s_+ \leq s_{c1}$ ou $s_+ \geq s_{c2}$, então rejeita H_0

Nota. As probabilidades não são exatas, porque S_+ é uma variável aleatória discreta, mas os valores críticos da Tabela 8 foram calculados de forma que $P(S_+ \leq s_c)$ não supera o nível de significância estabelecido.

Exemplo 10.8 Vamos retomar o Exemplo 10.6. Como há possibilidade de avaliar tanto a direção quanto a magnitude das diferenças, vamos usar o teste dos sinais por postos, com $\alpha = 0,05$. A tabela seguinte apresenta as diferenças e a atribuição dos postos.

Componente	Antes X_1	Depois X_2	Diferença $D = X_1 - X_2$	Postos
1	400	395	- 5	1
2	360	350	- 10	3
3	450	556	106	8
4	390	480	90	5
5	430	405	- 25	4
6	386	500	114	9
7	452	547	95	6
8	470	462	- 8	2
9	400	500	100	7
10	340	480	140	10

Somando os postos referentes às diferenças positivas, obtemos $s_+ = 45$. Como o teste é unilateral à direita, o valor crítico s_c é obtido na Tabela 8, tal que $P(S_+ \geq s_c) = 0,05$ ou, equivalentemente, $P(S_+ < s_c) = 0,95$. A ilustração, a seguir, mostra como obter o valor crítico s_c através da Tabela 8.

n	$P(S_+ \leq s_c)$ aproximadamente igual a				
	...	0,90	0,95	0,975	...
...
10	...	40	44	46	...
...

Verificamos que $s_c = 44$. Como $s_+ = 45 > s_c$, o teste rejeita H_0 ao nível de significância de 5%, mostrando evidência estatística de que o plano de manutenção programada aumentou o tempo mediano de falha dos componentes do sistema.

Note que a decisão desse teste foi diferente do teste dos sinais (Exemplo 10.6). Isso se deve ao teste dos sinais por postos (Teste de Wilcoxon) usar melhor a informação dos dados, computando, em cada par de observações, não só a direção da diferença, mas também o posto relativo às demais diferenças. Assim, esse teste é mais *poderoso* do que o anterior, isto é, tem maior chance de detectar a falsidade de H_0 , quando ela é realmente falsa.

Devemos dar preferência ao Teste de Wilcoxon, em relação ao teste dos sinais, sempre que for possível ordenar as diferenças da amostra pareada.

Amostra grande

A Tabela 8 fornece os valores críticos para $n \leq 20$. Para $n > 20$, a distribuição de S_+ é aproximadamente igual a uma distribuição normal, com média e variância especificadas em (10.9) e (10.10). Assim, podemos calcular:

$$z = \frac{4s_+ - n(n+1)}{\sqrt{n(n+1)(2n+1)}} \cdot \frac{\sqrt{24}}{4} \quad (10.11)$$

e usar a distribuição normal padrão para obter o valor crítico z_c , considerando o nível de significância α e o tipo de hipótese alternativa.

Exemplo 10.9 Vamos retomar ao Exemplo 10.7, usando o teste dos sinais por postos, com $\alpha = 0,05$. As hipóteses e a regra de decisão são as mesmas apresentadas anteriormente. A tabela seguinte apresenta as diferenças entre os calibradores:

Operário	Eletr.	Mec.	Dif.	Operário	Eletr.	Mec.	Dif.
1	27,0	27,0	0,0	14	22,0	29,0	- 7,0
2	25,0	30,1	- 5,1	15	16,0	16,0	0,0
3	22,0	28,0	- 6,0	16	22,0	20,6	1,4
4	34,0	34,0	0,0	17	29,5	25,0	4,5
5	23,0	24,5	- 1,5	18	36,0	33,0	3,0
6	22,0	28,0	- 6,0	19	35,0	41,0	- 6,0
7	25,0	28,0	- 3,0	20	35,0	35,0	0,0
8	32,3	30,0	2,3	21	27,5	28,0	- 0,5
9	34,0	36,0	- 2,0	22	29,0	31,8	- 2,8
10	23,5	29,0	- 5,5	23	27,0	28,0	- 1,0
11	34,2	39,0	- 4,8	24	24,3	29,3	- 5,0
12	31,8	30,0	1,8	25	30,1	30,8	- 0,7
13	28,4	20,0	8,4	26	29,3	35,6	- 6,3

Como temos $n = 22$ casos com diferenças não nulas, usaremos a distribuição aproximada normal. Excluímos os casos com diferença nula e fizemos a ordenação dos casos em termos de diferenças absolutas. A alocação dos postos foi feita em termos de diferenças absolutas, conforme mostra a tabela seguinte. Quando houve empate (como nos operários 7 e 18), colocamos, inicialmente, um posto para cada caso [coluna posto (a)] e, depois, calculamos a média [coluna posto (b)]. Nesta última coluna, colocamos o sinal da diferença entre parêntesis.

Op.	Dif.	Dif. abs.	Posto (a)	Posto (b)	Op.	Dif.	Dif. abs.	Posto (a)	Posto (b)
21	- 0,5	0,5	1	1 (-)	17	4,5	4,5	12	12 (+)
25	- 0,7	0,7	2	2 (-)	11	- 4,8	4,8	13	13 (-)
23	- 1,0	1,0	3	3 (-)	24	- 5,0	5,0	14	14 (-)
16	1,4	1,4	4	4 (+)	2	- 5,1	5,1	15	15 (-)
5	- 1,5	1,5	5	5 (-)	10	- 5,5	5,5	16	16 (-)
12	1,8	1,8	6	6 (+)	3	- 6,0	6,0	17*	18,0 (-)
9	- 2,0	2,0	7	7 (-)	6	- 6,0	6,0	18*	18,0 (-)
8	2,3	2,3	8	8 (+)	19	- 6,0	6,0	19*	18,0 (-)
22	- 2,8	2,8	9	9 (-)	26	- 6,3	6,3	20	20 (-)
7	- 3,0	3,0	10*	10,5 (-)	14	- 7,0	7,0	21	21 (-)
18	3,0	3,0	11*	10,5 (+)	13	8,4	8,4	22	22 (+)

* Procedimento para postos empatados.

Calculando os postos de diferenças positivas, temos: $s_+ = 62,5$. E o valor da estatística de teste para a amostra observada:

$$z = \frac{4s_+ - n(n+1)}{\sqrt{n(n+1)(2n+1)}} \cdot \frac{\sqrt{24}}{4} = \frac{4(62,5) - 22(23)}{\sqrt{22(23)(45)}} \cdot \frac{\sqrt{24}}{4} = -2,078$$

Considerando a regra de decisão baseada em $z_c = 1,645$ (valor da distribuição normal para $\alpha = 0,05$ e teste unilateral), o teste rejeita H_0 , levando à mesma decisão do teste dos sinais.

10.3.3 Teste de Mann-Whitney

O teste de Mann-Whitney ou teste de Wilcoxon-Mann-Whitney é usado para comparar a posição central de duas populações, com base em amostras independentes, extraídas aleatoriamente dessas populações. Veja os exemplos a seguir:

- um professor de Estatística quer comparar as médias obtidas em uma prova padrão do assunto por alunos de dois cursos diferentes;
- com o propósito de avaliar a qualidade de duas máquinas, são comparados os diâmetros das peças produzidas por elas;
- um administrador de rede quer saber se há diferenças significativas nos tempos de processamento entre computadores tipo PC e Macintosh (com configurações equivalentes).

Supostamente, os dados das duas amostras independentes são gerados por populações com distribuições contínuas, embora as variáveis não precisem ser observadas diretamente; pode ser observado apenas uma ordenação dos elementos. Sejam η_1 a mediana da população 1 e η_2 a mediana da população 2. As hipóteses podem ser colocadas como:

$$H_0: \eta_1 = \eta_2 \text{ e } H_1: \eta_1 \neq \eta_2,$$

sendo que na abordagem unilateral a hipótese alternativa é $H_1': \eta_1 > \eta_2$ ou $H_1'': \eta_1 < \eta_2$, dependendo do que se quer provar.

Sejam n_1 e n_2 os tamanhos das amostras 1 e 2. Os $n_1 + n_2$ elementos devem ser ordenados em ordem crescente, em termos da variável observada. Atribuímos posto 1 à menor observação, posto 2 à segunda menor observação e assim por diante, até o posto $n_1 + n_2$ à maior observação. Quando houver empates (valores iguais), adotaremos o mesmo procedimento do teste de sinais por postos.

Chamamos de W_1 a soma dos postos da amostra 1 e W_2 a soma dos postos da amostra 2. É possível mostrar a seguinte relação:

$$W_1 + W_2 = \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2} \quad (10.12)$$

que pode ser usada para verificar os cálculos de W_1 e W_2 .

A estatística do teste é basicamente a soma dos postos da amostra 1. Mas faremos uma pequena alteração, para facilitar a análise quando as amostras têm tamanhos diferentes. Consideraremos a estatística:

$$U = W_1 - \left[\frac{n_1(n_1 + 1)}{2} \right] \quad (10.13)$$

Sob H_0 , a estatística U tem as seguintes características:

$$E(U) = \frac{n_1 n_2}{2} \quad (10.14)$$

$$V(U) = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} \quad (10.15)$$

A hipótese nula do teste é de que não há diferença entre as posições centrais das duas populações. Sob essa hipótese, as somas dos postos referentes às duas amostras devem apresentar resultados aproximadamente iguais, com as devidas correções se as duas amostras tiverem tamanhos diferentes. Assim, se H_0 for verdadeira, observadas amostras independentes das duas populações, o valor calculado de U , que chamaremos de u , não deve estar distante da grandeza descrita em (10.14). Se for verificado u distante do valor esperado por H_0 , há indícios para rejeitar H_0 (considerando a hipótese alternativa e o nível de significância do teste).

Amostras pequenas

Para $n_1, n_2 \leq 20$, a Tabela 9 fornece valores críticos para u , permitindo construir a regra de decisão, em função da hipótese alternativa e do nível de significância α do teste. Veja o Quadro 10.2.

Quadro 10.2 Regra de decisão do teste de Mann-Whitney para amostras pequenas, usando a Tabela 9 do Apêndice A.

Hipótese alternativa	Regra de decisão
H_1' : $\eta_1 > \eta_2$	<ul style="list-style-type: none"> - Encontrar u_c tal que $P(U \leq u_c) \approx 1 - \alpha$ - Se $u \geq u_c$, então rejeita H_0
H_1'' : $\eta_1 < \eta_2$	<ul style="list-style-type: none"> - Encontrar u_c tal que $P(U \leq u_c) \approx \alpha$ - Se $u \leq u_c$, então rejeita H_0
H_1 : $\eta_1 \neq \eta_2$	<ul style="list-style-type: none"> - Encontrar u_{c1} tal que $P(U \leq u_{c1}) \approx \alpha/2$ - Encontrar u_{c2} tal que $P(U \leq u_{c2}) \approx 1 - \alpha/2$ - Se $u \leq u_{c1}$ ou $u \geq u_{c2}$, então rejeita H_0

Nota: As probabilidades não são exatas, porque U é uma variável aleatória discreta, mas os valores críticos da Tabela 9 foram calculados de forma que $P(U \leq u_c)$ não supera o nível de significância α estabelecido.

Exemplo 10.10 Um fabricante de vergalhões de ferro para estruturas afirma que seu novo produto apresenta resistência à tração superior ao modelo atualmente vendido, o que justificaria um preço maior. Um cliente não muito convencido quer realizar um teste estatístico para avaliar a afirmação do fabricante. Ele analisou 15 vergalhões de cada tipo. Os vergalhões foram submetidos à tração, em kgf, até o rompimento. Estudos anteriores sugerem que essa variável não segue uma distribuição normal, o que sugere o uso de teste não paramétrico. Os resultados são apresentados na tabela a seguir:

Vergalhão novo (1)			Vergalhão atual (2)		
276	380	237	119	696	240
231	127	143	461	298	246
144	234	260	473	327	566
151	165	237	380	293	232
195	198	174	287	199	108

As hipóteses são:

H_0 : a resistência mediana do vergalhão novo é igual à resistência mediana do vergalhão atual;

H_1 : a resistência mediana do vergalhão novo é maior que a resistência mediana do vergalhão atual.

Trata-se de um teste unilateral à direita, porque se espera que o novo vergalhão tenha *maior* resistência à tração que o atual. Adotaremos nível de significância de 5%. Primeiramente, ordenamos os valores, sem levar em conta de qual grupo cada um deles veio. Depois, atribuímos os postos, como é mostrado a seguir:

Grupo ¹	Resist.	Posto	Grupo ¹	Resist.	Posto
2	108	1	1	237	15,5
2	119	2	2	240	17
1	127	3	2	246	18
1	143	4	1	260	19
1	144	5	1	276	20
1	151	6	2	287	21
1	165	7	2	293	22
1	174	8	2	298	23
1	195	9	2	327	24
1	198	10	1	380	25,5
2	199	11	2	380	25,5
1	231	12	2	461	27
2	232	13	2	473	28
1	234	14	2	566	29
1	237	15,5	2	696	30

¹ Códigos: 1 = novo; 2 = atual.

Chamando a amostra referente ao novo tipo de vergalhão de amostra 1, calculamos a soma de postos: $w_1 = 173,5$. Logo:

$$u = 173,5 - \left[\frac{15(15 + 1)}{2} \right] = 53,5$$

Com base na hipótese alternativa (H_1' : $\eta_1 > \eta_2$) e no nível de significância $\alpha = 0,05$, podemos construir a regra de decisão, de acordo com o Quadro 10.2. É preciso encontrar o valor crítico u_c para o qual $P(U \leq u_c) \approx 1 - 0,05 = 0,95$. Procurando na Tabela 9, conforme esquema a seguir, encontramos $u_c = 152$.

$n_1 = 15$	$P(U \leq u_c)$ aproximadamente igual a				
n_2	...	0,90	0,95	0,975	...
...					
15	...	144	152	160	...
...					

Como a amostra produziu o valor $u = 53,5$, que é menor que o valor crítico, o teste aceita H_0 ao nível de significância de 5%. Então, o teste de Mann-Whitney *não encontrou* evidência estatística de que o novo vergalhão apresenta resistência à tração superior à do atual.

Amostras grandes

Para $n_1 > 20$ ou $n_2 > 20$, a distribuição de U aproxima-se de uma normal, com média e variância especificadas em (10.14) e (10.15). Assim, podemos calcular:

$$z = \frac{u - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} \quad (10.16)$$

e usar a distribuição normal padrão (Tabela 3) para obter o valor crítico z_c , considerando o nível de significância α e o tipo de hipótese alternativa.

Exemplo 10.11 Um administrador de rede tem recebido insistentes reclamações de usuários de que os tempos de processamento dos dois servidores da rede são diferentes, no que tange ao acesso às correspondências eletrônicas. Intrigado, porque se supunha que não haveria razão para diferenças (suas configurações são praticamente iguais), ele coletou dados dos dois servidores, registrando os tempos de acesso (em segundos) de 30 usuários em cada servidor, cujos resultados estão na tabela a seguir. Sabe-se que os tempos de acesso não seguem uma distribuição normal, e não há hipótese *a priori* sobre qual servidor é mais rápido.

Servidor 1			Servidor 2		
5,83	3,78	6,79	2,27	6,24	2,44
0,99	1,40	2,70	7,41	4,73	4,17
6,07	5,88	3,05	3,21	9,34	5,01
6,53	3,52	2,44	7,76	4,33	16,68
0,04	3,42	3,74	2,24	4,63	2,97
4,96	0,99	2,66	1,93	3,97	13,45
6,86	1,72	3,09	6,07	4,61	5,35
2,55	4,05	2,03	3,80	5,02	1,80
2,63	1,70	4,65	2,93	6,40	2,97
1,97	6,48	4,26	9,04	4,51	10,75

As hipóteses são:

$$H_0: \eta_1 = \eta_2 \text{ e } H_1: \eta_1 \neq \eta_2$$

onde η_1 é a mediana do tempo de acesso do servidor 1 e η_2 é a mediana do tempo de acesso do servidor 2. Faremos o teste ao nível de significância de 5%. Assim, a regra de decisão, para uma estatística com distribuição normal padrão, é apresentada na Figura 10.8.

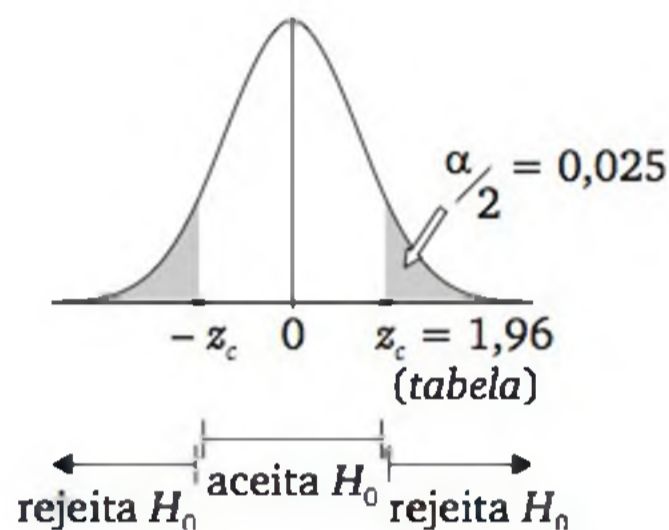


Figura 10.8 Regra de decisão para um teste bilateral, $\alpha = 0,05$, utilizando a distribuição normal padrão.

Calculando a soma dos postos da amostra 1, obtemos $w_1 = 754$. Resultando no seguinte valor para a estatística U :

$$u = w_1 - \left[\frac{n_1(n_1 + 1)}{2} \right] = 754 - \frac{30(30 + 1)}{2} = 289$$

Aplicando (10.15):

$$z = \frac{u - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} = \frac{289 - \frac{(30)(30)}{2}}{\sqrt{\frac{(30)(30)(30 + 30 + 1)}{12}}} = -2,38$$

Comparando o valor calculado de z com a regra de decisão da Figura 10.8, verificamos que o teste rejeita H_0 ao nível de significância de 5%. O teste de Mann-Whitney encontrou evidência estatística de que os tempos medianos de acesso aos dois servidores são diferentes.

EXERCÍCIOS

9. Um banco pretende trocar os terminais que seus caixas usam para atender ao público, com o intuito de conseguir diminuir o tempo de atendimento. Para avaliar se vale a pena a mudança, foi realizado um estudo em que os mesmos caixas atendiam a clientes com o terminal atual e com o novo (após treinamento), e os tempos médios de atendimento (em minutos) de cada caixa foram:

Caixa	Terminal atual	Terminal novo
1	1,6	0,5
2	3,5	0,3
3	0,8	5,9
4	1,1	1,6
5	7,1	1,7
6	8,4	2,2
7	4,8	1,0
8	5,9	1,4
9	2,4	1,0
10	5,5	2,1

- a) O terminal novo reduz o tempo de atendimento? Use o teste dos sinais com $\alpha = 0,05$.
- b) Repita o item (a), utilizando o teste dos sinais por postos.
- c) O que você conclui com base nos itens (a) e (b)? Por quê?
10. O mesmo banco do Exercício 9 acredita que o terminal novo pode vir a reduzir o número de pessoas nas filas. No mesmo estudo descrito anteriormente, o número de pessoas na fila foi monitorado:

Caixa	Terminal atual	Terminal novo
1	5	1
2	3	3
3	2	6
4	6	6
5	3	2
6	6	6
7	2	2
8	5	2
9	3	2
10	4	2

- a)** O terminal novo reduz o tempo de atendimento? Use o teste dos sinais com $\alpha = 0,05$.
- b)** Repita o item (a) utilizando o teste dos sinais por postos.

- 11.** Certa rede local de computadores tem como principal problema a baixa taxa de transmissão de dados (em kbps). Estuda-se a possibilidade de empregar um novo cabeamento, que supostamente aumentaria a taxa de transmissão de dados. Devido aos custos envolvidos, o gerente da rede decidiu realizar um experimento: os mesmos conjuntos de instruções foram executados pelo mesmo cliente, com o cabeamento atual e o proposto, e suas taxas de transmissão de dados medidas. Os resultados estão na tabela a seguir:

Instruções	Atual	Proposto	Instruções	Atual	Proposto
1	15,3	101,7	17	1,0	73,2
2	93,7	47,3	18	54,4	4,0
3	307,6	1.570,6	19	118,3	84,2
4	322,8	342,8	20	1,5	473,5
5	14,8	9,7	21	0,9	474,6
6	15,0	34,0	22	20,5	582,7
7	135,8	31,6	23	119,2	54,0
8	21,3	1.050,9	24	3,9	895,4
9	3,1	282,3	25	10,4	32,2
10	15,2	26,8	26	5,8	3.514,7
11	2,2	2,6	27	304,7	1.522,5
12	39,6	1.406,4	28	182,2	67,4
13	1,6	136,0	29	12,0	176,2
14	58,0	2.264,0	30	16,5	80,3
15	1,6	25,0	31	9,5	9,2
16	27,2	1,0	32	5,2	1,6

- a)** Pelo teste dos sinais, o cabeamento proposto aumenta a taxa de transmissão dos dados? Use nível de significância de 1%.
- b)** Repita o item (a), utilizando o teste dos sinais por postos.
- 12.** Em um processo de moldagem de peças plásticas, é preciso escolher quais dos dois métodos atualmente utilizados é o que obtém menor desperdício, que é mensurado pelo peso, em gramas, do polímero que “escapa” do mol-

de. Nem sequer se sabe se os dois métodos (chamados de 1 e 2) são realmente diferentes. Para tentar resolver o problema, o engenheiro responsável selecionou 25 operadores (treinados em ambos os métodos) e mediu os valores desperdiçados em cada método. Os valores, em gramas, foram:

Operador	Método 1	Método 2	Operador	Método 1	Método 2
1	0	0,642	14	0,001	0
2	0,247	0,106	15	0,182	19,657
3	0,678	0	16	0,073	0
4	1,577	0,142	17	0,314	8,651
5	0,002	0,001	18	0,001	0
6	0,041	1,025	19	5,890	1,324
7	0,593	0	20	0,016	0
8	0	0,322	21	0,190	8,346
9	4,078	0	22	0,366	0,007
10	0,049	3,635	23	0,010	0
11	1,010	24,113	24	0,002	7,858
12	0,066	0,338	25	2,566	0
13	0	19,746			

- a) Pelo teste dos sinais, usando um nível de significância de 5%, o que você conclui?
- b) Repita o item (a) pelo teste dos sinais por postos.
- c) Qual é sua conclusão? Por quê?
- 13.** Como parte de seu trabalho de conclusão de curso, um aluno de computação resolveu verificar se dois tipos de arquitetura (chamadas de A e B) diferem quanto à velocidade de processamento, num dado sistema computacional. Sob a arquitetura A, o aluno executou uma série de instruções e mediu os tempos de processamento, em milissegundos. Fez o mesmo sob a arquitetura B, com uma série de instruções análoga à que foi usada na arquitetura A. Os resultados, em milissegundos, foram:

Rede A		Rede B	
4,56	0,19	0,68	43,88
0,72	0,25	1,90	6,40
5,89	0,68	1,31	1,89
0,12	1,92	0,21	3,44
2,54	12,77	1,78	1,15
0,89	3,36	0,43	11,11

Para o nível de significância de 5%, as duas arquiteturas diferem quanto ao tempo de processamento?

- 14.** Dois tipos de concreto (X e Y) estão tendo suas resistências à compressão, em kgf, avaliadas por ensaios. Suspeita-se que o concreto Y seja mais resistente, o que poderia justificar seu maior preço. Sabe-se que as distribuições das trações, para ambos os tipos, não podem ser aproximadas por uma normal. Com base nos resultados apresentados a seguir, e considerando um nível de significância de 1%, a suspeita é confirmada?

Concreto X :

243,37 244,03 249,8 209,84 244,79
233,11 241,34 214,33 200,19 214,55

Concreto Y :

300,53 355,98 305,12 280,34 502,06 250,75 224,65

- 15.** O tempo de rompimento dos elos fusíveis é um fator crucial para a proteção dos sistemas de distribuição de energia elétrica: uma vez ocorrido um curto circuito, o fusível deve romper-se no menor tempo possível. A concessionária de energia elétrica de um estado brasileiro está estudando a adoção de um novo modelo, que o fabricante declara ter menor tempo de rompimento do que o modelo atual. Como parte do processo de decisão, foram realizados ensaios com os dois modelos. Eles foram submetidos a correntes de curto circuito, e seus tempos até o rompimento foram monitorados. Com base nos resultados (descritos a seguir), e usando um nível de significância de 5%, deve-se cogitar a adoção do novo modelo de elo fusível?

Tempos de rompimento (em segundos) – Atual elo fusível:

0,89 3,32 1,20 3,07 1,57 0,15 0,88 0,29 1,59 3,74 4,57
1,17 0,62 6,82 0,28 7,45 1,05 8,48 0,48 0,35 0,60 0,97

Tempo de rompimento (em segundos) – Novo elo fusível:

0,68 3,44 1,69 0,69 0,61 1,31 1,82 0,37 2,72 0,02 1,03
0,03 0,02 1,22 4,88 0,17 1,6 1,91 0,35 4,3 0,82 0,24
1,06 0,17 0,74

- 16.** No processo de produção de papel, a degradação da lignina (uma enzima) é um aspecto fundamental e precisa ser feito rapidamente, exigindo, usualmente, a utilização de cloro, o que pode ser danoso ao meio ambiente. Recentemente, foram realizadas pesquisas que procuram avaliar a viabilidade de degradação da lignina através da ação de fungos, em bio-reatores, redu-

zindo os danos ao meio ambiente e os gastos com tratamento de efluentes das fábricas de papel. Os departamentos de engenharia química e botânica de uma universidade brasileira resolveram realizar um experimento para avaliar duas espécies de fungo, medindo o tempo gasto para degradar um pequeno cubo de madeira de eucalipto. Acredita-se que a espécie 1 consiga degradar a lignina em menos tempo. Os resultados obtidos (em dias) estão descritos a seguir:

Espécie 1

6,5 11,0 16,0 13,5 13,0 16,5 28,5 6,0 7,0 10,0
 6,0 7,5 17,5 10,5 14,5 15,0 16,0 4,0 10,5 27,5
 5,5 8,5 37,0 25,0 19,0

Espécie 2

51,5 22,5 17,5 16,0 46,5 32,0 5,5 14,0 15,5 38,5
 36,5 46,0 17,0 13,0 19,0 34,5 20,0 59,5 14,5 20,5
 12,0 66,0 29,5 59,0 19,0

A espécie 1 realmente é mais rápida? Use nível de significância de 5%.

EXERCÍCIOS COMPLEMENTARES

- 17.** Uma concessionária de energia elétrica pretende calcular intervalos de confiança para o consumo mensal de kWh de pequenos consumidores. Medições anteriores indicam que o consumo segue uma distribuição normal com média de 85 kWh e desvio padrão de 15 kWh. Uma amostra aleatória de 25 consumidores foi selecionada, registrando seu consumo mensal:

82,01	67,62	80,48	77,35	109,02
92,51	70,53	86,35	60,65	82,36
132,11	66,18	88,56	69,53	101,34
99,45	70,00	80,50	74,87	68,96
101,08	87,87	100,2	92,74	97,55

- a)** Use o teste qui-quadrado de aderência para verificar se a suposição baseada nas medidas anteriores ainda é válida, em nível de significância de 5%. (Sugestão: defina as categorias utilizando os quartis da distribuição normal suposta.)
- b)** Refaça o problema, usando o teste de aderência de Kolmogorov-Smirnov.

- 18.** Certa empresa tem 3 unidades fabris. Suspeita-se que a existência de defeitos nos produtos estaria associada à unidade fabril que os produziu. Para testar isso, coletou-se uma amostra, obtendo-se:

Fábrica 1: 110 produtos sem defeito e 40 produtos com defeito.

Fábrica 2: 50 produtos sem defeito e 20 produtos com defeito.

Fábrica 3: 85 produtos sem defeito e 15 produtos com defeito.

Teste a suspeita da empresa ao nível de significância de 5%.

- 19.** Um instituto de metrologia está fazendo uma avaliação da qualidade de certo produto de limpeza, que é produzido por 3 filiais diferentes da mesma empresa. O produto precisa ter uma concentração adequada de desinfetante. Os testes identificaram 3 tipos de concentração: Adequada, Regular e Inadequada. Os resultados estão na tabela a seguir:

	Filial 1	Filial 2	Filial 3
Adequada	350	100	80
Regular	90	30	40
Inadequada	100	10	150

A concentração de desinfetante no produto é homogênea nas 3 filiais? Use nível de significância de 1%.

- 20.** O gerente de produção de um conglomerado industrial quer avaliar se as mudanças no programa da Qualidade foram aprovadas pelos funcionários. Uma amostra aleatória de 400 pessoas foi selecionada, e sua opinião sobre o programa foi registrada, antes e após a introdução das mudanças. Os resultados foram:

Antes das mudanças	Após as mudanças	
	Aprovam	Não aprovam
Aprovavam	170	30
Não aprovavam	190	10

Usando um teste adequado, verifique se as mudanças do programa da Qualidade foram aprovadas pelos funcionários. Use $\alpha = 0,05$.

- 21.** Estamos interessados em verificar se há diferenças entre o desempenho escolar de homens e o de mulheres no curso de engenharia de produção. Aplicamos a mesma prova a um grupo de 15 alunas e a um grupo de 13 alunos, que obtiveram os resultados a seguir:

Alunas 4,5 8,9 6,8 7,8 5,4 10,0 5,7 8,0
5,0 6,0 7,6 7,5 4,0 6,8 9,0

Alunos 3,5 6,5 8,0 2,0 9,5 5,6 4,5
7,0 8,8 7,5 8,0 1,5 6,5

Testes de aderência concluíram que não é possível usar um teste paramétrico. Usando nível de significância de 1%, qual é a sua conclusão?

22. Uma olaria de SC estava preocupada com a qualidade dos seus produtos (pois estava perdendo muitos clientes para a concorrência). Resolveu, então, implantar um programa de qualidade, objetivando diminuir o número de tijolos defeituosos produzidos por seus 12 operários. Os números de tijolos defeituosos produzidos antes e após o programa estão na tabela a seguir:

Operário	1	2	3	4	5	6	7	8	9	10	11	12
Antes	47	56	54	49	36	48	51	38	61	49	56	52
Depois	71	34	45	34	50	24	42	35	53	20	67	50

Testes de aderência indicam a necessidade de utilizar testes não paramétricos. Usando nível de significância de 2,5%, o programa de qualidade deu resultados?

23. Há interesse em avaliar o comportamento de um tipo de pneu especial, supostamente de grande durabilidade. Há dois fabricantes, e 30 pneus de cada um foram selecionados, ao acaso, para avaliar sua durabilidade. Os resultados foram:

	101900	40900	170600	151100	92600	91100
	281700	113300	78200	141700	244200	129600
Fabricante 1:	36500	72100	49000	227400	103100	72000
	58300	128100	132400	34800	165900	94200
	113200	87900	92100	97400	99700	87500
	218500	108000	46000	106000	49200	144200
	153400	59600	82800	94900	235400	91800
Fabricante 2:	204400	211100	114600	173700	195300	144300
	493400	586700	59000	235100	137700	102400
	40600	240100	145600	128900	328000	65500

Testes de aderência mostraram a inadequação da distribuição normal para descrever a duração, mas é razoável supor que a distribuição seja simétrica. Pode-se dizer que as vidas médias são diferentes nas populações de onde vieram os pneus, ao nível de significância de 5%?

Correlação e Regressão

11.1 CORRELAÇÃO

Numa população de pessoas, podemos dizer que as variáveis *peso* e *altura* são *correlacionadas positivamente*, pois a maioria dos indivíduos altos também é pesada, enquanto a maioria dos indivíduos baixos é leve. De forma análoga, o *faturamento* de uma empresa e o *nível de utilização* do seu sistema computacional devem ter *correlação positiva*. Já a quantidade de memória RAM e o tempo de processamento devem ter *correlação negativa*.

Dizemos que duas variáveis, X e Y , estão *positivamente correlacionadas* quando elas *caminham num mesmo sentido*, ou seja, elementos com valores pequenos de X tendem a ter valores pequenos de Y e elementos com valores grandes de X tendem a ter valores grandes de Y . Estão *negativamente correlacionadas* quando elas *caminham em sentidos opostos*, ou seja, elementos com valores pequenos de X tendem a ter valores grandes de Y e elementos com valores grandes de X tendem a ter valores pequenos de Y .

É importante ressaltar que o conceito de *correlação* refere-se a uma associação numérica entre duas variáveis, não implicando, necessariamente, relação de *causa-e-efeito*, ou mesmo uma estrutura com interesses práticos. A análise de dados para verificar correlações é usualmente feita em termos exploratórios; verificação de uma correlação serve como elemento auxiliar na análise do problema em estudo. Ou seja, o estudo da correlação numérica entre as observações de duas variáveis é geralmente um passo intermediário na análise de um problema.

Diagramas de dispersão

Uma forma de visualizarmos se duas variáveis apresentam-se correlacionadas é através do *diagrama de dispersão*, onde os valores das variáveis são representados por pontos, num sistema cartesiano.

Exemplo 11.1 No processo de queima de massa cerâmica para pavimento, corpos de prova foram avaliados por três variáveis: X_1 = retração linear (%), X_2 = resistência mecânica (MPa) e X_3 = absorção de água (%). Os resultados de 18 ensaios são apresentados a seguir:

Ensaio	X_1	X_2	X_3	Ensaio	X_1	X_2	X_3
1	8,70	38,42	5,54	10	13,24	60,24	0,58
2	11,68	46,93	2,83	11	9,10	40,58	3,64
3	8,30	38,05	5,58	12	8,33	41,07	5,87
4	12,00	47,04	1,10	13	11,34	41,94	3,32
5	9,50	50,90	0,64	14	7,48	35,53	6,00
6	8,58	34,10	7,25	15	12,68	38,42	0,36
7	10,68	48,23	1,88	16	8,76	45,26	4,14
8	6,32	27,74	9,92	17	9,93	40,70	5,48
9	8,20	39,20	5,63	18	6,50	29,66	8,98

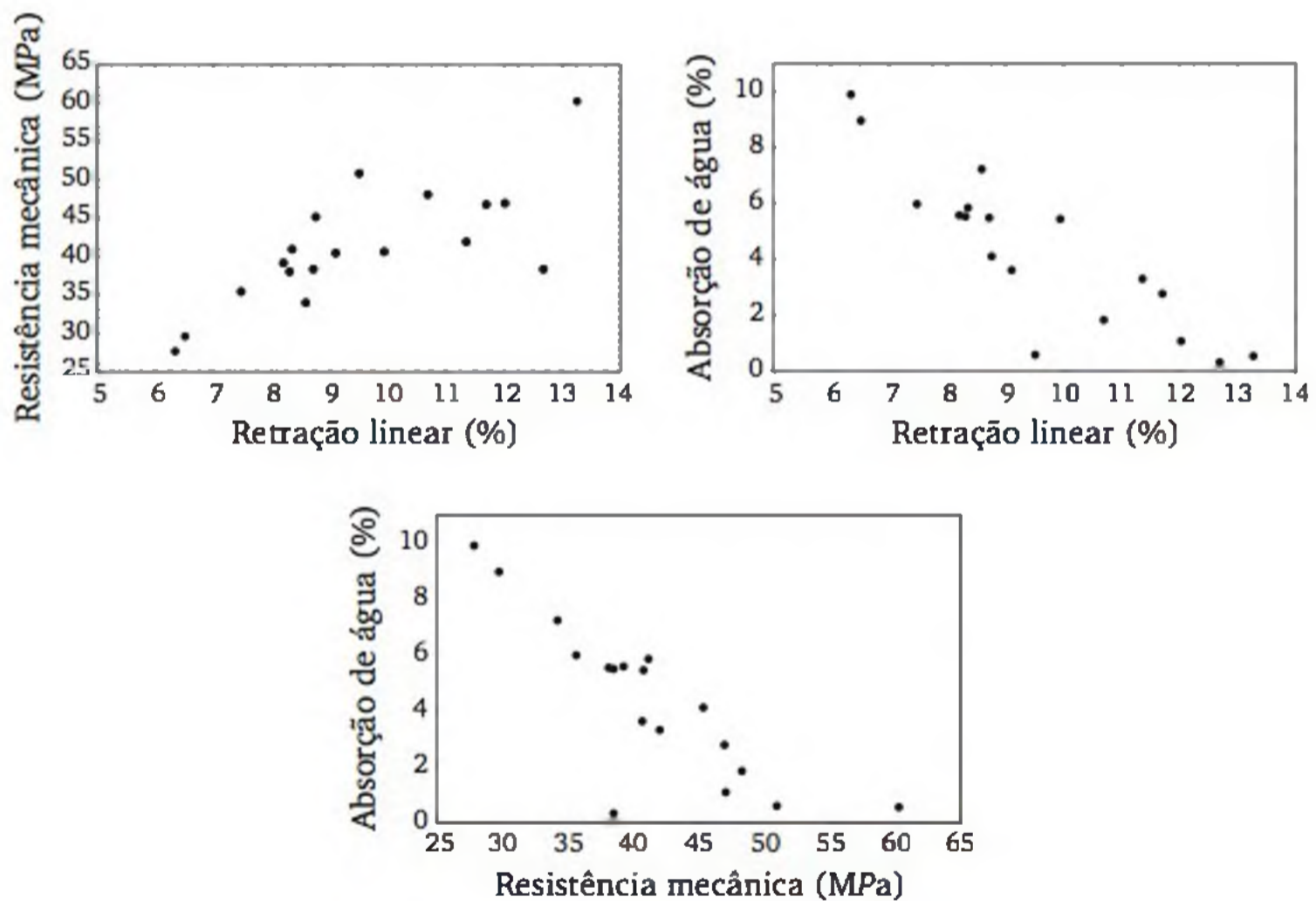


Figura 11.1 Diagramas de dispersão de uma amostra de 18 observações das variáveis retração linear, resistência mecânica e absorção de água, em massas cerâmicas.

A Figura 11.1 sugere que existe correlação positiva entre resistência mecânica e retração linear. E correlação negativa entre absorção de água e retração linear; e entre resistência mecânica e absorção de água.

11.2 COEFICIENTE DE CORRELAÇÃO LINEAR DE PEARSON

Nesta seção, apresentaremos o chamado *coeficiente de correlação (linear) de Pearson*, que descreve a correlação linear dos dados de duas variáveis aleatórias.

A ideia da construção do coeficiente de correlação

O valor do coeficiente de correlação não deve depender da unidade de medida dos dados. Por exemplo, o coeficiente de correlação entre as variáveis *peso* e *altura*, observadas num certo conjunto de indivíduos, deve acusar o mesmo valor, independentemente se o peso for medido em *gramas* ou *quilogramas* e a altura em *metros* ou *centímetros*.

Para evitar o efeito da unidade de medida, consideramos os dados em termos da quantidade de desvio padrão que se afastam da média. Assim, a padronização de $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ é feita da seguinte forma:

$$x'_i = \frac{x_i - \bar{x}}{s_x} \qquad y'_i = \frac{y_i - \bar{y}}{s_y} \quad (i = 1, 2, \dots, n) \qquad (11.1)$$

- onde: \bar{x} : média de x_1, x_2, \dots, x_n ;
- s_x : desvio padrão de x_1, x_2, \dots, x_n ;
- \bar{y} : média de y_1, y_2, \dots, y_n ;
- s_y : desvio padrão de y_1, y_2, \dots, y_n .

A Figura 11.2 ilustra dois diagramas de dispersão entre as variáveis *resistência mecânica* e *retração linear*. O primeiro na escala original e o segundo na escala padronizada.

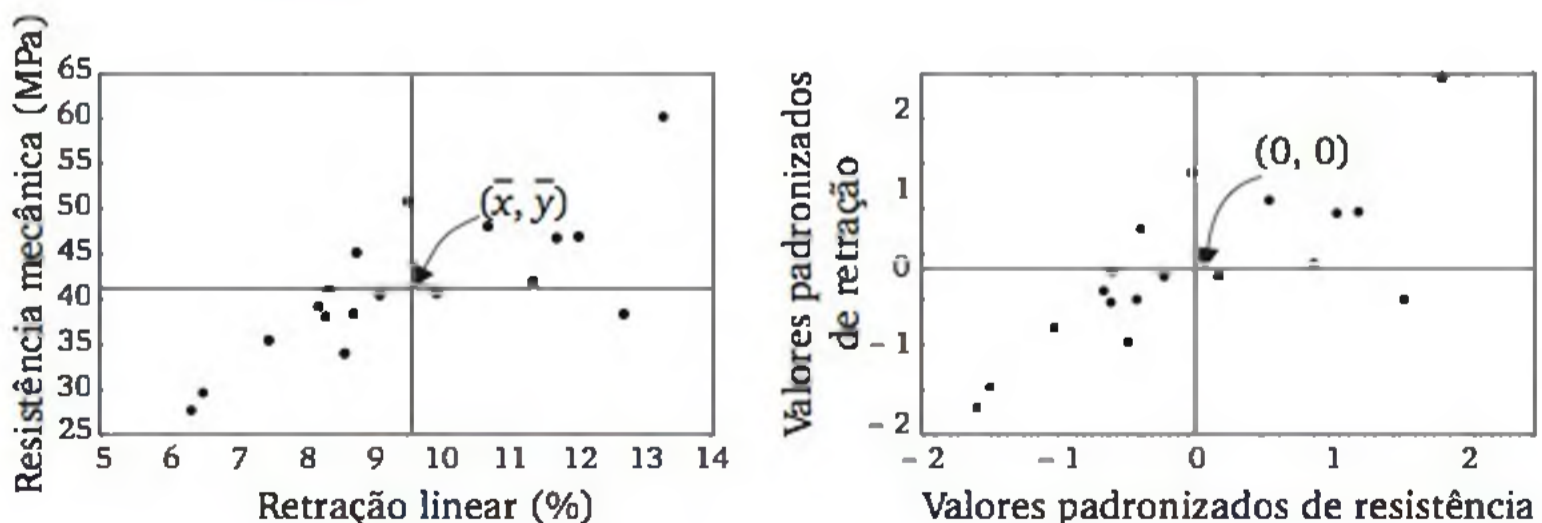


Figura 11.2 Ilustração do efeito da padronização das escalas.

Olhando a Figura 11.2, é possível notar que, no caso de correlação positiva, os pontos tendem a se localizar nos quadrantes I e III (x' e y' com o mesmo sinal). De forma análoga, quando houver correlação negativa, os pontos tendem a se localizar nos quadrantes II e IV (x' e y' com sinais trocados). Assim, a soma dos produtos $x'_i y'_i$ ($i = 1, 2, \dots, n$) deve ter sinal positivo ou negativo, dependendo do sentido da correlação. Baseado nessa ideia, o coeficiente de correlação linear de Pearson, r , é definido pela seguinte expressão, em termos dos valores padronizados:

$$r = \frac{\sum_{i=1}^n (x'_i y'_i)}{n - 1} \quad (11.2)$$

Para ilustrar a obtenção do coeficiente r , considere 3 observações do par de variáveis aleatórias (X, Y) : $(3, 6)$, $(4, 4)$, $(5, 2)$. Temos: $\bar{x} = 3$, $\bar{y} = 4$, $s_x = 1$ e $s_y = 2$. Daí, decorrem os seguintes pares de valores padronizados: $(-1, 1)$, $(0, 0)$, $(1, -1)$. A Figura 11.3 mostra os pontos nos dois sistemas de coordenadas.

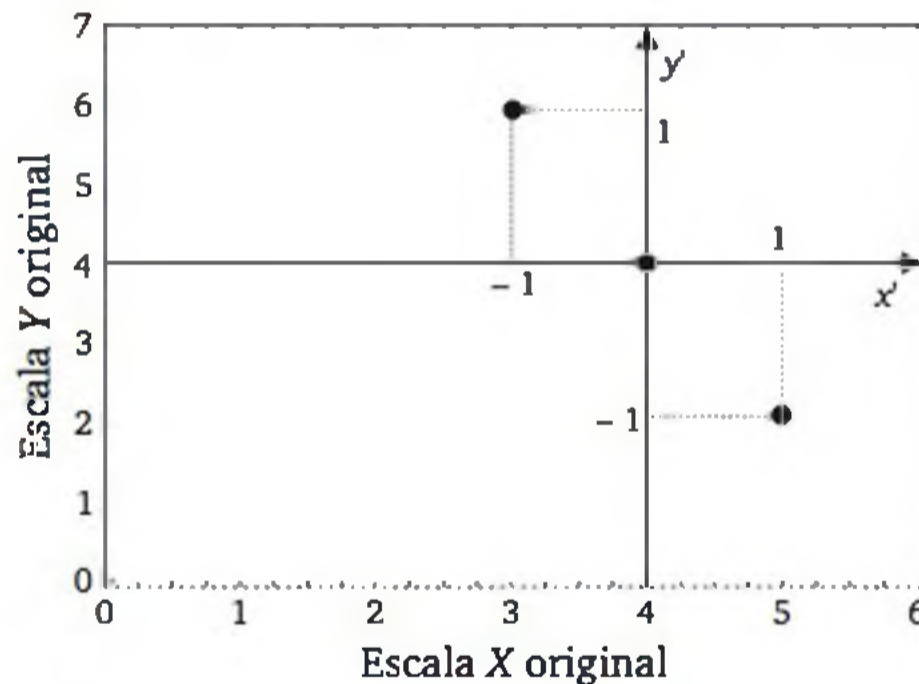


Figura 11.3 Padronização de três observações de (X, Y) .

Calculando o coeficiente de correlação de Pearson:

$$r = \frac{\sum_{i=1}^n (x'_i y'_i)}{n - 1} = \frac{(-1)(1) + (0)(0) + (1)(-1)}{2} = -1$$

Interpretação do coeficiente de correlação

Para qualquer conjunto de dados, podemos demonstrar que o valor do coeficiente de correlação de Pearson, r , estará no intervalo de -1 a 1 . Será posi-

tivo quando os dados apresentarem correlação linear positiva; será *negativo* quando os dados apresentarem correlação linear negativa.

O valor de r será *tão mais próximo* de 1 (ou -1) quanto mais *forte* for a correlação nos dados observados. Teremos $r = +1$ se os pontos estiverem exatamente sobre uma reta ascendente (*correlação positiva perfeita*). Por outro lado, teremos $r = -1$ se os pontos estiverem exatamente sobre uma reta descendente (*correlação negativa perfeita*). Quando não houver correlação nos dados, r acusará um valor próximo de 0 (zero).

A Figura 11.4 mostra os possíveis valores de r e a interpretação em termos do sentido (positivo ou negativo) e da força (fraca, moderada ou forte) da correlação.

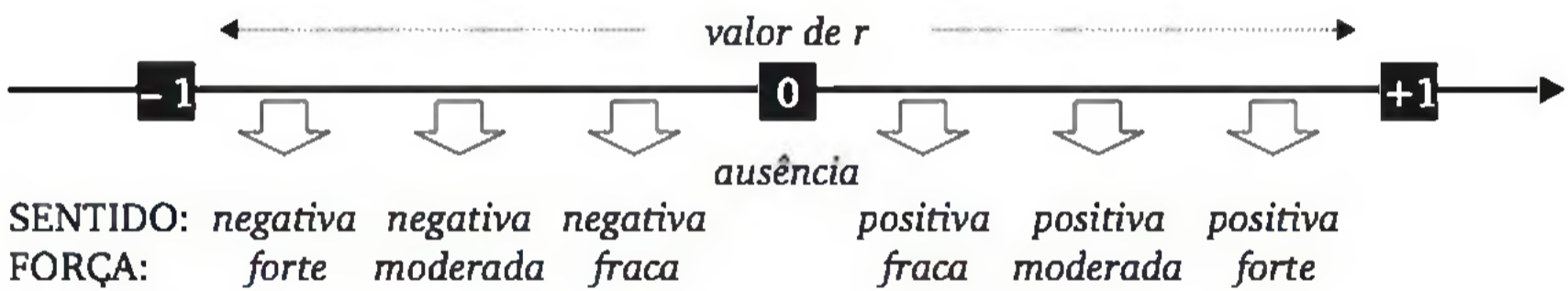


Figura 11.4 Sentido e força da correlação, em termos do coeficiente r .

Exemplo 11.1 (continuação) Com as 18 observações das variáveis retração linear (%), resistência mecânica (MPa) e absorção de água (%), calculamos o coeficiente de correlação de Pearson para cada par das variáveis, como é mostrado a seguir:

	Retração linear	Resistência mecânica	Absorção de água
Retração linear	1,00	0,75	- 0,88
Resistência mecânica	0,75	1,00	- 0,84
Absorção de água	- 0,88	- 0,84	1,00

Observamos que, entre resistência mecânica e retração linear, temos correlação positiva de moderada a forte. Entre retração linear e absorção de água, e entre resistência mecânica e absorção de água, temos correlações negativas fortes.

Cálculo do coeficiente de correlação de Pearson

Efetuar o cálculo do coeficiente de correlação r através dos valores padronizados, além de ser bastante trabalhoso, tem o inconveniente de incorporar erros de arredondamentos no cálculo dos valores padronizados, podendo comprometer o resultado final. Em geral, é conveniente usar a expressão a seguir, que é matematicamente igual à expressão (11.2):¹

$$r = \frac{n \sum (x_i \cdot y_i) - (\sum x_i)(\sum y_i)}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \cdot \sqrt{n \sum y_i^2 - (\sum y_i)^2}} \quad (11.3)$$

Ilustraremos o uso de (11.3) com as 3 observações (3, 6), (4, 4), (5, 2):

i	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	3	6	9	36	18
2	4	4	16	16	16
3	5	2	25	4	10
Soma	12	12	50	56	44

$$r = \frac{3(44) - (12)(12)}{\sqrt{3(50) - (12)^2} \cdot \sqrt{3(56) - (12)^2}} = \frac{-12}{\sqrt{6} \cdot \sqrt{24}} = -1$$

Coeficiente de correlação populacional

De forma análoga, como definimos a medida descritiva de correlação entre as observações em (11.2) e (11.3), podemos definir, em termos probabilísticos, o *parâmetro* correlação entre duas variáveis aleatórias, X e Y , pelo valor esperado do produto destas variáveis padronizadas, ou seja:

$$\rho = \text{Corr}(X, Y) = E \left\{ \left(\frac{X - \mu_X}{\sigma_X} \cdot \frac{Y - \mu_Y}{\sigma_Y} \right) \right\} \quad (11.4)$$

onde $\mu_X = E(X)$, $\mu_Y = E(Y)$, $\sigma_X = \sqrt{V(X)}$ e $\sigma_Y = \sqrt{V(Y)}$

1 Por simplicidade, daqui para frente excluiremos os índices dos somatórios.

Inferência sobre ρ

Dada uma amostra aleatória simples $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ de n observações do par de variáveis aleatórias (X, Y) , o coeficiente r , calculado por (11.2) ou (11.3), pode ser considerado uma *estimativa* do verdadeiro e desconhecido coeficiente ρ .

É comum o interesse em verificar as seguintes hipóteses:

$H_0: \rho = 0$ (as variáveis X e Y são *não correlacionadas*);

$H_1: \rho \neq 0$ (as variáveis X e Y são *correlacionadas*).

podendo, ainda, a hipótese alternativa indicar o sentido da correlação (teste unilateral), tal como $H_1': \rho > 0$ (X e Y são *correlacionadas positivamente*) ou $H_1'': \rho < 0$ (X e Y são *correlacionadas negativamente*). O teste unilateral é aplicado nos casos em que esperamos o coeficiente de correlação com determinado sinal (+ ou -).

Restringindo-se à verificação de correlação linear e supondo X e Y com distribuições normais, podemos realizar o teste, calculando

$$t = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}} \quad (11.5)$$

e usando como distribuição de referência a t de Student com $gl = n - 2$. A Tabela 10 do apêndice apresenta uma forma mais rápida para esse teste. Para cada n , a Tabela 10 apresenta o valor absoluto mínimo de r para se rejeitar H_0 .

Exemplo 11.1 (continuação) Calculamos, anteriormente, as seguintes correlações baseadas em $n = 18$ observações:

	Retração linear	Resistência mecânica	Absorção de água
Retração linear	1,00	0,75	- 0,88
Resistência mecânica	0,75	1,00	- 0,84
Absorção de água	- 0,88	- 0,84	1,00

Considerando testes bilaterais ao nível de significância de 5%, verificamos na Tabela 10 que, para $n = 18$, o valor absoluto mínimo para a correlação ser significativa (rejeitar H_0) é 0,468. Como os três coeficientes calculados são, em valor absoluto, superiores a 0,468, concluímos que as três medidas usadas para avaliar a qualidade da cerâmica são realmente correlacionadas.

EXERCÍCIOS

1. Calcule o coeficiente de correlação de Pearson entre retração linear (%) e resistência mecânica (MPa) para as 5 primeiras observações apresentadas no Exemplo 11.1. Apenas com estas observações, o teste estatístico detecta correlação real entre as duas variáveis?
2. Com respeito aos 23 alunos de uma turma de estatística, foram observadas as variáveis *número de faltas* e *nota final na disciplina*. Esses dados acusaram a seguinte correlação, descrita pelo coeficiente de correlação de Pearson: $r = -0,56$. Comente as seguintes frases relativas à turma em estudo e ao coeficiente obtido.
 - a) “Como $r = -0,56$ (correlação negativa moderada), nenhum aluno com grande número de faltas tirou nota alta.”
 - b) “Como as duas variáveis são correlacionadas, bastaria usar uma delas como critério de avaliação, pois uma acarreta a outra.”
 - c) “Os dados observados mostraram uma leve tendência de que a nota final se relaciona inversamente com o número de faltas, ou seja, os alunos *frequentadores* tiveram, em geral, melhor desempenho nas avaliações do que os alunos que faltaram muito.”
3. Sejam $X = \text{nota na prova do vestibular de matemática}$ e $Y = \text{nota final na disciplina de cálculo}$. Estas variáveis foram observadas em 20 alunos, ao final do primeiro período letivo de um curso de engenharia. Os dados são apresentados a seguir:

X	Y	X	Y	X	Y	X	Y	X	Y
39	65	43	78	21	52	64	82	65	88
57	92	47	89	28	73	75	98	47	71
34	56	52	75	35	50	30	50	28	52
40	70	70	50	80	90	32	58	67	88

- a) Calcule a correlação entre a *nota no vestibular de matemática* e a *nota na disciplina de cálculo*. Interprete o resultado.
- b) Construa um diagrama de dispersão e verifique se algum aluno *foge* ao comportamento geral dos demais (ponto discrepante).
- c) Retire o valor discrepante detectado no item anterior e calcule novamente o coeficiente r . Interprete.
- d) Verifique se a correlação encontrada no item anterior é significativa. Faça o teste ao nível de significância de 5% e interprete o resultado.

4. No desenvolvimento computacional de um escalonador, foram realizados alguns testes em 16 condições experimentais diferentes. O desempenho do escalonador foi observado através da quantidade de trabalho executado, num certo período de tempo: em processamento de textos (X_1), em processamento interativo de dados (X_2) e em processamento de dados em *batch* (X_3). Os coeficientes de correlação calculados sobre as 16 observações foram:

	X_1	X_2	X_3
X_1	1,00	0,18	0,86
X_2	0,18	1,00	0,02
X_3	0,86	0,02	1,00

Que informações podem ser extraídas dessa matriz de correlações? Há evidências de que nas situações em que desempenho é melhor para um atributo (por exemplo, processamento de texto), ele também tende a ser melhor em outro atributo (por exemplo, processamento interativo de dados)?

11.3 REGRESSÃO LINEAR SIMPLES

O termo *regressão* surgiu com os trabalhos de Galton no final do século XIX. Esses trabalhos procuravam explicar certas características de um indivíduo a partir das características de seus pais, como, por exemplo, predizer a altura de um indivíduo em função das alturas de seus pais. O modelo matemático-estatístico foi aperfeiçoado e hoje é utilizado nas mais variadas áreas, em particular na engenharia e na informática.

Iniciaremos o estudo de regressão com a formulação mais simples, relacionando uma variável Y , chamada de variável *resposta* ou *dependente*, com uma variável X , denominada de variável *explicativa* ou *independente*.² Veja o Quadro 11.1.

² Dizemos que a regressão é *simples*, porque estamos considerando apenas uma variável independente.

Quadro 11.1 Exemplos de aplicação da regressão linear simples.

Variável independente, X	Variável dependente, Y
Temperatura do forno (°C)	Resistência mecânica da cerâmica (MPa)
Quantidade de aditivo (%)	Octanagem da gasolina
Renda (R\$)	Consumo (R\$)
Memória RAM do computador (Gb)	Tempo de resposta do sistema (s)
Área construída do imóvel (m ²)	Preço do imóvel (R\$)

Observe que os exemplos do Quadro 11.1 se distinguem dos exemplos sobre correlação por suporem uma relação de causalidade entre X e Y . É esta a diferença básica de um estudo de correlação e uma análise de regressão. A aplicação da análise de regressão é geralmente feita sob um referencial teórico, que justifique uma relação matemática de causalidade. Além disso, a variável X normalmente é controlada (não aleatória) e Y é uma variável aleatória.

Assim como num estudo de correlação, a análise de regressão também parte de um conjunto de observações pareadas $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, relativas às variáveis X e Y .

Exemplo 11.2 Considere um experimento em que se analisa a octanagem da gasolina (Y) em função da adição de um novo aditivo (X). Para isso, foram realizados ensaios com os percentuais de 1, 2, 3, 4, 5 e 6% de aditivo. Os resultados são mostrados na Figura 11.5.

X	Y
1	80,5
2	81,6
3	82,1
4	83,7
5	83,9
6	85,0

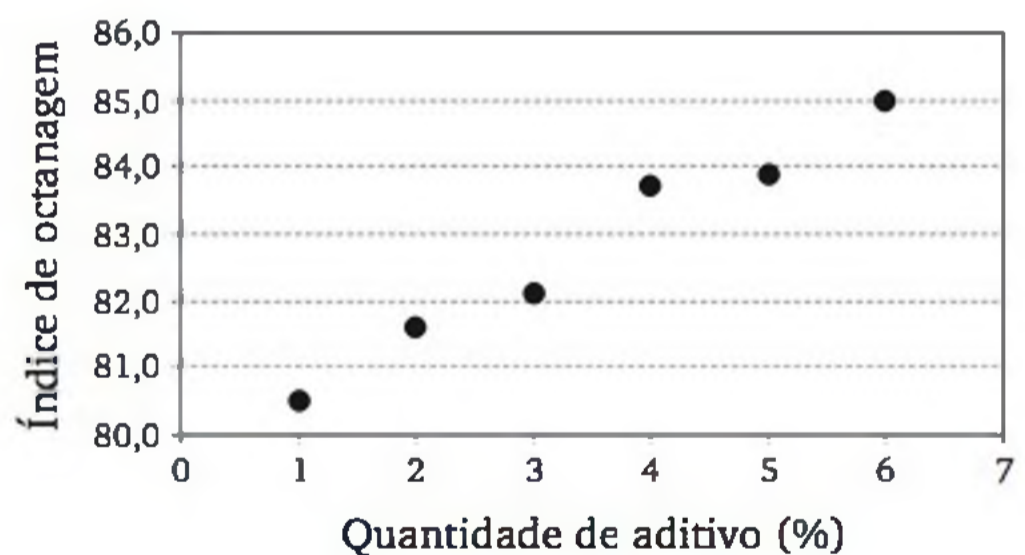


Figura 11.5 Dados experimentais do efeito de um aditivo (X) na octanagem da gasolina (Y).

Observe que é razoável supor uma relação aproximadamente linear entre X e Y para os níveis de aditivo ensaiados (de 1 a 6%). Contudo, os pontos não

estão exatamente sobre uma reta, provavelmente por causa da existência de fatores não controláveis no processo. Vamos supor, então, que o valor esperado de Y varie com X , de acordo com uma equação de primeiro grau, ou seja:

$$E\{Y\} = \alpha + \beta X \quad (11.6)$$

onde α e β são os parâmetros do modelo.

Seja um conjunto de observações $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. O chamado modelo de *regressão linear simples* para as observações é dado por

$$Y_i = \alpha + \beta x_i + \varepsilon_i \quad (11.7)$$

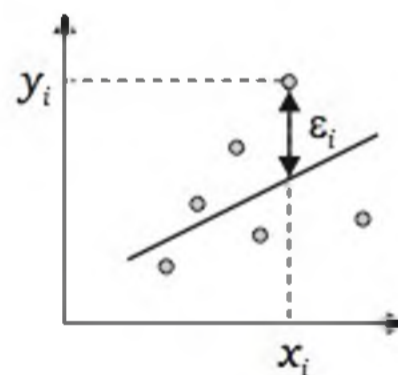
onde: Y_i é a variável aleatória associada à i -ésima observação de Y ; e

ε_i é o *erro aleatório* da i -ésima observação, isto é, o efeito de uma infinidade de fatores que estão afetando a observação de Y de forma aleatória.

Note que é razoável supor $E\{\varepsilon_i\} = 0$ ($i = 1, 2, \dots, n$). Assim, considerando X uma variável controlável (não aleatória), temos, pelas propriedades do valor esperado (Capítulo 5): $E\{Y_i\} = \alpha + \beta x_i$, compatível com (11.6) para a i -ésima observação.

11.3.1 Método dos mínimos quadrados

Para a construção do modelo descrito em (11.6), precisamos obter estimativas para α e β , a partir de um conjunto de observações $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. Ou seja, queremos encontrar a reta que passe o *mais próximo possível* dos pontos observados.



Há vários métodos para estimar os parâmetros do modelo. O mais usual é o *método de mínimos quadrados*, que consiste em fazer com que a soma dos erros quadráticos seja a menor possível. Considerando o modelo (11.7), temos que o erro aleatório da i -ésima observação ($i = 1, 2, \dots, n$) é dado por

$$\varepsilon_i = Y_i - (\alpha + \beta x_i) \quad (11.8)$$

conforme ilustramos anteriormente.

O método consiste em obter os valores de α e β que minimizam a expressão:

$$S = \sum \varepsilon_i^2 = \sum \{Y_i - (\alpha + \beta x_i)\}^2 \quad (11.9)$$

que pode ser feito igualando as derivadas parciais a zero, ou seja:³

$$\frac{\partial S}{\partial \alpha} = 0 \text{ e } \frac{\partial S}{\partial \beta} = 0 \quad (11.10)$$

resultando nas seguintes estimativas para α e β , as quais chamaremos de a e b , respectivamente:⁴

$$b = \frac{n \cdot \sum (x_i y_i) - (\sum x_i) \cdot (\sum y_i)}{n \cdot \sum x_i^2 - (\sum x_i)^2} \quad (11.11)$$

$$a = \frac{\sum y_i - b \sum x_i}{n} \quad (11.12)$$

onde $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ é a amostra efetivamente observada.

A chamada *equação (reta) de regressão* é dada por

$$\hat{y} = a + bx \quad (11.13)$$

Para cada valor x_i ($i = 1, 2, \dots, n$), temos, pela equação de regressão, o valor *predito*:

$$\hat{y}_i = a + bx_i \quad (11.14)$$

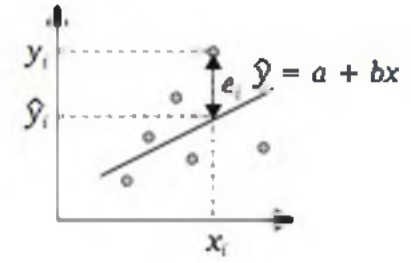
A diferença entre os valores observados e os preditos é chamada de *resíduo*:

$$e_i = y_i - \hat{y}_i \quad (11.15)$$

³ Pode-se mostrar que a segunda derivada é negativa, acarretando que a solução encontrada é realmente um ponto de mínimo da função (11.9).

⁴ É possível mostrar que os *estimadores* associados às estimativas a e b são não viciados e os mais eficientes, dentre as possíveis relações lineares com as observações de Y .

O resíduo relativo à i -ésima observação (e_i) pode ser considerado uma estimativa do erro aleatório (ε_i) desta observação (veja ilustração ao lado).



Exemplo 11.2 (continuação)

Tabela 11.1 *Dados do experimento e cálculos intermediários para obter a equação de regressão.*

Dados			Cálculos intermediários	
Ensaio (i)	x_i	y_i	x_i^2	$x_i y_i$
1	1	80,5	1	80,5
2	2	81,6	4	163,2
3	3	82,1	9	246,3
4	4	83,7	16	334,8
5	5	83,9	25	419,5
6	6	85,0	36	510,0
Soma	21	496,8	91	1.754,3

$$b = \frac{6 \cdot (1754,3) - (21) \cdot (496,8)}{6 \cdot (91) - (21)^2} = \frac{93}{105} = 0,886$$

$$a = \frac{496,8 - (0,886) \cdot (21)}{6} = 79,7$$

Assim, temos a seguinte reta de regressão:

$$\hat{y} = 79,7 + (0,886)x$$

Para traçar a reta no plano, basta atribuir dois valores para x e calcular os correspondentes valores de \hat{y} (veja a Figura 11.6).

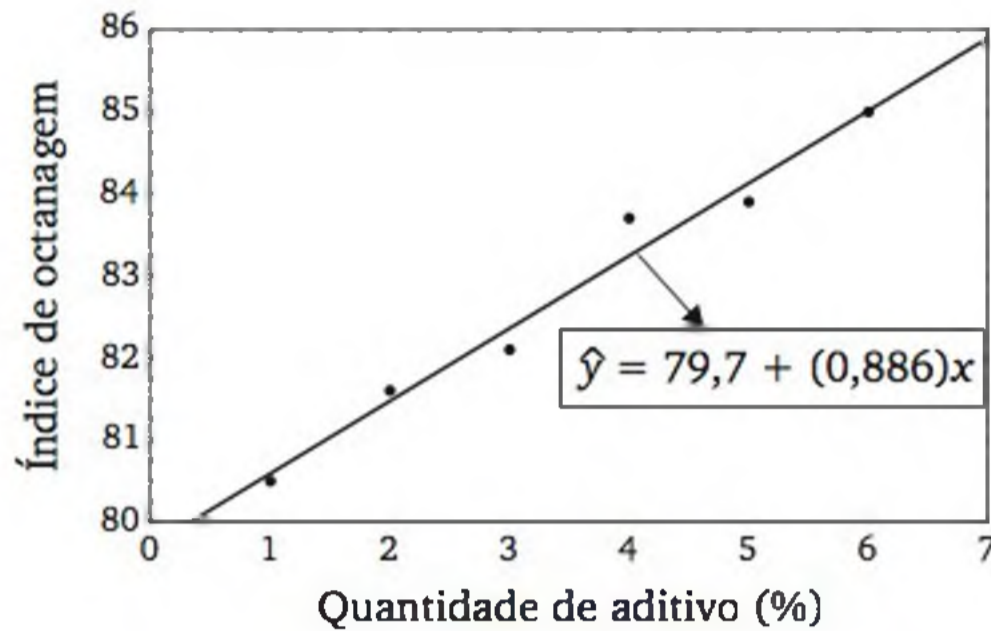


Figura 11.6 Diagrama de dispersão dos dados do Exemplo 11.2 e a reta de regressão ajustada a esses dados.

A partir dos seis ensaios experimentais, construímos um modelo, o qual nos permite *prever* o índice de octanagem da gasolina (\hat{y}) a partir de uma quantidade do novo aditivo (x).⁵ Por exemplo, se for adicionado $x = 5,5\%$ de aditivo, esperamos um índice de octanagem de $\hat{y} = 79,7 + (0,886).(5,5) = 84,573$. A Tabela 11.2 mostra que os valores preditos pelo modelo estão bastante próximos dos valores observados no experimento.

Tabela 11.2 Valores preditos [$\hat{y}_i = 79,7 + (0,886)x_i$] e resíduos ($e_i = y_i - \hat{y}_i$).

x_i	y_i	\hat{y}_i	e_i
1	80,5	80,586	- 0,086
2	81,6	81,472	0,128
3	82,1	82,358	- 0,258
4	83,7	83,244	0,456
5	83,9	84,130	- 0,230
6	85,0	85,016	- 0,016

O coeficiente b fornece uma estimativa da variação esperada de Y , a partir da variação de *uma* unidade em X . O sinal deste coeficiente indica o sentido da variação. No exemplo, podemos dizer: a cada 1% a mais do novo aditivo, esperamos um aumento de 0,886 no índice de octanagem.

⁵ Ressaltamos que o modelo só deve ser usado para realizar previsões no intervalo de X ensaiado (de 1 a 6% de aditivo), pois não há informação sobre o relacionamento entre X e Y fora deste intervalo.

11.3.2 Análise de variância do modelo

Se X não influencia Y , então o valor esperado de Y pode ser estimado simplesmente pela média aritmética (\bar{y}) das observações de Y . Mas se existe influência de X sobre Y , então deve haver algum ganho em considerar a equação de regressão ($\hat{y} = a + bx$). Este ganho pode ser avaliado ao comparar os resíduos nas duas situações (Figura 11.7).

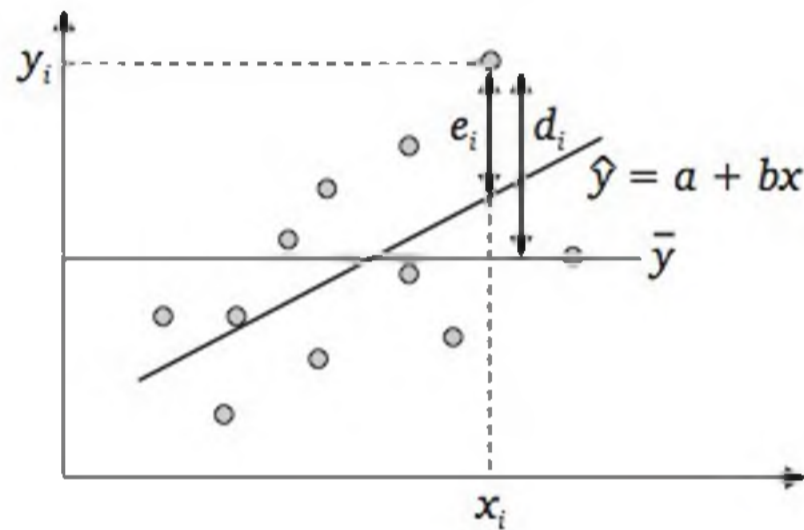


Figura 11.7 Ilustração de resíduos em relação à média aritmética ($d_i = y_i - \bar{y}$) e à equação de regressão ($e_i = y_i - \hat{y}_i$).

Para $i = 1, 2, \dots, n$, sejam:

- a) $y_i - \bar{y}$ (desvios em relação à média aritmética – não levam em consideração a relação entre Y e X);
- b) $y_i - \hat{y}_i$ (desvios em relação aos valores preditos pela equação de regressão – consideram uma relação linear entre Y e X);
- c) $\hat{y}_i - \bar{y}$ (desvios dos valores preditos em relação à média aritmética).
Veja a Figura 11.8.

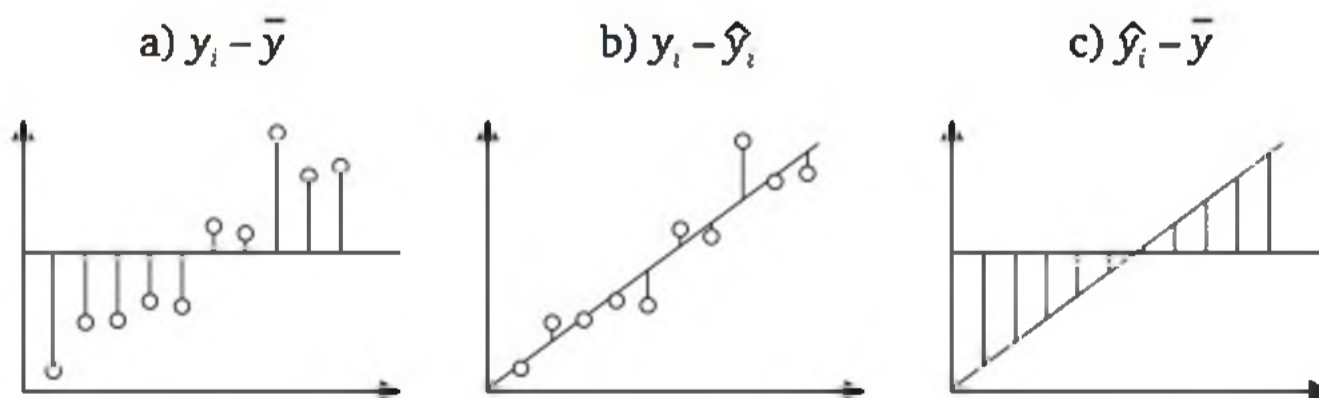


Figura 11.8 Ilustração dos desvios numa situação hipotética.

As somas dos quadrados dos desvios satisfazem à seguinte equação:

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2 \quad (11.16)$$

variação total *variação explicada pela equação de regressão* *variação não explicada*

Chamaremos de *coeficiente de determinação* a seguinte razão:

$$R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = \frac{\text{variação explicada}}{\text{variação total}} \quad (11.17)$$

O coeficiente de determinação é uma medida descritiva da proporção da variação de Y que pode ser *explicada* por variações em X , segundo o modelo especificado.

Exemplo 11.2 (continuação)

x_i	y_i	\bar{y}	\hat{y}_i	$y_i - \bar{y}$	$y_i - \hat{y}_i$	$\hat{y}_i - \bar{y}$	$(y_i - \bar{y})^2$	$(y_i - \hat{y}_i)^2$	$(\hat{y}_i - \bar{y})^2$
1	80,5	82,8	80,59	- 2,3	- 0,09	- 2,21	5,29	0,01	4,90
2	81,6	82,8	81,47	- 1,2	0,13	- 1,33	1,44	0,02	1,77
3	82,1	82,8	82,36	- 0,7	- 0,26	- 0,44	0,49	0,07	0,20
4	83,7	82,8	83,24	0,9	0,46	0,44	0,81	0,21	0,20
5	83,9	82,8	84,13	1,1	- 0,23	1,33	1,21	0,05	1,77
6	85,0	82,8	85,01	2,2	- 0,01	2,21	4,84	0,00	4,90
Soma de Quadrados:							14,08	0,35	13,73

Coeficiente de determinação:

$$R^2 = \frac{13,73}{14,08} = 0,975 = 97,5\%$$

Em termos dos 6 ensaios realizados, a variância da octanagem da gasolina é explicada, em parte, pela variação da quantidade de aditivo adicionado ($R^2 = 97,5\%$ de explicação) e em parte ($1 - R^2 = 2,5\%$) devido a outros fatores intervenientes no processo.

No caso do modelo de regressão linear simples, R^2 coincide, numericamente, com o quadrado do coeficiente de correlação r de Pearson, estudado na seção 11.2.

Processo simplificado de cálculo

Soma de quadrados totais (corrigida pela média aritmética):

$$SQT = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n} \quad (11.18)$$

Soma de quadrados do erro ou soma de quadrados dos resíduos:

$$SQE = \sum (y_i - \hat{y}_i)^2 = \sum y_i^2 - a \sum y_i - b \sum x_i y_i \quad (11.19)$$

Soma de quadrados da regressão:

$$SQR = SQT - SQE \quad (11.20)$$

Coefficiente de determinação:

$$R^2 = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT} \quad (11.21)$$

O leitor poderá refazer os cálculos do Exemplo 11.2 com as expressões (11.18) a (11.21). Os resultados devem ser os mesmos encontrados anteriormente.

Cada soma de quadrados está associada a certo número de *graus de liberdade*. Os desvios de cada observação em relação às estimativas de $E\{Y\}$ têm graus de liberdade iguais a n subtraído do número de parâmetros estimados em $E\{Y\}$. Assim, os desvios $y_i - \bar{y}$ têm $n - 1$ graus de liberdade; e os desvios $y_i - \hat{y}_i$ têm $n - 2$ graus de liberdade.

A soma de quadrados dividida pelo correspondente grau de liberdade fornece o *quadrado médio* ou *variância*. E a razão entre o *quadrado médio da regressão* e o *quadrado médio do erro* resulta na chamada *razão F* (veja Tabela 11.3).

Tabela 11.3 *Análise de variância (Anova) da regressão linear simples.*

Fonte de variação	g_i	SQ	QM	Razão F
Regressão	1	$SQR = \sum (\hat{y}_i - \bar{y})^2$	$QMR = \frac{SQR}{1}$	$f = \frac{QMR}{QME}$
Erro	$n - 2$	$SQE = \sum (y_i - \hat{y}_i)^2$	$QME = \frac{SQE}{n - 2}$	
Total	$n - 1$	$SQT = \sum (y_i - \bar{y})^2$	$QMT = \frac{SQT}{n - 1}$	

Exemplo 11.2 (continuação) Tabela da Anova:

Fonte de variação	gl	SQ	MQ	Razão F
Regressão	1	13,73	13,729	156,26
Erro	4	0,35	0,088	
Total	5	14,08		

11.3.3 Inferências sobre o modelo de regressão

Com respeito ao Exemplo 11.2, se fizéssemos outro experimento sob as mesmas condições e com os mesmos valores de X , não deveríamos encontrar exatamente os mesmos seis valores de Y descritos no Exemplo 11.2, porque o erro experimental deverá estar presente no processo. Porém, se o experimento for bem controlado, deveremos encontrar valores bastante próximos e, conseqüentemente, uma reta de regressão também bastante próxima da encontrada anteriormente.

Podemos imaginar que existe uma *equação de regressão verdadeira*, a qual poderia ser obtida se pudéssemos realizar o experimento com infinitos ensaios. Possíveis amostras de (X, Y) devem seguir, aproximadamente, o comportamento da *equação de regressão verdadeira*, de tal forma que a variação dos pontos em torno desta equação pode ser caracterizada por uma distribuição de probabilidades, conforme mostra a Figura 11.9.

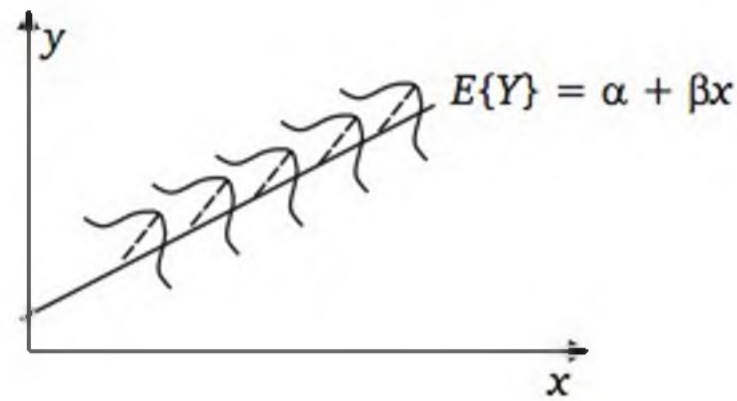


Figura 11.9 Ilustração da distribuição de probabilidades em torno da verdadeira regressão.

Suposições do modelo

Considerando novamente o modelo de regressão linear simples:

$$Y_i = \alpha + \beta x_i + \varepsilon_i \quad (11.22)$$

vamos supor:

- 1) os termos de *erro* ($\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$) são variáveis aleatórias *independentes*;
- 2) $E\{\varepsilon_i\} = 0$;
- 3) $V\{\varepsilon_i\} = \sigma^2$; e
- 4) ε_i tem distribuição normal ($i = 1, 2, \dots, n$).

A primeira suposição exige que o i -ésimo ensaio (ou observação) não seja afetado pelos ensaios (ou observações) anteriores; a segunda exige que a escolha da forma da equação da média do processo – no caso, uma reta – deve ser correta; a terceira exige que a variação aleatória não se altere ao longo das observações; e a quarta aponta para uma forma particular de distribuição dos termos aleatórios. Na Seção 11.3.4, estudaremos como verificar a validade dessas suposições.

Teste de significância do modelo

Com as suposições (1) – (4) é possível testar formalmente se o modelo de regressão propicia ganhos significativos em relação à utilização da média aritmética simples, na estimação de $E\{Y\}$. Considerando o modelo

$$E\{Y\} = \alpha + \beta \cdot X \quad (11.23)$$

formulamos as hipóteses

$$H_0: \beta = 0 \quad \text{e} \quad H_1: \beta \neq 0$$

Note que, sob H_0 , o valor esperado de Y (a média do processo) não é afetado por X . Esse teste pode ser realizado através da razão F , que, sob H_0 e as suposições (1) – (4), tem *distribuição F* com $gl = 1$ no numerador e $gl = n - 2$ no denominador.

No Exemplo 11.2, temos $gl = 1$ no numerador e $gl = 4$ no denominador. Para o nível de significância de 5%, a Tabela 6 do Apêndice fornece o valor crítico $f_c = 7,71$. Como $f = 156,26 > f_c = 7,71$, o teste rejeita H_0 , indicando que o modelo de regressão propicia ganhos significativos em relação à simples utilização da média aritmética dos valores de Y , na avaliação da média do processo.

Inferência sobre o coeficiente angular

A grandeza

$$s_e = \sqrt{QME} = \sqrt{\frac{SQE}{n-2}} \quad (11.24)$$

ou, de forma mais operacional,

$$s_e = \sqrt{\frac{\sum y_i^2 - a \sum y_i - b \sum x_i y_i}{n-2}} \quad (11.25)$$

corresponde ao *desvio padrão dos resíduos*, ou seja, uma estimativa do desvio padrão do *erro aleatório*, σ . A partir de (11.25), podemos ter uma estimativa do *erro padrão* de b por

$$s_b = s_e \cdot \sqrt{\frac{n}{n \sum x_i^2 - (\sum x_i)^2}} \quad (11.26)$$

e testar as hipóteses do tipo $H_0: \beta = \beta_0$ vs. $H_1: \beta \neq \beta_0$ (β_0 é uma constante especificada), calculando

$$t = \frac{b - \beta_0}{s_b} \quad (11.27)$$

e comparando com o valor tabelado t_c da distribuição t (Tabela 4 do apêndice) com $gl = n - 2$. Se $\beta_0 = 0$, este teste é equivalente ao *teste F* visto anteriormente.⁶

⁶ O teste também pode ser unilateral, mas alertamos que a maioria dos pacotes computacionais apresenta o valor p associado a um teste bilateral.

Dado certo nível de confiança γ , podemos obter na Tabela 4 o valor correspondente de t_γ , com $gl = n - 2$, e construir um intervalo de confiança para β por

$$IC(\beta, \gamma) = b \pm t_\gamma \cdot s_b \quad (11.28)$$

Inferência sobre o coeficiente escalar

Uma estimativa do *erro padrão* de a é dada por

$$s_a = s_e \cdot \sqrt{\frac{1}{n} + \frac{(\sum x_i)^2}{n \sum x_i^2 - (\sum x_i)^2}} \quad (11.29)$$

Para testar $H_0: \alpha = \alpha_0$ vs. $H_1: \alpha \neq \alpha_0$ (α_0 é uma constante especificada), calcula-se:

$$t = \frac{a - \alpha_0}{s_a} \quad (11.30)$$

e compara com o valor crítico t_α , obtido na Tabela 4, com $gl = n - 2$. Em especial, se for adotado $\alpha_0 = 0$, este teste avalia se é razoável supor que a reta de regressão passe pela origem.

Podemos obter um intervalo de confiança para α por

$$IC(\alpha, \gamma) = a \pm t_\gamma \cdot s_a \quad (11.31)$$

onde t_γ é obtido na Tabela 4, em função do nível de confiança γ estabelecido e dos graus de liberdade ($gl = n - 2$).

Uso do computador

Exemplo 11.3 O anexo deste capítulo contém dados relativos à venda de 142 automóveis *seminovos*, incluindo o modelo, o preço de revenda (R\$), o preço do modelo novo (R\$), o tempo de uso (anos completos) e a quilometragem.

O preço de venda de um carro *seminovo* depende do preço do mesmo modelo 0 km. Assim, procuramos estabelecer um modelo de regressão entre o pre-

ço de revenda (Y) e o preço do correspondente modelo 0 km (X). Usando a planilha *Excel*, obtivemos os seguintes resultados:⁷

Estatística de regressão	
R múltiplo	0,889
R-quadrado	0,791
R-quadrado ajustado	0,789
Erro padrão	1778,484
Observações	142

ANOVA

	SQ	gl	QM	F	Valor p
Regressão	1,67E + 09	1	1,67E+09	528,5782	2,22E-49
Resíduo (erro)	4,43E + 08	140	3163004		
Total	2,11E + 09	141			

	Coeficientes	Erro padrão	Estat. t	Valor-p	Inferior 95,0%	Superior 95,0%
Interseção	2654,11	431,22	6,155	7,46E-09	1801,56	3.506,67
Valor novo	0,476	0,021	22,991	2,22E-49	0,43	0,52

A primeira tabela de resultados mostra algumas estatísticas e, em particular, o R^2 (*R-quadrado*) igual a 0,791. Este resultado indica que, na amostra observada, cerca de 79% da variação do preço de revenda pode ser *explicada* por uma relação linear com o preço do automóvel 0 km. Os demais 21% podem ser considerados como a variação provocada por outros fatores não considerados no modelo de regressão.

O *R-múltiplo*, no caso de regressão linear simples, corresponde ao Coeficiente de Correlação de Pearson. O erro padrão (s_e) já foi apresentado anteriormente e R^2 *ajustado* não será tratado neste texto.

A segunda tabela apresenta a análise de variância (Anova) do modelo. Temos: $f = 528$ [valor $p = (2,22) \cdot 10^{-49}$], indicando que o valor do carro novo (X) é significativo para *explicar* o preço do carro *seminovo* (Y).

⁷ Para fazer uma análise de regressão no Excel, é necessário instalar as Ferramentas de Análise (*ferramentas* → *suplementos*) e, depois, proceder a análise (*ferramentas* → *análise de dados* → *regressão*). Nas tabelas de resultados apresentadas neste texto, alguns termos foram adequados de acordo com nossa terminologia.

A primeira coluna da terceira tabela apresenta as estimativas dos coeficientes, donde, no presente exemplo, temos a seguinte equação de predição para o preço de revenda (Y), em função do preço do automóvel novo (X):

$$\hat{y} = 2654,11 + (0,476)x$$

ou seja, tendo o preço de um particular carro novo, x , podemos obter uma predição para o preço de revenda, \hat{y} . Por exemplo, um modelo no qual o preço de novo é R\$ 16.000,00 tem preço de revenda, predito pelo modelo, de

$$\hat{y} = 2654,11 + (0,476)(16000) = 10270$$

ou seja, R\$ 10.270,00.

Com a equação de regressão, observamos, também, que, a cada real de diferença no carro novo, esperamos uma diferença de 0,476 reais na revenda.⁸

A última tabela também fornece os resultados de testes estatísticos sobre cada um dos parâmetros do modelo. As duas últimas colunas desta tabela apresentam um intervalo de 95% de confiança para os dois parâmetros do modelo (o intercepto α e a inclinação β).

11.3.4 Análise dos resíduos

Na seção anterior, estabelecemos um modelo para um conjunto de observações $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, relativo às variáveis X e Y , da forma

$$Y_i = \alpha + \beta x_i + \varepsilon_i \quad (11.32)$$

onde: Y_i é a variável aleatória associada à i -ésima observação de Y ;

ε_i é o *erro aleatório* da i -ésima observação; e

α e β são parâmetros a serem estimados com os dados.

Assim, estamos assumindo que X causa Y através de uma relação linear, e toda a variação em torno dessa relação deve-se ao efeito (*erro*) aleatório. Além disso, para a validade dos intervalos de confiança e testes estatísticos, é necessário supor que as observações de Y sejam independentes, e o termo de *erro* tenha distribuição normal, com média nula e variância constante. Apresentaremos um processo gráfico para verificar se estas suposições podem ser válidas e, caso contrário, o que pode ser feito para corrigir as distorções.

⁸ É claro que um bom modelo para o preço de revenda deve levar em conta outros fatores, tais como a idade do veículo, estado de conservação etc. Na Seção 11.4 incluiremos novas variáveis no modelo.

Um primeiro gráfico pode ser feito antes de se aplicar a análise de regressão. É o diagrama de dispersão, conforme discutido na Seção 11.1. Por esse gráfico, podemos verificar se a função linear é adequada para representar a forma estrutural entre X e Y . Veja o gráfico à esquerda da Figura 11.10.

Após a estimação dos parâmetros α e β , podemos calcular os *resíduos*, ou seja,

$$e_i = y_i - \hat{y}_i \quad (i = 1, 2, \dots, n) \quad (11.33)$$

Um gráfico apresentando os pares (x_i, e_i) , que chamaremos de *gráfico dos resíduos*, é bastante útil na avaliação do modelo de regressão. Veja o gráfico à direita da Figura 11.10.

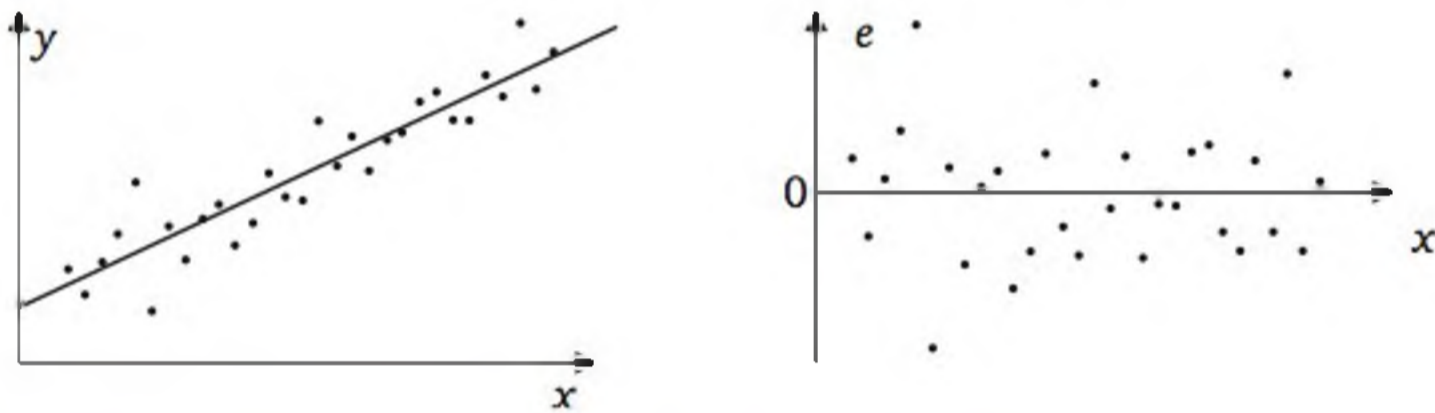


Figura 11.10 Gráficos indicando adequação do modelo.

Os gráficos da Figura 11.10 indicam uma situação onde as suposições do modelo estão aparentemente satisfeitas, pois os resíduos apresentam-se distribuídos de forma aleatória em torno da reta de regressão. No gráfico dos resíduos, a reta de regressão corresponde à linha horizontal sobre o valor zero.⁹ Já a Figura 11.11 apresenta uma situação onde existe um ponto discrepante. Esse ponto é visível nos dois gráficos, mas no gráfico dos resíduos ele aparece mais nitidamente.

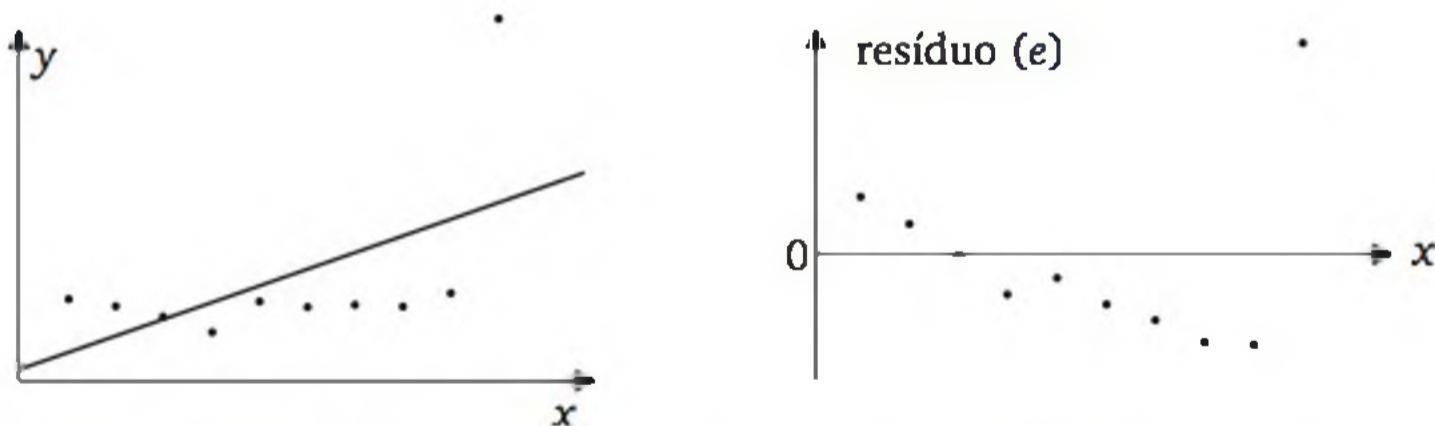


Figura 11.11 Gráficos indicando a presença de um valor discrepante.

⁹ O gráfico de probabilidade normal (Seção 6.5) também poderia ser aplicado aos resíduos para verificar a adequação da distribuição normal.

A Figura 11.11 mostra como um ponto discrepante pode *forçar* uma inclinação na reta, sugerindo uma tendência não compatível com as demais observações. Esse problema surge, principalmente, quando se tem uma amostra de observações pequena e o ponto discrepante estiver numa das extremidades do intervalo de observação de X . É prudente, nesse caso, buscar a razão da existência do ponto discrepante. Se sua causa for algum erro, alguma falha no experimento ou, ainda, puder ser considerada como uma situação atípica, devemos efetuar nova análise sem essa observação discrepante.

11.3.5 Transformações

Quando se trata de um estudo experimental, a variável X costuma ser estabelecida. Por exemplo, num estudo para verificar a relação entre o tempo de cozimento (X) e a maciez (Y) de um alimento, podemos estabelecer diferentes tempos de cozimento e verificar o resultado Y . Nos casos em que temos controle sobre a variável X , podemos variá-la uniformemente sobre o intervalo de estudo. Por exemplo, se pretendemos fazer a análise entre 20 e 30 minutos de cozimento, podemos fazer ensaios com os tempos de cozimentos 20, 21, 22, ... e 30 minutos.

Em estudos de levantamento, normalmente X e Y são observadas, sendo comum ocorrer uma distribuição assimétrica de valores de X . Por exemplo, considere o problema de se avaliar a relação entre *renda* (X) e *consumo* (Y). A maioria dos indivíduos tem renda baixa e, conseqüentemente, tende a consumir pouco, provocando distribuições assimétricas para X e Y . Nesta situação, os dados devem distribuir-se conforme mostra a Figura 11.12.

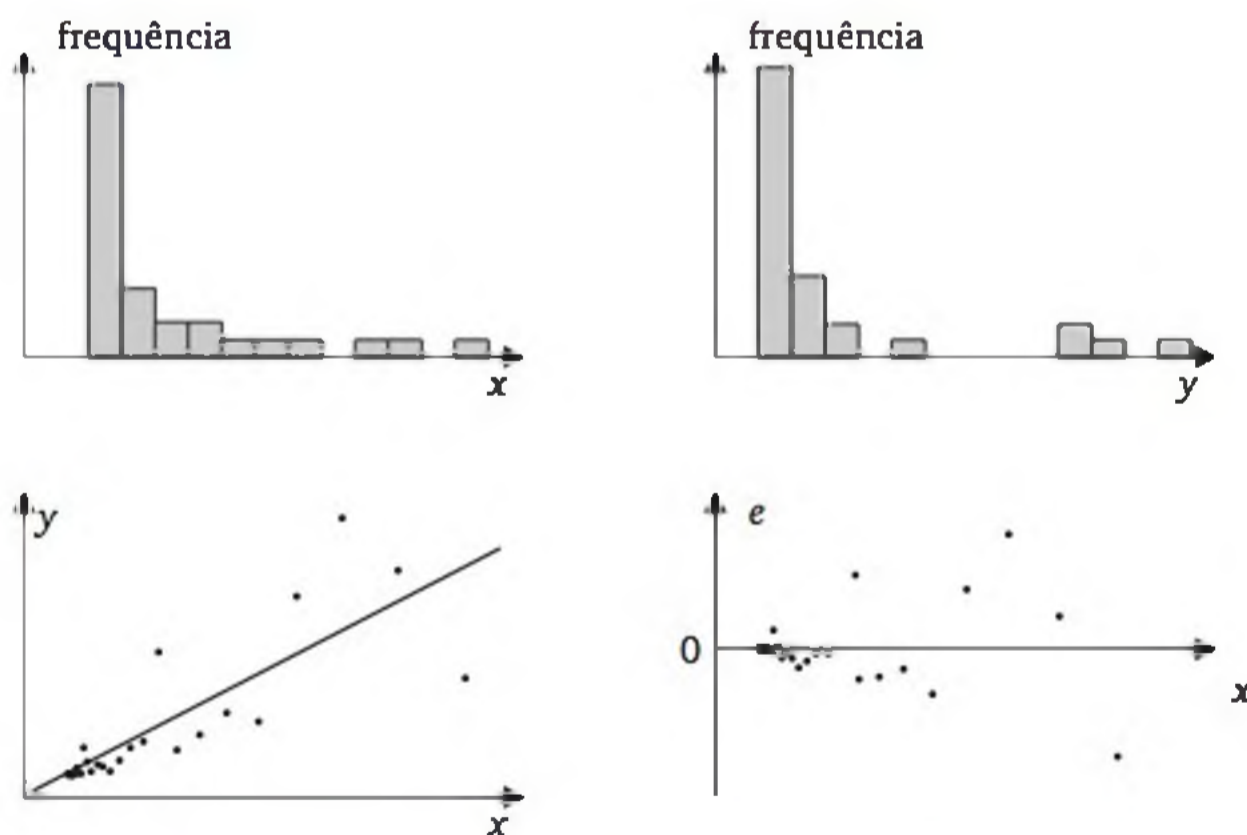


Figura 11.12 Gráficos indicando distribuições assimétricas de X e Y , além da variância de Y aumentar proporcionalmente com X .

Em situações como indicado na Figura 11.12, os valores grandes de X vão ter mais peso na determinação da inclinação da reta. Neste caso, recomendamos a aplicação da transformação logarítmica tanto nos valores de X como nos valores de Y , estabelecendo o seguinte modelo:¹⁰

$$\log(y_i) = \alpha + \beta \log(x_i) + \varepsilon_i \quad (11.34)$$

A transformação logarítmica aumenta as distâncias entre os valores pequenos e reduz as distâncias entre os valores grandes, tornando distribuições com assimetria positiva (cauda mais longa à direita) em distribuições aproximadamente simétricas. Com isso, temos uma situação mais adequada para estabelecer a reta de regressão. Em termos computacionais, devemos:

- a) calcular o logaritmo natural de cada valor x_i e de cada valor y_i ;
- b) aplicar a análise de regressão linear sobre os dados transformados $[\log(x_i), \log(y_i)]$; e
- c) construir novamente o gráfico de resíduos para verificar a adequação das suposições nesse novo modelo.

Na Equação 11.34, depois de estimados os parâmetros α e β por a e b , respectivamente, podemos aplicar a transformação inversa, resultando em

$$\hat{y} = e^{a + b \cdot \log(x)} \quad (11.35)$$

A Figura 11.13 apresenta uma situação que sugere relação *não linear*, com Y crescendo rapidamente para valores pequenos de X e crescendo lentamente para valores grandes de X . É uma situação típica onde se recomenda uma transformação logarítmica somente nos valores da variável X , ou seja, passamos a considerar o seguinte modelo para os dados:

$$y_i = \alpha + \beta \log(x_i) + \varepsilon_i \quad (11.36)$$

Note que o modelo (11.33) pode ser considerado linear em termos das variáveis $\log(x_i)$ e y_i (mas não mais entre x_i e y_i). Em termos computacionais, devemos:

- a) calcular o logaritmo de cada valor x_i ;
- b) aplicar a análise de regressão linear sobre os dados $(\log(x_i), y_i)$; e
- c) construir novamente o gráfico de resíduos para verificar a adequação das suposições nesse novo modelo.

10 É comum usar o logaritmo natural ou na base 10. Outra transformação que se presta ao mesmo propósito é a raiz quadrada. Esta segunda transformação é usada nas situações em que a inadequação do modelo não aparece de forma tão forte como visto na Figura 11.10. Observamos que estas transformações são possíveis somente quando todos os valores são positivos.

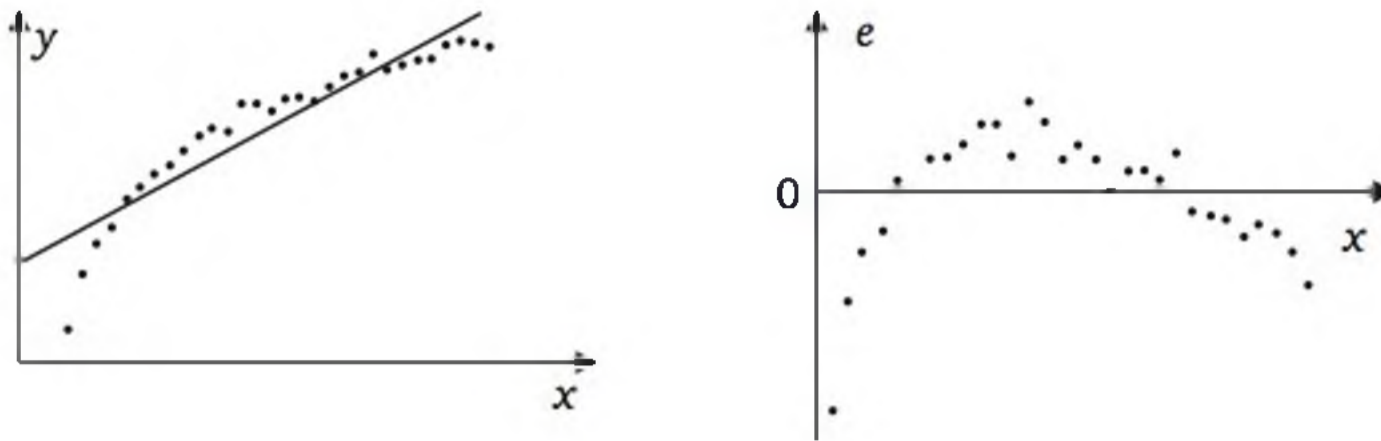


Figura 11.13 Gráficos indicando uma relação não linear, aparentemente logarítmica.

A Figura 11.14 sugere os seguintes problemas: (1) relação *não linear* e (2) aumento da variância à medida que X aumenta. Em casos assim, recomendamos uma transformação logarítmica nos valores da variável Y , ajustando o seguinte modelo aos dados:

$$\log(y_i) = \alpha + \beta x_i + \varepsilon_i \quad (11.37)$$

Ou seja,

- a) calcula-se o logaritmo de cada valor y_i ;
- b) aplica-se a análise de regressão linear sobre os dados $(x_i, \log(y_i))$; e
- c) refaz-se novamente o gráfico de resíduos para verificar se o novo modelo é mais adequado aos dados.

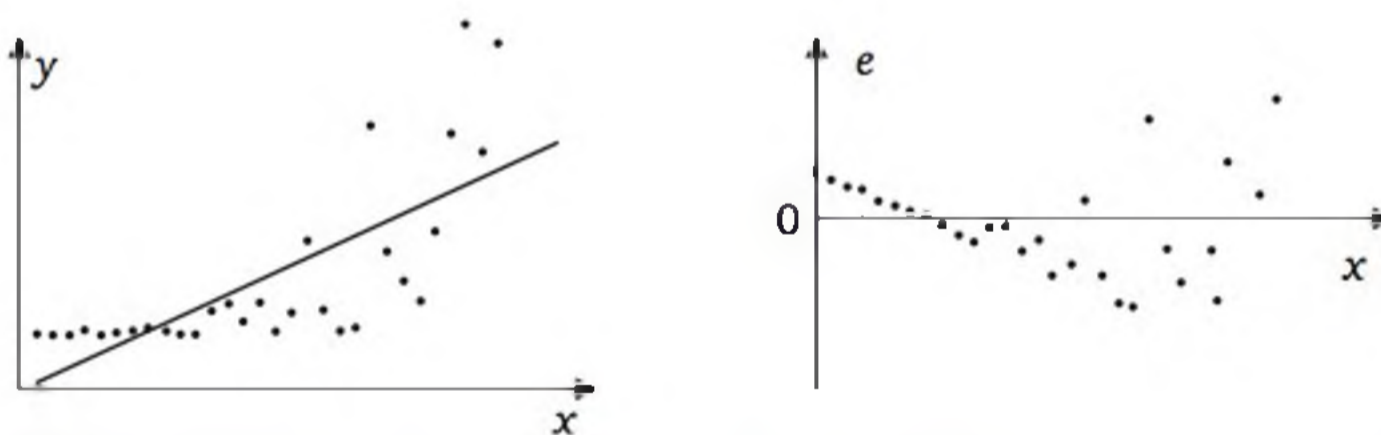


Figura 11.14 Gráficos indicando uma relação não linear – aparentemente exponencial – e variância não constante.

O uso de transformações auxilia o pesquisador a encontrar um modelo mais adequado para os dados, ainda que utilizando as expressões da regressão linear. A transformação logarítmica é muito usada por ter uma interpretação prática interessante, pois transforma variações percentuais de mesma magnitude em variações constantes. Por exemplo, se se considerar um aumento absoluto no salário de R\$ 100,00, seu significado vai ser muito diferente para quem

ganha R\$ 100,00 do que para quem ganha R\$ 1.000,00. Por isso, é mais comum se ouvir falar em aumentos percentuais de salários. Um aumento de 10% no salário representa um ganho de R\$ 10,00 para quem ganha R\$ 100,00 e um ganho de R\$ 100,00 para quem ganha R\$ 1.000,00. Na escala logarítmica, esses incrementos são iguais. Por essa razão, é muito comum usar a escala (ou transformação) logarítmica em variáveis econômicas ou medidas de valor e tamanho em geral.

Exemplo 11.3 (continuação) Na seção anterior, construímos uma regressão entre o preço de revenda de carros *seminovos* (Y) e o preço do correspondente modelo 0 km (X), considerando uma amostra de 142 automóveis apresentada no anexo deste capítulo. A Figura 11.15 apresenta o diagrama de dispersão e o gráfico dos resíduos desse modelo.

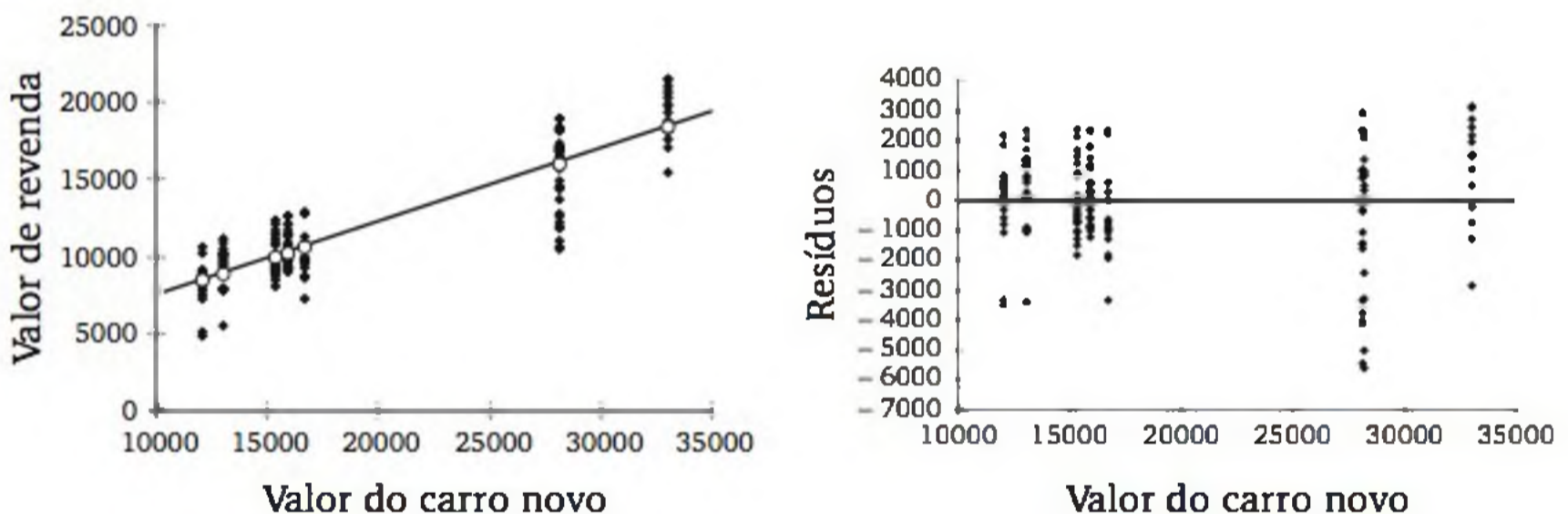


Figura 11.15 Gráfico de dispersão com o ajuste da reta de regressão e gráfico dos resíduos – Exemplo 11.3.

Observamos na Figura 11.15 que X só assume alguns determinados valores. Isto porque os automóveis em estudo são de 7 modelos e, para cada modelo, o preço 0 km é único. Por outro lado, não parece haver fortes violações nas suposições do modelo de regressão, a não ser a ocorrência maior de valores pequenos com respeito às duas variáveis, o que sugere tentarmos uma transformação logarítmica em X e em Y .

Realizamos a transformação logarítmica nos valores das duas variáveis e refizemos a análise de regressão, contudo o R^2 reduziu e o gráfico dos resíduos apontou uma distribuição assimétrica, com cauda mais longa à esquerda. Em função desses resultados, preferimos manter o modelo original. Na verdade, o preço de um carro *seminovo* depende de vários outros fatores, levando a um modelo de regressão múltipla, o qual discutiremos na próxima seção.

EXERCÍCIOS

5. Um administrador de uma grande sorveteria anotou por um longo período de tempo a *temperatura média diária*, em °C (X), e o *volume de vendas diárias de sorvete*, em kg (Y). Com os dados, foi ajustada a seguinte equação de regressão:

$$y = 0,5 + 1,8x, \text{ com } R^2 = 0,80$$

Pergunta-se:

- a) Qual é o consumo esperado de sorvete num dia de 27°C?
- b) Qual é o incremento esperado nas vendas de sorvete a cada 1°C de aumento da temperatura?
6. No processo de queima de massa cerâmica, avaliou-se o efeito da temperatura do forno (X) sobre a resistência mecânica da massa queimada (Y). Foram realizados 6 ensaios com níveis de temperatura equidistantes, os quais designaremos por 1, 2, 3, 4, 5 e 6. Os valores obtidos de resistência mecânica (MPa) foram: 41, 42, 50, 53, 54, 60, respectivamente. Pede-se:
- a) As estimativas de α e β da equação de regressão $E(Y) = \alpha + \beta x$.
- b) O coeficiente R^2 .
- c) O desvio padrão dos resíduos, s_e .
- d) O teste estatístico $H_0: \beta = 0$
7. A tabela a seguir relaciona os pesos (em centenas de kg) e as taxas de rendimento de combustível em rodovia (km/litro), numa amostra de 10 carros de passeio novos.

Peso	12	13	14	14	16	18	19	22	24	26
Rendimento	16	14	14	13	11	12	09	09	08	06

- a) Calcule o coeficiente de correlação de Pearson.
- b) Considerando o resultado do item (a), como você avalia o relacionamento entre peso e rendimento, na amostra observada?
- c) Para estabelecer uma equação de regressão, qual deve ser a variável dependente e qual deve ser a variável independente? Justifique a sua resposta.
- d) Estabeleça a equação de regressão, considerando a resposta do item (c).
- e) Apresente o diagrama de dispersão e a reta de regressão obtida em (d).

- f) Você considera adequado o ajuste do modelo de regressão do item (d)? Dê uma medida dessa adequação, interpretando-a.
- g) Qual é o rendimento esperado para um carro de 2.000 kg? Justifique sua resposta. Lembrete: os dados de peso na tabela estão em *centenas* de kg.
- h) Você considera seu estudo capaz de predizer o rendimento esperado de um veículo com peso de 7.000 kg? Justifique sua resposta.
8. Para verificar a viabilidade de incluir os resíduos da queima de carvão mineral na composição do cimento, foram feitos ensaios com cimento contendo de 0 a 9% de cinza de carvão; e medida a resistência à compressão (em MPa), após 28 dias. Os resultados foram os seguintes:

Carvão (%):	0	1	2	3	4	5	6	7	8	9
Resistência (MPa):	38,5	40,2	42,1	37,5	41,1	36,9	38,2	36,7	39,5	35,9

- a) Estabeleça a equação de regressão.
- b) Calcule R^2 .
- c) Teste se o coeficiente angular pode ser zero. Use $\alpha = 0,05$.
- d) Os resultados mostram evidência de que o uso de cinza de carvão mineral na composição do cimento diminui a sua resistência aos 28 dias?
9. Um estudo foi desenvolvido para verificar o quanto o comprimento de um cabo da porta serial de microcomputadores influencia na qualidade da transmissão de dados, medida através do número de falhas em 100.000 lotes de dados transmitidos (taxa de falha). Os resultados foram:

Comp. do cabo (m)	8	8	9	9	10	10	11	11	12
Taxa de falha	2,2	2,1	3,0	2,9	4,1	4,5	6,2	5,9	9,8

Comp. do cabo (m)	12	13	13	14	14	15
Taxa de falha	8,7	12,5	13,1	19,3	17,4	28,2

- a) Estabeleça a equação (reta) de regressão.
- b) Faça a análise dos resíduos e verifique se o modelo linear é adequado.
- c) Qual é a transformação sugerida pelo gráfico dos resíduos?
- d) Faça uma análise de regressão com os dados transformados.

10. A partir de um levantamento de 397 apartamentos em Criciúma-SC, realizou-se uma regressão entre área total (m²) e valor (\$).¹¹ Como essas variáveis apresentavam uma distribuição de frequências muito assimétricas à direita, foram feitas transformações logarítmicas em ambas as variáveis, antes de proceder à regressão. O modelo ajustado aos dados foi:

$$\log_{10}(\text{valor}) = 1,60 + (1,38) \log_{10}(\text{área}) \quad R^2 = 0,89$$

Interprete o valor de R^2 e o coeficiente 1,38.

11.4 INTRODUÇÃO À REGRESSÃO MÚLTIPLA

Em geral, uma variável dependente Y depende de várias variáveis independentes (X_1, X_2, \dots, X_k). Na análise de regressão múltipla, procuramos construir um modelo estatístico-matemático para se estudar, objetivamente, a relação entre as variáveis independentes e a variável dependente e, a partir do modelo, conhecer a influência de cada variável independente, como também, prever a variável dependente em função do conhecimento das variáveis independentes. Veja o Quadro 11.2.

Quadro 11.2 Exemplos de aplicação da regressão linear múltipla.

Variáveis independentes (X_1, X_2, \dots, X_k)	→	Variável dependente Y
X_1 = renda (R\$) X_2 = poupança (R\$) X_3 = taxa de juros (%)	→	Y = consumo (R\$)
X_1 = memória RAM (Gb) X_2 = sistema operacional X_3 = tipo de processador	→	Y = tempo de resposta do sistema computacional (segundos)
X_1 = área construída do imóvel (m ²) X_2 = padrão de qualidade (custo do m ² , R\$) X_3 = localização	→	Y = preço de um imóvel novo (R\$)
X_1 = valor do modelo novo (R\$) X_2 = quilometragem X_3 = idade do veículo (anos) X_4 = estado de conservação X_5 = opcionais	→	Y = valor de revenda de carro seminovo (R\$)

11 Os exercícios 10 e 12 baseiam-se em dados extraídos da dissertação de mestrado de Evelise C. Zancan, PPGEP/UFSC, 1995.

Para estabelecer o modelo clássico de regressão múltipla, suporemos que Y seja uma variável quantitativa contínua e X_1, X_2, \dots, X_k sejam variáveis quantitativas ou indicadoras de certos atributos. A variável indicadora deve ter valor 1, quando o atributo está presente; e 0, quando não está presente. Por exemplo, a variável $X_4 = \text{estado de conservação do veículo}$ pode ter valor 1 quando o veículo for considerado “bom” e 0 quando for considerado “ruim”.

Será suposto que Y é uma variável aleatória, isto é, somente será conhecida após a observação do elemento (indivíduo, carro etc.), enquanto X_1, X_2, \dots, X_k também podem provir de observações ou serem estabelecidas *a priori*. Especificados os valores x_1, x_2, \dots, x_k para as variáveis independentes, suporemos a seguinte equação para o valor esperado de Y :

$$E\{Y\} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k \quad (11.38)$$

A análise de regressão múltipla parte de um conjunto de observações $(x_{11}, x_{21}, \dots, x_{k1}, y_1), (x_{12}, x_{22}, \dots, x_{k2}, y_2), \dots, (x_{1n}, x_{2n}, \dots, x_{kn}, y_n)$ relativas às variáveis X_1, X_2, \dots, X_k e Y . Diremos que uma dada observação Y depende, em parte, dos correspondentes valores x_1, x_2, \dots, x_k e de uma infinidade de outros fatores, representados pelo termo de *erro*, ε . Mais especificamente, suporemos o seguinte modelo para as observações:

$$Y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \varepsilon_i \quad (i = 1, 2, \dots, n) \quad (11.39)$$

onde $\alpha, \beta_1, \beta_2, \dots$ e β_k são parâmetros a serem estimados com os dados e ε_i representa o *efeito (erro) aleatório* da i -ésima observação. As demais suposições são análogas à regressão simples, acrescentando a suposição de que as variáveis independentes X_1, X_2, \dots, X_k não devem ter correlações altas entre si.

Exemplo 11.5 Considerando os dados de 142 automóveis (anexo), vamos construir um modelo de regressão para tentar *explicar* $Y = \text{preço de revenda de automóveis seminovos}$ (em R\$), em função de:

- $X_1 = \text{preço do correspondente modelo 0 km}$ (em R\$);
- $X_2 = \text{tempo de uso}$ (em anos completos); e
- $X_3 = \text{quilometragem}$ (em milhares de km).

Usando a planilha *Excel*, obtivemos os seguintes resultados:

Estatística de regressão	
R múltiplo	0,961
R-Quadrado	0,923
R-quadrado ajustado	0,921
Erro padrão	1087
Observações	142

ANOVA

	gl	SQ	QM	F	Valor p
Regressão	3	1,95E + 09	6,51E+08	550,27	1,52E-76
Resíduo	138	1,63E + 08	1182186		
Total	141	2,11E+09			

	Coefi- cientes	Erro padrão	Estat t	Valor p	Inferior 95,0%	Superior 95,0%
Interseção	6240,13	352,11	17,722	2,25E-37	5543,89	6936,36
Valor novo	0,48	0,01	37,448	3,61E-74	0,45	0,50
Tempo uso	- 432,92	136,64	- 3,168	0,0019	- 703,10	- 162,75
Quilometragem	- 45,11	9,00	- 5,014	1,61E-06	- 62,90	- 27,32

Observamos, na primeira tabela, o valor de R^2 (R-quadrado) igual a 0,923. Este resultado indica que, na amostra observada, cerca de 92% da variação do preço de revenda podem ser *explicados* por uma relação linear que envolve o preço do automóvel 0 km (X_1), o tempo de uso (X_2) e a quilometragem (X_3). Um resultado razoavelmente maior do que os 79,1%, que foi obtido quando usamos apenas X_1 como variável independente (Exemplo 11.3).¹²

A segunda tabela (Anova) fornece o resultado estatístico da seguinte hipótese nula:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$$

Pela hipótese H_0 , o conjunto de variáveis independentes em estudo não tem poder de *explicação* sobre a variável dependente.¹³ A razão F resultou no valor $f = 550,27$, com correspondente valor $p = (1,52)10^{-76}$. Como o valor p é extremamente pequeno, o teste estatístico rejeita H_0 , indicando que as variáveis independentes escolhidas são significativas para *explicar* a variância de Y .

A terceira tabela fornece as estimativas dos coeficientes, incluindo intervalos de confiança e testes estatísticos para cada particular coeficiente. A primeira coluna apresenta as estimativas dos coeficientes. Assim, no presente exemplo, temos a seguinte equação:

12 O cálculo do R^2 na regressão múltipla é análogo ao da regressão simples.

13 Cabe observar que o teste estatístico refere-se à população. Quando se tem uma amostra muito pequena, pode-se obter um valor alto de R^2 e o teste aceitar $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$.

$$\hat{y} = 6240 + 0,48x_1 - 433x_2 - 45,1x_3$$

Com essa equação, tendo para um particular carro usado o preço de novo (x_1), o tempo de uso (x_2) e a quilometragem (x_3), podemos obter uma predição para o seu preço de revenda, \hat{y} . Por exemplo, um modelo, cujo preço do carro novo é R\$ 16.000,00, que tenha 2 anos de uso e 50 mil quilômetros rodados, seu preço de revenda, predito pelo modelo, é

$$\hat{y} = 6240 + (0,48) (16000) - (433)(2) - (45,1) (50) = 10779$$

ou seja, R\$ 10.779,00.

Com a equação de regressão, observamos, também, que a cada real de diferença no carro novo, esperamos uma diferença de 48 centavos de reais na revenda (mantendo-se constantes o tempo de uso e a quilometragem). A cada ano de envelhecimento do automóvel, esperamos R\$ 433,00 a menos na revenda (mantendo-se constantes o valor do carro novo e a quilometragem). E, também, a cada 1.000 quilômetros rodados, esperamos R\$ 45,11 a menos na revenda (mantendo-se constantes o valor de novo e o tempo de uso).¹⁴

A última tabela também fornece os resultados de testes estatísticos individuais, relativos a cada um dos coeficientes da equação de regressão. Ou seja, temos os resultados dos quatro seguintes testes:

1. $H_0: \alpha = 0$;
2. $H_0: \beta_1 = 0$;
3. $H_0: \beta_2 = 0$; e
4. $H_0: \beta_3 = 0$.

Como em todos os quatro casos, os valores p foram inferiores ao nível de significância usual de 0,05, rejeitamos as quatro hipóteses nulas, concluindo que nenhuma das variáveis independentes pode ser excluída do modelo.

Assim como na regressão simples, podemos calcular os *resíduos* para verificar a adequação do modelo de regressão. Calculamos, inicialmente, os valores preditos, \hat{y} , associados a cada conjunto de valores (x_1, x_2, \dots, x_k) usado na análise.

14 Dois comentários são pertinentes no momento:

a) É sabido que a desvalorização do automóvel não é linear com o tempo de uso. Uma transformação logarítmica em Y deve tornar o modelo mais realista.

b) As variáveis independentes, nessa aplicação, são correlacionadas. Por exemplo, um automóvel mais velho deve ter maior quilometragem. Logo, a interpretação “mantendo as demais variáveis constantes” pode ficar prejudicada. Além disso, os valores dos coeficientes das variáveis independentes correlacionadas não são bem estimados (observe a magnitude dos intervalos de confiança nas duas últimas colunas da terceira tabela).

se. No exemplo dos automóveis, os valores preditos são calculados pela equação $\hat{y} = 6240 + 0,48x_1 - 433x_2 - 45,1x_3$, com x_1 , x_2 e x_3 associados a cada um dos 142 automóveis avaliados. Os resíduos são obtidos através da diferença entre os valores observados e os valores preditos, isto é

$$e_i = y_i - \hat{y}_i \quad (i = 1, 2, \dots, n) \quad (11.40)$$

Os resíduos podem ser apresentados num diagrama de dispersão com cada variável independente ou com os valores preditos, os quais correspondem a uma combinação linear das variáveis independentes. A Figura 11.16 apresenta o diagrama de dispersão dos pares ordenados (\hat{y}_i, e_i) .

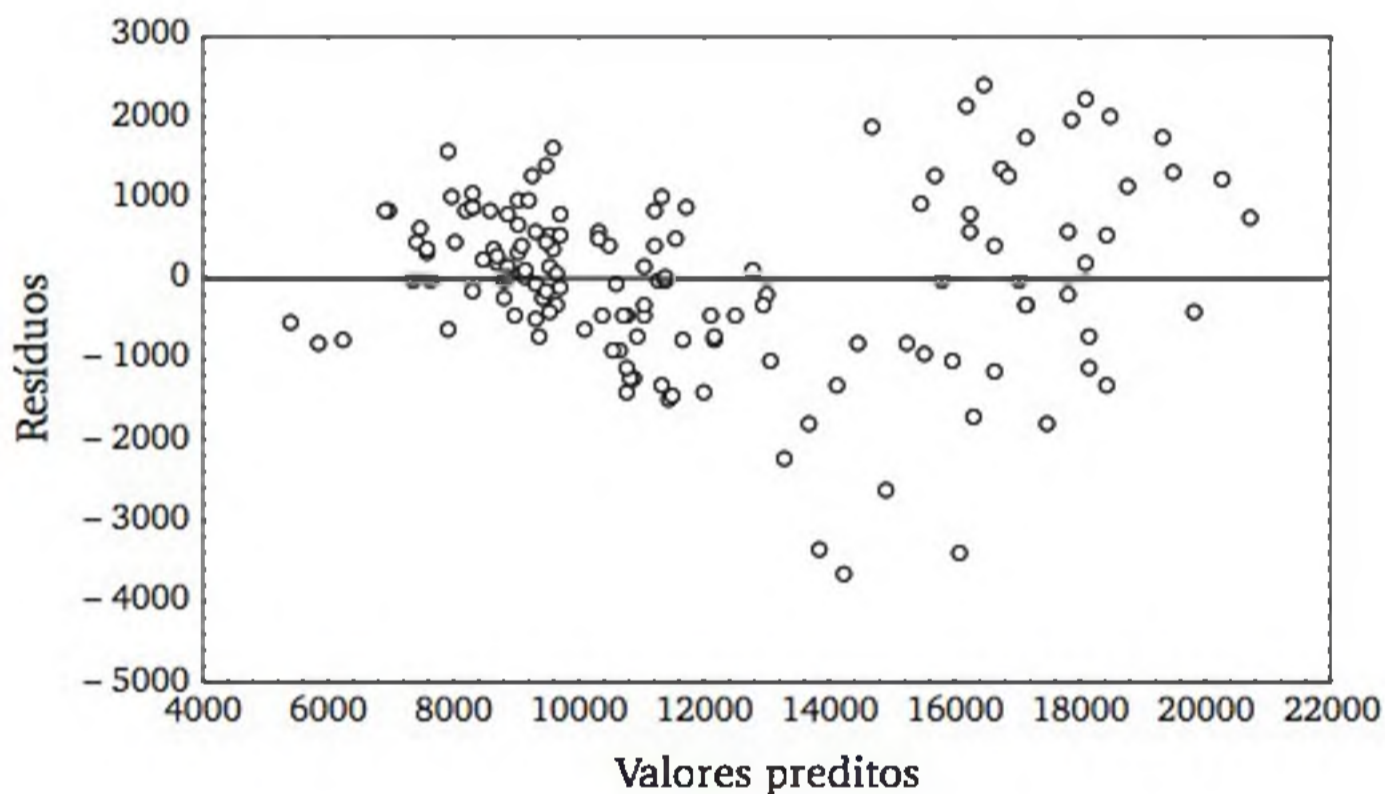


Figura 11.16 *Gráfico dos resíduos com os valores preditos.*

A análise do gráfico de resíduos (Figura 11.16) mostra um certo padrão. Para valores preditos pequenos, os resíduos tendem a ser positivos, depois eles tendem a ser negativos e, para valores preditos grandes, eles tendem a ser positivos de novo. Além disso, observamos que a dispersão aumenta para os valores preditos maiores. Conforme visto na seção anterior, essas características sugerem a aplicação de uma transformação logarítmica na variável dependente.

Raciocinando em termos da relação entre tempo de uso (X_2) e o valor do automóvel (Y), é mais natural considerar que a cada ano de uso, o automóvel tenha uma redução percentual do seu valor, reforçando a transformação sugerida pelo gráfico dos resíduos. A construção de um modelo mais adequado para esses dados é deixada para o leitor.

EXERCÍCIOS

- 11.** Com o auxílio de um computador, refaça o Exemplo 13.7, mas considerando como variável dependente o $\log(Y)$, onde Y = valor de revenda do automóvel. Observe o gráfico dos resíduos. Exclua as três observações que aparecem como discrepantes. Refaça novamente a análise.
- 12.** Considerando novamente os dados de 397 apartamentos em Criciúma, realizou-se uma regressão múltipla, onde $Y = \log_{10}(\text{valor em \$})$, $X_1 = \text{idade (em anos)}$, $X_2 = \text{gasto com energia elétrica dos moradores do apartamento (kw)}$, $X_3 = \log_{10}(\text{área em m}^2)$ e $X_4 = \text{local (1 = centro e bairros nobres, 0 = outros bairros)}$. A equação ajustada, com $R^2 = 0,92$, foi:

$$Y = 1,982 - (0,0073)X_1 - (0,0002)X_2 + (1,228)X_3 + (0,108)X_4$$

$(\mathbf{p} < 0,01)$ $(\mathbf{p} = 0,30)$ $(\mathbf{p} < 0,01)$ $(\mathbf{p} < 0,01)$

Os valores \mathbf{p} abaixo da equação estão associados aos testes de significância de cada coeficiente. Esses testes indicam que alguma variável pode ser excluída do modelo sem prejuízos? Interprete os coeficientes significativos.

ANEXO

Os dados que seguem foram coletados em 2001, pelo Prof. Manoel R. Lino, do Departamento de Informática e Estatística – UFSC, e fornecem informações sobre a venda de 142 automóveis *seminovos*, incluindo o modelo, o preço de revenda (R\$), o preço do modelo novo (R\$), o tempo de uso do automóvel (anos completos) e a quilometragem.

Auto	Modelo	Preço de rev.	Preço novo	Tempo de uso	Km	Auto	Modelo	Preço de rev.	Preço novo	Tempo de uso	Km
1	Mille	4.890	12.081	5	98	72	Gol	10.340	15.945	3	39
2	Mille	5.604	12.081	5	88	73	Gol	9.680	15.945	3	39
3	Mille	7.820	12.081	4	73	74	Gol	11.640	15.945	2	39
4	Mille	7.320	12.081	4	65	75	Gol	11.350	15.945	2	36
5	Mille	8.100	12.081	4	62	76	Gol	11.380	15.945	2	36
6	Mille	7.590	12.081	4	59	77	Gol	12.050	15.945	2	32
7	Mille	8.950	12.081	3	61	78	Gol	11.430	15.945	2	18

Auto	Modelo	Preço de rev.	Preço novo	Tempo de uso	Km	Auto	Modelo	Preço de rev.	Preço novo	Tempo de uso	Km
8	Mille	8.590	12.081	3	42	79	Gol	12.570	15.945	1	38
9	Mille	8.530	12.081	3	38	80	Gol	12.040	15.945	1	20
10	Mille	9.040	12.081	2	50	81	Gol	12.580	15.945	1	11
11	Mille	8.790	12.081	2	41	82	Fiorino	7.270	16.711	5	92
12	Mille	9.200	12.081	2	38	83	Fiorino	8.790	16.711	5	72
13	Mille	10.240	12.081	1	19	84	Fiorino	9.510	16.711	4	75
14	Mille	10.560	12.081	1	12	85	Fiorino	8.659	16.711	4	69
15	Fiesta	5.500	13.050	5	90	86	Fiorino	9.660	16.711	4	66
16	Fiesta	7.780	13.050	5	75	87	Fiorino	9.870	16.711	3	57
17	Fiesta	7.850	13.050	5	64	88	Fiorino	9.749	16.711	3	50
18	Fiesta	7.900	13.050	5	60	89	Fiorino	9.340	16.711	3	48
19	Fiesta	7.980	13.050	5	60	90	Fiorino	9.643	16.711	3	45
20	Fiesta	9.450	13.050	4	63	91	Fiorino	11.230	16.711	2	46
21	Fiesta	9.040	13.050	4	56	92	Fiorino	9.970	16.711	2	42
22	Fiesta	8.900	13.050	4	45	93	Fiorino	10.900	16.711	2	37
23	Fiesta	8.970	13.050	4	45	94	Fiorino	10.589	16.711	2	30
24	Fiesta	9.990	13.050	3	48	95	Fiorino	12.910	16.711	1	22
25	Fiesta	10.150	13.050	3	44	96	Fiorino	12.830	16.711	1	17
26	Fiesta	9.150	13.050	3	36	97	Parati	12.000	28.137	5	99
27	Fiesta	10.200	13.050	3	33	98	Parati	11.880	28.137	5	85
28	Fiesta	10.530	13.050	2	52	99	Parati	10.590	28.137	4	82
29	Fiesta	10.900	13.050	2	47	100	Parati	12.280	28.137	4	67
30	Fiesta	11.200	13.050	2	45	101	Parati	14.410	28.137	4	60
31	Fiesta	9.680	13.050	2	23	102	Parati	14.580	28.137	4	54
32	Fiesta	10.200	13.050	2	15	103	Parati	15.750	28.137	4	48
33	Fiesta	9.580	13.050	1	27	104	Parati	14.960	28.137	3	53
34	Fiesta	9.980	13.050	1	16	105	Parati	18.340	28.137	3	48
35	Fiesta	10.050	13.050	1	12	106	Parati	14.580	28.137	3	46

Auto	Modelo	Preço de rev.	Preço novo	Tempo de uso	Km	Auto	Modelo	Preço de rev.	Preço novo	Tempo de uso	Km
36	Corsa	8.450	15.337	5	75	107	Parati	17.200	28.137	3	39
37	Corsa	8.120	15.337	5	69	108	Parati	12.680	28.137	2	60
38	Corsa	8.680	15.337	5	65	109	Parati	17.020	28.137	2	39
39	Corsa	8.900	15.337	5	56	110	Parati	16.800	28.137	2	37
40	Corsa	9.200	15.337	4	78	111	Parati	16.800	28.137	2	37
41	Corsa	8.960	15.337	4	65	112	Parati	15.680	28.137	2	29
42	Corsa	9.350	15.337	4	62	113	Parati	18.360	28.137	1	32
43	Corsa	9.180	15.337	4	59	114	Parati	18.960	28.137	1	18
44	Corsa	9.260	15.337	4	59	115	Parati	17.090	28.137	1	18
45	Corsa	9.250	15.337	4	56	116	Escort	11.050	28.168	5	94
46	Corsa	9.680	15.337	3	75	117	Escort	10.480	28.168	5	82
47	Corsa	10.100	15.337	3	60	118	Escort	13.650	28.168	5	68
48	Corsa	9.950	15.337	3	59	119	Escort	12.800	28.168	4	85
49	Corsa	9.580	15.337	3	57	120	Escort	16.570	28.168	4	72
50	Corsa	9.460	15.337	3	48	121	Escort	16.400	28.168	4	55
51	Corsa	10.900	15.337	2	49	122	Escort	16.950	28.168	3	60
52	Corsa	11.200	15.337	2	36	123	Escort	16.860	28.168	3	47
53	Corsa	10.750	15.337	2	36	124	Escort	17.050	28.168	3	47
54	Corsa	12.050	15.337	2	33	125	Escort	18.120	28.168	2	44
55	Corsa	12.350	15.337	1	40	126	Escort	18.900	28.168	2	37
56	Corsa	11.640	15.337	1	23	127	Escort	18.280	28.168	1	26
57	Corsa	11.400	15.337	1	22	128	Escort	17.400	28.168	1	25
58	Gol	9.200	15.945	5	75	129	Vectra	18.830	32.995	5	75
59	Gol	9.340	15.945	5	75	130	Vectra	18.120	32.995	5	68
60	Gol	9.000	15.945	5	68	131	Vectra	15.490	32.995	4	80
61	Gol	9.340	15.945	5	45	132	Vectra	17.600	32.995	4	54
62	Gol	9.450	15.945	4	78	133	Vectra	17.050	32.995	4	47
63	Gol	9.680	15.945	4	69	134	Vectra	19.800	32.995	3	63

Auto	Modelo	Preço de rev.	Preço novo	Tempo de uso	Km	Auto	Modelo	Preço de rev.	Preço novo	Tempo de uso	Km
64	Gol	9.920	15.945	4	62	135	Vectra	20.300	32.995	3	58
65	Gol	0.320	15.945	4	59	136	Vectra	20.500	32.995	3	49
66	Gol	9.950	15.945	4	58	137	Vectra	19.880	32.995	3	43
67	Gol	9.680	15.945	4	55	138	Vectra	21.050	32.995	2	40
68	Gol	10.500	15.945	3	63	139	Vectra	20.810	32.995	2	36
69	Gol	10.860	15.945	3	50	140	Vectra	19.400	32.995	2	29
70	Gol	10.780	15.945	3	50	141	Vectra	21.500	32.995	1	28
71	Gol	10.560	15.945	3	43	142	Vectra	21.440	32.995	1	19

RESPOSTAS DE EXERCÍCIOS

CAPÍTULO 1

3. a) 6,5 b) 3,5 c) 1,87
4. As empacotadoras A e B aparentam estar calibradas, pois enchem os pacotes de café, *em média*, com o peso declarado; contudo a variância relativamente alta para a empacotadora B sugere que os pacotes de café provindos desta empacotadora têm pesos muito diferentes entre eles – falta de uniformidade no produto. A empacotadora C aparenta estar descalibrada.
5. Os valores são todos iguais.

CAPÍTULO 2

1. {José da Silva, Bartolomeu, Arnaldo, Geraldo, Joana, Aristóteles}
2. {S, B, L, C}
3. Não, basta extrair 100 números da tabela, com quatro algarismos, pertencentes ao conjunto {1.650, 1.651, ..., 8.840}, sem repetição.
9. $s_a^2 = 0,02687$ com 8 graus de liberdade.
11. a) $ef(A) = 13,65$, $ef(B) = 3,72$, $ef(C) = 14,72$, $ef(D) = 7,03$, $ef(E) = -0,15$ e $ef(F) = -2,41$.

Os fatores A, C e D são os que alteram mais o nível médio da resposta. Maior volume específico é obtido com A, C e D em seus níveis superiores.

b) $ef(A) = 6,80$, $ef(B) = 0,33$, $ef(C) = 2,77$, $ef(D) = -3,59$, $ef(E) = -2,07$
e $ef(F) = -3,78$.

O fator A provoca maior alteração na variabilidade da resposta. Menor variabilidade é obtida no nível inferior de A.

CAPÍTULO 3

3. a) 7,6 b) 2,37

4. 0,944 e 1,047

5. $m_d = 74,8$; $q_i = 74,3$; $q_s = 75,875$

8. d) C = 1: $\bar{x} = 5,54$ e $s = 0,50$;
C = 4: $\bar{x} = 9,92$ e $s = 0,73$;
C = 8: $\bar{x} = 4,14$ e $s = 0,21$.

f) $m_d = 9,4$; $q_i = 8,7$; $q_s = 12,55$

CAPÍTULO 4

1. a) $\Omega = \{\text{cara, coroa}\}$; b) $\Omega = \{0,1,2, \dots\}$; c) $\Omega = \{0, 1, 2, \dots\}$;

d) $\Omega = \{v, \text{ tal que } v \geq 0\}$;

e) $\Omega = \{t, \text{ tal que } -\infty \leq t \leq +\infty\}$. Poderia também desconsiderar resultados absurdos, como $\Omega = \{t, \text{ tal que } -5 \leq t \leq 40\}$.

2. a) $A = \{t, \text{ tal que } 5 < t \leq 10\}$; $B = \{t, \text{ tal que } t > 10\}$;
 $C = \{t, \text{ tal que } t > 8\}$; $D = \{t, \text{ tal que } t > 5\}$;
 $E = \phi$; $F = \{t, \text{ tal que } 8 < t \leq 10\}$;
 $G = \{t, \text{ tal que } t \leq 5 \text{ ou } t > 10\}$

3. a) $3/4$; b) $22/52$

4. 2,5%

5. a) $2/5$, $15/50$, $7/50$, $1/10$, $3/50$ b) $3/5$

6. a) $1/e$ b) $1 - 5/2e$
8. $1/1024$
9. $6.500/6.850$; $530/6.850$; $500/6.850$; $500/530$; $500/6.500$
10. a) $0,2095$ b) $0,1364$
11. a) $0,02875$ b) $0,5565$
12. $0,85$
13. a) $0,72$ b) $0,98$
14. $0,9639$
15. a) $0,54$ b) $0,04$ c) $0,42$
16. a)

HH	HM	MH	MM
$36/105$	$27/105$	$27/105$	$15/105$

 b) $51/105$ c) $36/51$
17. a)

P	A
$3/7$	$4/7$

 b)

PP	PA	AP	AA
$9/49$	$12/49$	$12/49$	$16/49$

 c) $29/161$ $40/161$ $40/161$ $52/161$ d) $0,01157$ e) $0,2593$
18. a) $9/25$ b) $17/60$ c) $4/17$
19. a) $0,7$ b) $0,94$ c) $0,8333$ d) $0,875$
20. $0,9990$
21. $0,4865$

CAPÍTULO 5

1. a)

x:	0	1
p(x):	$1/2$	$1/2$

 b)

x:	0	1	2
p(x):	$1/4$	$1/2$	$1/4$

 c)

x:	0	1	2
p(x):	$0,36$	$0,48$	$0,16$

 d)

x:	0	1	2	3
p(x):	$0,216$	$0,432$	$0,288$	$0,064$

$$3. \quad F(x) = \begin{cases} 0 & \text{se } x < 0 \\ 0,216 & \text{se } 0 \leq x < 1 \\ 0,648 & \text{se } 1 \leq x < 2 \\ 0,936 & \text{se } 2 \leq x < 3 \\ 1 & \text{se } x \geq 3 \end{cases}$$

4. a) 0,5 e 0,25 b) 1 e 0,5 c) 0,8 e 0,48 d) 1,2 e 0,72

5. a) 4 e 9,6 b) 5 e 9,6 c) 8 e 38,4

6. 1.000 g e 10,7703 g

7. a) 0,6415 b) 0,1887 c) 0,0754 d) 1 e) 19

9. 0,0640

10. 0,5 e 0,3879

11. a) 0,0843 b) 0,2052

12. 0,0527

13. 150 e 27,39

14. 0,0702

15. b) 0,6415 c) 0,1886 d) 1 erro

16. proposta 1

17. a) 0,737 b) 0,337

18. a) 0,9990 b) 0,9989

19. 0,0137

20. 0,6321

21. a) 0,6723 b) 0,7183

22. a) 0,7408 b) 0,0369 c) 0,6135

23. Alternativa 2 ($\mu = \text{R\$ } 111,80$)

24. a) 0,0498;

b) Proposta: lucro esperado R\$ 57,47; por categoria: R\$ 76,00 (não);

c) 1064,85

25. a) 0,8009 b) 0,8428

CAPÍTULO 6

1. a) $f(x) = \begin{cases} 1, & \text{para } 0 \leq x \leq 1 \\ 0, & \text{para } x \notin [0,1] \end{cases}$ b) $F(x) = \begin{cases} 0, & \text{para } x < 0 \\ x, & \text{para } 0 \leq x < 1 \\ 1, & \text{para } x \geq 1 \end{cases}$
- c) $\frac{1}{2}$ e $\frac{1}{12}$
2. $f(x) = \begin{cases} 1/4, & \text{para } 20 \leq x \leq 24 \\ 0, & \text{para } x \notin [20,24] \end{cases}$ a) $\frac{1}{4}$ b) 22 c) $\frac{4}{3}$
3. $f(x) = \begin{cases} \frac{x}{4} - 5, & \text{para } 20 \leq x < 22 \\ 6 - \frac{x}{4}, & \text{para } 22 \leq x < 24 \\ 0, & \text{para } x \notin [20,24] \end{cases}$ a) $\frac{1}{8}$ b) 22 c) $\frac{2}{3}$
4. $f(x) = \begin{cases} e^{-x} & \text{para } x \geq 0 \\ 0, & \text{para } x < 0 \end{cases}$
5. a) 1 b) $\frac{1}{2}$ c) $\frac{59}{72}$ d) 1 e) $\frac{1}{6}$
6. a) 0,3679 b) 0,4135 c) 0,3679
8. a) 0,0495 b) 0,9505 c) 0,6826 d) 0,955 e) 0,9974
 f) 0,000 g) 1,65 h) 2,58
9. a) 0,6915 b) 0,7333
10. a) 0,0314 b) 0,9372
11. a) 0,1052 b) 0,2033
12. 0,1056
13. 0,4724
14. 0,6321
15. 2.877 horas
17. a) 0,0071 b) 0,15 (usando aproximação normal)
18. a) 0,8472 b) 0,3859

19. a) 0,0781 b) ≈ 0
 20. 0,0202
 21. 27,63 segundos
 22. a) 0,8858 b) 0,153 mm
 23. a) 0,1335 b) 0,6632 c) 0,1541 d) 5.504 kg/cm²
 24. M2, pois $E(M1) = R\$ 83,64$ e $E(M2) = R\$ 163,30$.

CAPÍTULO 7

1.

\hat{p}	$p(\hat{p})$
0	$\frac{1}{2}$
$\frac{1}{2}$	$\frac{1}{2}$

2. \bar{x} : 3,80 3,85 3,90 3,95 4,00 4,05 4,10
 $p(\bar{x})$: $\frac{1}{16}$ $\frac{2}{16}$ $\frac{3}{16}$ $\frac{4}{16}$ $\frac{3}{16}$ $\frac{2}{16}$ $\frac{1}{16}$ média = 3,95 var. = 0,00625
4. a) 0,5 b) 0,3174 c) 0,0456 d) 0,95 e) 0,03
5. 0,0722 (usando a correção de continuidade)
6. 0,0694 (usando a correção de continuidade)
7. a) $\frac{1}{2}$ e $\frac{1}{12}$ b) 0,06 c) $\frac{1}{2}$ e $\frac{1}{1200}$
 d) aprox. normal e) 0,7016
8. a) 0,32 b) 0,0037
10. 55% \pm 6,9%
11. 3,00% \pm 1,36%
12. 98,0 mm \pm 1,8 mm
13. 28
14. 11
15. 12

16. a) 865 b) 1.494
17. $35,9 \pm 11,2$
18. a) $14,00 \pm 0,98$ b) 184
19. a) 102 b) $5,30 \pm 0,45$
- c) Não, pois o intervalo onde deve estar a verdadeira média abrange também valores menores que cinco.
- d) $68,6\% \pm 7,4\%$
20. 77
21. 176
22. 2.500 (adotando $z \approx 2$)
23. a) 15,443 e 2,074 b) $15,44 \pm 1,20$ c) 157
24. a) $95\% \pm 2,1\%$ b) 12.298

CAPÍTULO 8

1. a) 0,0062 b) 0,3874 c) 0,0062
2. a) Rejeita H_0 b) Aceita H_0 c) Rejeita H_0
3. Rejeita H_0 se ocorrer 0, 1, 2, 13, 14 ou 15 caras. Caso contrário, aceita H_0 .
4. É possível. Por exemplo, se no teste para verificar se uma moeda é honesta ocorrer $Y = 2$ caras em $n = 12$ lançamentos, temos $\mathbf{p} = 0,0384$, que rejeita H_0 ao nível de 5%, mas aceita ao nível de 1%. O inverso nunca acontece.
5. a) Decide-se por H_1 , pois o valor \mathbf{p} é menor do que o nível de significância adotado. Dada a evidência da amostra, o risco dele estar tomando a decisão errada é de 0,0001.
- b) Decide-se por H_0 , pois o valor \mathbf{p} é maior do que o nível de significância adotado. Quando se aceita H_0 , o valor \mathbf{p} não oferece qualquer informação sobre o risco de se estar tomando a decisão errada.
- c) Quanto menor o valor \mathbf{p} , existe maior evidência para a rejeição de H_0 (e consequente aceitação de H_1).
6. a) bilateral b) unilateral c) unilateral d) unilateral

7. a) H_0 : Em média, a produtividade com treinamento é igual do que a produtividade sem treinamento. H_1 : Em média, a produtividade com treinamento é maior do que a produtividade sem treinamento (teste unilateral).
- b) H_0 : Em média, a velocidade é igual ao valor anunciado. H_1 : Em média, a velocidade é menor do que o valor anunciado (teste unilateral).
- c) H_0 : As produtividades médias são iguais para os dois métodos de treinamento. H_1 : As produtividades médias são diferentes para os dois métodos de treinamento (teste bilateral).
8. a) 0,0031 b) 0,1937 c) 0,6127
9. Rejeita H_0 se ocorrer 10 ou 11 caras. Caso contrário, aceita H_0 .
10. Hipóteses: $H_0: p = 0,5$ e $H_1: p > 0,5$ (p = probabilidade de a criança acertar uma dada questão). Decisão: rejeita H_0 , isto é, há evidência de que a criança tem algum conhecimento sobre o assunto ($\mathbf{p} = 0,0031$).
11. a) $H_0: p = 0,25$ e $H_1: p > 0,25$; b) $\mu = 3$ c) $\mathbf{p} = 0,1576$
- d) Aceita H_0 . **Não** há evidência de que a criança tem algum conhecimento sobre o assunto.
12. a) 0,0094 b) 0,3844 c) 0,0094
14. a) Aceita H_0 : a moeda é honesta ($\mathbf{p} = 0,2892$).
- b) Rejeita H_0 , isto é, decide-se que a moeda é viciada ($\mathbf{p} \approx 0,0000068$, uso da aproximação normal).
15. Decisão: rejeita H_0 , isto é, há evidência de que o sistema “inteligente” adquiriu algum conhecimento sobre o assunto ($\mathbf{p} = 0,0071$, uso da aproximação normal).
16. Não, pois $z = 1,11 \Rightarrow \mathbf{p} = 0,267 \Rightarrow$ aceita H_0 .
17. Não, pois um teste unilateral aceita $H_0: p = 0,90$ ($z = -1,18$, $\mathbf{p} = 0,119$)
18. Sim ($z = -3,13$; $\mathbf{p} < 0,00135$)
19. Sim ($t = 5,59$; $\mathbf{p} < 0,0005$)
20. Sim ($q^2 = 12,89$; $\mathbf{p} < 0,01$)
21. Não ($q^2 = 21,3$; $\mathbf{p} > 0,20$)
22. $n = 53$
23. $n = 18$ (mais 10 unidades)

24. a) 71,06 e 7,49; b) $H_0: \mu = 70$ e $H_1: \mu > 70$
 c) Não ($t = 0,57$; $\mathbf{p} > 0,25$)
 d) Aceitar H_0 quando falsa e) 0,564
25. a) 498,94 e 4,07 b) Não ($t = -1,04$; $0,10 < \mathbf{p} < 0,25$)
 c) Sim ($q^2 = 24,85$; $0,05 < \mathbf{p} < 0,10$)
26. a) Não ($z = -0,968$; $\mathbf{p} = 0,166$) b) Não, $n = 138$ c) 0,751
27. a) Não ($z = -0,86$; $\mathbf{p} = 0,1949$) b) 0,74 c) 9.450
28. a) 20; b) Não ($t = -0,596$; $\mathbf{p} > 0,25$)
 c) Sim ($q^2 = 42,75$; $\mathbf{p} < 0,0025$)

CAPÍTULO 9

1. Não. Usando teste t unilateral para amostras independentes: $t = 1,51$ ($0,05 < \mathbf{p} < 0,10$)
2. Sim. Usando teste t unilateral para dados pareados: $t = 3,10$ ($0,01 < \mathbf{p} < 0,025$)
3. b) Rejeita H_0 ao nível de 5%, pois $t = 2,70 \implies 0,01 < \mathbf{p} < 0,025$ (teste unilateral).
4. a) Rejeita H_0 ao nível de 5%, pois, $t = 3,04 \implies 0,005 < \mathbf{p} < 0,010$ (teste unilateral).
5. Não, $t = -0,36$, $\mathbf{p} > 0,25$ (teste bilateral).
6. Há diferença significativa ($t = 294$, $\mathbf{p} < 0,01$).
7. Sim. Teste t bilateral para amostras independentes: $t = -4,40$ ($\mathbf{p} < 0,001$).
8. Sim ($t = 5,175$; $0,002 < \mathbf{p} < 0,005$).
9. $n \approx 22$.
10. a) Rejeita a hipótese de igualdade entre os circuitos ($f = 11,27 > f_c = 3,89$).

11. Não ($f = 3,08 < f_c = 7,15$, então aceita a hipótese nula de igualdade entre as variâncias).

12. Sim ($f = 28,40 > f_c = 4,76$).

13. a)

Fonte de variação	SQ	gl	QM	f
Memória <i>cache</i>	1026675	1	1026675	1207,9
Memória principal	1944075	1	1944075	2287,1
Interações	151875	1	151875	178,7
Erro	6800	8	850	
Total	3129425	11		

b) Todos (memória *cache*, memória principal e a interação entre elas).

c) Não, devido a presença de interação.

14. a)

Fonte de variação	SQ	gl	QM	f	fc
Processador	1028	2	514	60,76	3,89
Tipo de carga	18	3	6	0,71	3,49
Interação	286	6	47,67	5,63	3,00
Erro	102	12	8,50		
Total	1934	23			

A carga de trabalho e a interação são significativas.

15. a) Sim ($t = 6,30$; $\mathbf{p} < 0,0005$) b) Não ($q^2 = 11,16$; $\mathbf{p} > 0,10$)

16. a) Sim ($f = 10,72$; $\mathbf{p} < 0,01$) b) Sim ($f = 52,17$; $\mathbf{p} < 0,01$)

17. Somente parâmetro é significativo, pois:

Ruído: $f = 0,389$; $\mathbf{p} > 0,25$; Parâmetro: $f = 8,469$; $\mathbf{p} < 0,01$;

Interação: $f = 0,043$; $\mathbf{p} > 0,25$.

18.

Fator	Efeito	f
A	0,34	0,019
B	- 1,94	0,613
C	4,58	3,419
D	- 1,23	0,245

$f_c = 2,96$, então somente o fator C apresenta efeito significativo a 10%.

CAPÍTULO 10

1. Aceita H_0 . Pedido do laboratório A não é corroborado pelos dados.
2. Rejeita H_0 . A suspeita dos engenheiros tem fundamento.
3. Rejeita H_0 no teste de aderência à normal, e aceita H_0 no teste de aderência à exponencial. O sócio está certo.
4. Rejeita H_0 no teste de aderência à normal. Não é possível utilizar os gráficos de controle de médias de Shewhart para monitorar o processo.
5. a) Aceita H_0 . Não há evidência suficiente para considerar que a dimensão não siga uma distribuição normal.
b) Rejeita H_0 . Há evidência suficiente para considerar que a dimensão não segue uma distribuição normal com média 15 e desvio padrão 0,4.
6. Rejeita H_0 . Os percentuais não podem ser considerados iguais.
7. Aceita H_0 . Não há evidência suficiente para considerar os percentuais diferentes.
8. Rejeita H_0 . A suspeita tem fundamento, há evidência suficiente para sugerir que os percentuais são diferentes.
9. a) Aceita H_0 . Não há evidência suficiente de que os tempos com o novo terminal sejam menores do que aqueles obtidos com o atual.
b) Rejeita H_0 . O novo terminal possibilitou tempos de atendimento menores, há evidência suficiente de que os tempos são menores.
c) Rejeita H_0 , resultado do item (b). Porque o teste dos sinais por postos é mais sensível (tem maior poder de detectar a falsidade de H_0 quando

ela é realmente falsa), pois leva em conta o sinal e a magnitude das diferenças entre os grupos.

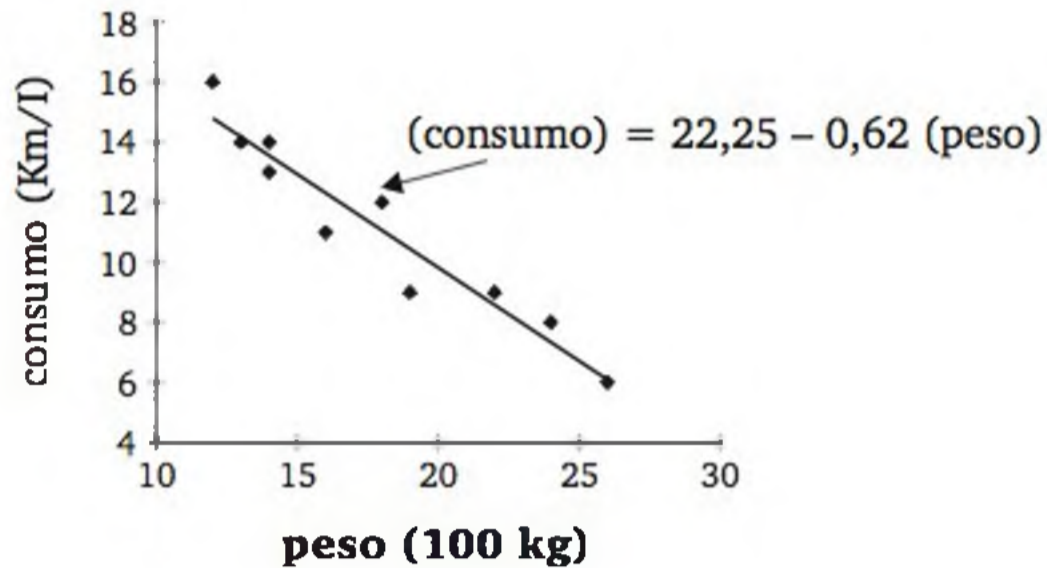
10. a) Rejeita H_0 . O número de pessoas na fila com o novo terminal é menor do que aquele obtido com o atual. Há evidência estatística suficiente de que o número de pessoas é menor.
b) Aceita H_0 . Não há evidência suficiente de que o número de pessoas na fila com o novo terminal seja menor do que aquele obtido com o atual.
11. a) Aceita H_0 . Não há evidência suficiente de que as taxas de transmissão de dados com o novo cabeamento sejam maiores do que aquelas obtidas com o atual.
b) Rejeita H_0 . O novo cabeamento possibilitou taxas de transmissão maiores, há evidência suficiente de que os taxas de transmissão são maiores.
12. a) Aceita H_0 . Não há evidência suficiente de que os dois métodos produzam resultados diferentes.
b) Aceita H_0 . Não há evidência suficiente de que os dois métodos produzam resultados diferentes.
c) Os métodos não produzem resultados diferentes, pois tanto teste dos sinais quanto o dos sinais por postos aceitaram H_0 .
13. Aceita H_0 . Não há evidência suficiente de que as arquiteturas causem diferenças nos tempos de processamento.
14. Rejeita H_0 . A suspeita é confirmada. Há evidência suficiente para considerar o concreto Y mais resistente à compressão do que o X .
15. Aceita H_0 . Não há evidência suficiente de que os tempos de rompimento do novo modelo de elo fusível sejam menores, não se deve cogitar sua aquisição.
16. Rejeita H_0 . A espécie 1 degrada a lignina mais rapidamente. Há evidência suficiente para considerar que a espécie 1 degrada a lignina mais rápido do que a espécie 2.
17. a) Aceita H_0 . Não há evidência suficiente para considerar que o consumo de energia elétrica não siga uma distribuição normal com média de 85 kWh e desvio padrão de 15 kWh.
b) Aceita H_0 . Não há evidência suficiente para considerar que o consumo de energia elétrica não siga uma distribuição normal com média de 85 kWh e desvio padrão de 15 kWh.

18. Aceita H_0 . Não há evidência suficiente de que haja associação entre os percentuais de defeitos e as fábricas onde as peças foram produzidas.
19. Rejeita H_0 . A concentração de desinfetante não é homogênea nas três filiais. Há evidência suficiente para considerar que a concentração difere dependendo da filial que produz o produto de limpeza.
20. Rejeita H_0 . As mudanças no programa da Qualidade foram aprovadas pelos funcionários. Há evidência suficiente de que aqueles que desaprovavam antes agora aprovam.
21. Aceita H_0 . Não há evidência suficiente de que haja diferença entre os índices de alunos e alunas.
22. Aceita H_0 . Não há evidência suficiente de que o número de defeituosos diminuiu, ou seja, que o programa da Qualidade teve resultado.
23. Aceita H_0 . Não há evidência suficiente para considerar que as médias das durações dos pneus vendidos pelos dois fabricantes sejam diferentes.

CAPÍTULO 11

1. 0,64 (Não, a correlação não é significativamente diferente de zero.)
3. a) 0,69 c) 0,86
d) Correlação é positiva forte e significativamente diferente de zero.
5. a) 49,1 kg b) 1,8 kg
6. a) $a = 36,6$ e $b = 3,83$ b) $R^2 = 0,95$ c) $s_e = 1,836$
d) Rejeita H_0 , pois $f = 76,1$ ($\mathbf{p} < 0,0001$) ou $t = 8,72$ ($\mathbf{p} < 0,0001$).
7. a) $r = -0,96$ b) Correlação negativa forte.
c) Variável dependente: consumo; e variável independente: peso.
d) $(\text{consumo}) = 22,25 - 0,62 (\text{peso})$.

e)



- f) Sim, verifica-se pelo gráfico do item (e) que uma relação linear parece adequar-se bem ao presente problema. Além disso, tem-se um coeficiente de determinação próximo de 1 ($R^2 = 0,92$)
- g) 9,85 km/l
- h) Não, pois os veículos estudados estavam na faixa de 1200 a 2600 kg e, portanto, a equação de regressão deve ser usada apenas nesta faixa.
8. a) $\hat{y} = 40,22 - (0,35)x$ b) 0,27
- c) Aceita $H_0: \beta = 0$ (Teste bilateral: $t = -1,72$; $0,10 < p < 0,20$).
- d) Não.
9. c) Transformação logarítmica em Y.
10. Em termos do modelo estabelecido, 89% da variância de $\log(\text{valor})$ pode ser explicada pelos diferentes valores de $\log(\text{área})$. A cada unidade a mais no $\log(\text{área})$, sendo área medida em m^2 , o $\log(\text{valor})$ aumenta 1,38 unidade, sendo valor medido em \$.

APÊNDICE: Tabelas Estatísticas

Tabela 1 *Distribuição binomial: probabilidade de cada valor x em função de n e p.*

n	x	p																		
		0,05	0,1	0,15	0,2	0,25	0,3	0,35	0,4	0,45	0,5	0,55	0,6	0,65	0,7	0,75	0,8	0,85	0,9	0,95
1	0	0,9500	0,9000	0,8500	0,8000	0,7500	0,7000	0,6500	0,6000	0,5500	0,5000	0,4500	0,4000	0,3500	0,3000	0,2500	0,2000	0,1500	0,1000	0,0500
	1	0,0500	0,1000	0,1500	0,2000	0,2500	0,3000	0,3500	0,4000	0,4500	0,5000	0,5500	0,6000	0,6500	0,7000	0,7500	0,8000	0,8500	0,9000	0,9500
2	0	0,9025	0,8100	0,7225	0,6400	0,5625	0,4900	0,4225	0,3600	0,3025	0,2500	0,2025	0,1600	0,1225	0,0900	0,0625	0,0400	0,0225	0,0100	0,0025
	1	0,0950	0,1800	0,2550	0,3200	0,3750	0,4200	0,4550	0,4800	0,4950	0,5000	0,4950	0,4800	0,4550	0,4200	0,3750	0,3200	0,2550	0,1800	0,0950
	2	0,0025	0,0100	0,0225	0,0400	0,0625	0,0900	0,1225	0,1600	0,2025	0,2500	0,3025	0,3600	0,4225	0,4900	0,5625	0,6400	0,7225	0,8100	0,9025
3	0	0,8574	0,7290	0,6141	0,5120	0,4219	0,3430	0,2746	0,2160	0,1664	0,1250	0,0911	0,0640	0,0429	0,0270	0,0156	0,0080	0,0034	0,0010	0,0001
	1	0,1354	0,2430	0,3251	0,3840	0,4219	0,4410	0,4436	0,4320	0,4084	0,3750	0,3341	0,2880	0,2389	0,1890	0,1406	0,0960	0,0574	0,0270	0,0071
	2	0,0071	0,0270	0,0574	0,0960	0,1406	0,1890	0,2389	0,2880	0,3341	0,3750	0,4084	0,4320	0,4436	0,4410	0,4219	0,3840	0,3251	0,2430	0,1354
	3	0,0001	0,0010	0,0034	0,0080	0,0156	0,0270	0,0429	0,0640	0,0911	0,1250	0,1664	0,2160	0,2746	0,3430	0,4219	0,5120	0,6141	0,7290	0,8574
4	0	0,8145	0,6561	0,5220	0,4096	0,3164	0,2401	0,1785	0,1296	0,0915	0,0625	0,0410	0,0256	0,0150	0,0081	0,0039	0,0016	0,0005	0,0001	0,0000
	1	0,1715	0,2916	0,3685	0,4096	0,4219	0,4116	0,3845	0,3456	0,2995	0,2500	0,2005	0,1536	0,1115	0,0756	0,0469	0,0256	0,0115	0,0036	0,0005
	2	0,0135	0,0486	0,0975	0,1536	0,2109	0,2646	0,3105	0,3456	0,3675	0,3750	0,3675	0,3456	0,3105	0,2646	0,2109	0,1536	0,0975	0,0486	0,0135
	3	0,0005	0,0036	0,0115	0,0256	0,0469	0,0756	0,1115	0,1536	0,2005	0,2500	0,2995	0,3456	0,3845	0,4116	0,4219	0,4096	0,3685	0,2916	0,1715
	4	0,0000	0,0001	0,0005	0,0016	0,0039	0,0081	0,0150	0,0256	0,0410	0,0625	0,0915	0,1296	0,1785	0,2401	0,3164	0,4096	0,5220	0,6561	0,8145
5	0	0,7738	0,5905	0,4437	0,3277	0,2373	0,1681	0,1160	0,0778	0,0503	0,0313	0,0185	0,0102	0,0053	0,0024	0,0010	0,0003	0,0001	0,0000	0,0000
	1	0,2036	0,3281	0,3915	0,4096	0,3955	0,3602	0,3124	0,2592	0,2059	0,1563	0,1128	0,0768	0,0488	0,0284	0,0146	0,0064	0,0022	0,0005	0,0000

Tabela 1 (Continuação).

<i>n</i>	<i>x</i>	<i>p</i>																		
		0,05	0,1	0,15	0,2	0,25	0,3	0,35	0,4	0,45	0,5	0,55	0,6	0,65	0,7	0,75	0,8	0,85	0,9	0,95
2	2	0,0214	0,0729	0,1382	0,2048	0,2637	0,3087	0,3364	0,3456	0,3369	0,3125	0,2757	0,2304	0,1811	0,1323	0,0879	0,0512	0,0244	0,0081	0,0011
	3	0,0011	0,0081	0,0244	0,0512	0,0879	0,1323	0,1811	0,2304	0,2757	0,3125	0,3369	0,3456	0,3364	0,3087	0,2637	0,2048	0,1382	0,0729	0,0214
	4	0,0000	0,0005	0,0022	0,0064	0,0146	0,0284	0,0488	0,0768	0,1128	0,1563	0,2059	0,2592	0,3124	0,3602	0,3955	0,4096	0,3915	0,3281	0,2036
	5	0,0000	0,0000	0,0001	0,0003	0,0010	0,0024	0,0053	0,0102	0,0185	0,0313	0,0503	0,0778	0,1160	0,1681	0,2373	0,3277	0,4437	0,5905	0,7738
	6	0,7351	0,5314	0,3771	0,2621	0,1780	0,1176	0,0754	0,0467	0,0277	0,0156	0,0083	0,0041	0,0018	0,0007	0,0002	0,0001	0,0000	0,0000	0,0000
6	1	0,2321	0,3543	0,3993	0,3932	0,3560	0,3025	0,2437	0,1866	0,1359	0,0938	0,0609	0,0369	0,0205	0,0102	0,0044	0,0015	0,0004	0,0001	0,0000
	2	0,0305	0,0984	0,1762	0,2458	0,2966	0,3241	0,3280	0,3110	0,2780	0,2344	0,1861	0,1382	0,0951	0,0595	0,0330	0,0154	0,0055	0,0012	0,0001
	3	0,0021	0,0146	0,0415	0,0819	0,1318	0,1852	0,2355	0,2765	0,3032	0,3125	0,3032	0,2765	0,2355	0,1852	0,1318	0,0819	0,0415	0,0146	0,0021
	4	0,0001	0,0012	0,0055	0,0154	0,0330	0,0595	0,0951	0,1382	0,1861	0,2344	0,2780	0,3110	0,3280	0,3241	0,2966	0,2458	0,1762	0,0984	0,0305
	5	0,0000	0,0001	0,0004	0,0015	0,0044	0,0102	0,0205	0,0369	0,0609	0,0938	0,1359	0,1866	0,2437	0,3025	0,3560	0,3932	0,3993	0,3543	0,2321
	6	0,0000	0,0000	0,0000	0,0001	0,0002	0,0007	0,0018	0,0041	0,0083	0,0156	0,0277	0,0467	0,0754	0,1176	0,1780	0,2621	0,3771	0,5314	0,7351
	7	0,6983	0,4783	0,3206	0,2097	0,1335	0,0824	0,0490	0,0280	0,0152	0,0078	0,0037	0,0016	0,0006	0,0002	0,0001	0,0000	0,0000	0,0000	0,0000
7	1	0,2573	0,3720	0,3960	0,3670	0,3115	0,2471	0,1848	0,1306	0,0872	0,0547	0,0320	0,0172	0,0084	0,0036	0,0013	0,0004	0,0001	0,0000	0,0000
	2	0,0406	0,1240	0,2097	0,2753	0,3115	0,3177	0,2985	0,2613	0,2140	0,1641	0,1172	0,0774	0,0466	0,0250	0,0115	0,0043	0,0012	0,0002	0,0000
	3	0,0036	0,0230	0,0617	0,1147	0,1730	0,2269	0,2679	0,2903	0,2918	0,2734	0,2388	0,1935	0,1442	0,0972	0,0577	0,0287	0,0109	0,0026	0,0002
	4	0,0002	0,0026	0,0109	0,0287	0,0577	0,0972	0,1442	0,1935	0,2388	0,2734	0,2918	0,2903	0,2679	0,2269	0,1730	0,1147	0,0617	0,0230	0,0036
	5	0,0000	0,0002	0,0012	0,0043	0,0115	0,0250	0,0466	0,0774	0,1172	0,1641	0,2140	0,2613	0,2985	0,3177	0,3115	0,2753	0,2097	0,1240	0,0406
	6	0,0000	0,0000	0,0001	0,0004	0,0013	0,0036	0,0084	0,0172	0,0320	0,0547	0,0872	0,1306	0,1848	0,2471	0,3115	0,3670	0,3960	0,3720	0,2573
	7	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0006	0,0016	0,0037	0,0078	0,0152	0,0280	0,0490	0,0824	0,1335	0,2097	0,3206	0,4783	0,6983

Tabela 1 (Continuação).

<i>n</i>	<i>x</i>	<i>p</i>																		
		0,05	0,1	0,15	0,2	0,25	0,3	0,35	0,4	0,45	0,5	0,55	0,6	0,65	0,7	0,75	0,8	0,85	0,9	0,95
8	0	0,6634	0,4305	0,2725	0,1678	0,1001	0,0576	0,0319	0,0168	0,0084	0,0039	0,0017	0,0007	0,0002	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000
	1	0,2793	0,3826	0,3847	0,3355	0,2670	0,1977	0,1373	0,0896	0,0548	0,0313	0,0164	0,0079	0,0033	0,0012	0,0004	0,0001	0,0000	0,0000	0,0000
	2	0,0515	0,1488	0,2376	0,2936	0,3115	0,2965	0,2587	0,2090	0,1569	0,1094	0,0703	0,0413	0,0217	0,0100	0,0038	0,0011	0,0002	0,0000	0,0000
	3	0,0054	0,0331	0,0839	0,1468	0,2076	0,2541	0,2786	0,2787	0,2568	0,2188	0,1719	0,1239	0,0808	0,0467	0,0231	0,0092	0,0026	0,0004	0,0000
	4	0,0004	0,0046	0,0185	0,0459	0,0865	0,1361	0,1875	0,2322	0,2627	0,2734	0,2627	0,2322	0,1875	0,1361	0,0865	0,0459	0,0185	0,0046	0,0004
	5	0,0000	0,0004	0,0026	0,0092	0,0231	0,0467	0,0808	0,1239	0,1719	0,2188	0,2568	0,2787	0,2786	0,2541	0,2076	0,1468	0,0839	0,0331	0,0054
	6	0,0000	0,0000	0,0002	0,0011	0,0038	0,0100	0,0217	0,0413	0,0703	0,1094	0,1569	0,2090	0,2587	0,2965	0,3115	0,2936	0,2376	0,1488	0,0515
	7	0,0000	0,0000	0,0000	0,0001	0,0004	0,0012	0,0033	0,0079	0,0164	0,0313	0,0548	0,0896	0,1373	0,1977	0,2670	0,3355	0,3847	0,3826	0,2793
	8	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0007	0,0017	0,0039	0,0084	0,0168	0,0319	0,0576	0,1001	0,1678	0,2725	0,4305	0,6634
9	0	0,6302	0,3874	0,2316	0,1342	0,0751	0,0404	0,0207	0,0101	0,0046	0,0020	0,0008	0,0003	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	1	0,2985	0,3874	0,3679	0,3020	0,2253	0,1556	0,1004	0,0605	0,0339	0,0176	0,0083	0,0035	0,0013	0,0004	0,0001	0,0000	0,0000	0,0000	0,0000
	2	0,0629	0,1722	0,2597	0,3020	0,3003	0,2668	0,2162	0,1612	0,1110	0,0703	0,0407	0,0212	0,0098	0,0039	0,0012	0,0003	0,0000	0,0000	0,0000
	3	0,0077	0,0446	0,1069	0,1762	0,2336	0,2668	0,2716	0,2508	0,2119	0,1641	0,1160	0,0743	0,0424	0,0210	0,0087	0,0028	0,0006	0,0001	0,0000
	4	0,0006	0,0074	0,0283	0,0661	0,1168	0,1715	0,2194	0,2508	0,2600	0,2461	0,2128	0,1672	0,1181	0,0735	0,0389	0,0165	0,0050	0,0008	0,0000
	5	0,0000	0,0008	0,0050	0,0165	0,0389	0,0735	0,1181	0,1672	0,2128	0,2461	0,2600	0,2508	0,2194	0,1715	0,1168	0,0661	0,0283	0,0074	0,0006
	6	0,0000	0,0001	0,0006	0,0028	0,0087	0,0210	0,0424	0,0743	0,1160	0,1641	0,2119	0,2508	0,2716	0,2668	0,2336	0,1762	0,1069	0,0446	0,0077
	7	0,0000	0,0000	0,0000	0,0003	0,0012	0,0039	0,0098	0,0212	0,0407	0,0703	0,1110	0,1612	0,2162	0,2668	0,3003	0,3020	0,2597	0,1722	0,0629
	8	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0013	0,0035	0,0083	0,0176	0,0339	0,0605	0,1004	0,1556	0,2253	0,3020	0,3679	0,3874	0,2985
9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0008	0,0020	0,0046	0,0101	0,0207	0,0404	0,0751	0,1342	0,2316	0,3874	0,6302	
10	0	0,5987	0,3487	0,1969	0,1074	0,0563	0,0282	0,0135	0,0060	0,0025	0,0010	0,0003	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	
	1	0,3151	0,3874	0,3474	0,2684	0,1877	0,1211	0,0725	0,0403	0,0207	0,0098	0,0042	0,0016	0,0005	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000

Tabela 1 (Continuação).

<i>n</i>	<i>x</i>	<i>p</i>																		
		0,05	0,1	0,15	0,2	0,25	0,3	0,35	0,4	0,45	0,5	0,55	0,6	0,65	0,7	0,75	0,8	0,85	0,9	0,95
2	2	0,0746	0,1937	0,2759	0,3020	0,2816	0,2335	0,1757	0,1209	0,0763	0,0439	0,0229	0,0106	0,0043	0,0014	0,0004	0,0001	0,0000	0,0000	0,0000
	3	0,0105	0,0574	0,1298	0,2013	0,2503	0,2668	0,2522	0,2150	0,1665	0,1172	0,0746	0,0425	0,0212	0,0090	0,0031	0,0008	0,0001	0,0000	0,0000
	4	0,0010	0,0112	0,0401	0,0881	0,1460	0,2001	0,2377	0,2508	0,2384	0,2051	0,1596	0,1115	0,0689	0,0368	0,0162	0,0055	0,0012	0,0001	0,0000
	5	0,0001	0,0015	0,0085	0,0264	0,0584	0,1029	0,1536	0,2007	0,2340	0,2461	0,2340	0,2007	0,1536	0,1029	0,0584	0,0264	0,0085	0,0015	0,0001
	6	0,0000	0,0001	0,0012	0,0055	0,0162	0,0368	0,0689	0,1115	0,1596	0,2051	0,2384	0,2508	0,2377	0,2001	0,1460	0,0881	0,0401	0,0112	0,0010
	7	0,0000	0,0000	0,0001	0,0008	0,0031	0,0090	0,0212	0,0425	0,0746	0,1172	0,1665	0,2150	0,2522	0,2668	0,2503	0,2013	0,1298	0,0574	0,0105
	8	0,0000	0,0000	0,0000	0,0001	0,0004	0,0014	0,0043	0,0106	0,0229	0,0439	0,0763	0,1209	0,1757	0,2335	0,2816	0,3020	0,2759	0,1937	0,0746
	9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0016	0,0042	0,0098	0,0207	0,0403	0,0725	0,1211	0,1877	0,2684	0,3474	0,3874	0,3151
	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010	0,0025	0,0060	0,0135	0,0282	0,0563	0,1074	0,1969	0,3487	0,5987
	11	0	0,5688	0,3138	0,1673	0,0859	0,0422	0,0198	0,0088	0,0036	0,0014	0,0005	0,0002	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
1		0,3293	0,3835	0,3248	0,2362	0,1549	0,0932	0,0518	0,0266	0,0125	0,0054	0,0021	0,0007	0,0002	0,0000	0,0000	0,0000	0,0000	0,0000	
2		0,0867	0,2131	0,2866	0,2953	0,2581	0,1998	0,1395	0,0887	0,0513	0,0269	0,0126	0,0052	0,0018	0,0005	0,0001	0,0000	0,0000	0,0000	
3		0,0137	0,0710	0,1517	0,2215	0,2581	0,2568	0,2254	0,1774	0,1259	0,0806	0,0462	0,0234	0,0102	0,0037	0,0011	0,0002	0,0000	0,0000	
4		0,0014	0,0158	0,0536	0,1107	0,1721	0,2201	0,2428	0,2365	0,2060	0,1611	0,1128	0,0701	0,0379	0,0173	0,0064	0,0017	0,0003	0,0000	
5		0,0001	0,0025	0,0132	0,0388	0,0803	0,1321	0,1830	0,2207	0,2360	0,2256	0,1931	0,1471	0,0985	0,0566	0,0268	0,0097	0,0023	0,0003	
6		0,0000	0,0003	0,0023	0,0097	0,0268	0,0566	0,0985	0,1471	0,1931	0,2256	0,2360	0,2207	0,1830	0,1321	0,0803	0,0388	0,0132	0,0025	
7		0,0000	0,0000	0,0003	0,0017	0,0064	0,0173	0,0379	0,0701	0,1128	0,1611	0,2060	0,2365	0,2428	0,2201	0,1721	0,1107	0,0536	0,0158	
8		0,0000	0,0000	0,0000	0,0002	0,0011	0,0037	0,0102	0,0234	0,0462	0,0806	0,1259	0,1774	0,2254	0,2568	0,2581	0,2215	0,1517	0,0710	
9		0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0018	0,0052	0,0126	0,0269	0,0513	0,0887	0,1395	0,1998	0,2581	0,2953	0,2866	0,2131	
10		0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0007	0,0021	0,0054	0,0125	0,0266	0,0518	0,0932	0,1549	0,2362	0,3248	0,3835	
11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0005	0,0014	0,0036	0,0088	0,0198	0,0422	0,0859	0,1673	0,3138		

Tabela 1 (Continuação).

<i>n</i>	<i>x</i>	<i>p</i>																		
		0,05	0,1	0,15	0,2	0,25	0,3	0,35	0,4	0,45	0,5	0,55	0,6	0,65	0,7	0,75	0,8	0,85	0,9	0,95
12	0	0,5404	0,2824	0,1422	0,0687	0,0317	0,0138	0,0057	0,0022	0,0008	0,0002	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	1	0,3413	0,3766	0,3012	0,2062	0,1267	0,0712	0,0368	0,0174	0,0075	0,0029	0,0010	0,0003	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	2	0,0988	0,2301	0,2924	0,2835	0,2323	0,1678	0,1088	0,0639	0,0339	0,0161	0,0068	0,0025	0,0008	0,0002	0,0000	0,0000	0,0000	0,0000	0,0000
	3	0,0173	0,0852	0,1720	0,2362	0,2581	0,2397	0,1954	0,1419	0,0923	0,0537	0,0277	0,0125	0,0048	0,0015	0,0004	0,0001	0,0000	0,0000	0,0000
	4	0,0021	0,0213	0,0683	0,1329	0,1936	0,2311	0,2367	0,2128	0,1700	0,1208	0,0762	0,0420	0,0199	0,0078	0,0024	0,0005	0,0001	0,0000	0,0000
	5	0,0002	0,0038	0,0193	0,0532	0,1032	0,1585	0,2039	0,2270	0,2225	0,1934	0,1489	0,1009	0,0591	0,0291	0,0115	0,0033	0,0006	0,0000	0,0000
	6	0,0000	0,0005	0,0040	0,0155	0,0401	0,0792	0,1281	0,1766	0,2124	0,2256	0,2124	0,1766	0,1281	0,0792	0,0401	0,0155	0,0040	0,0005	0,0000
	7	0,0000	0,0000	0,0006	0,0033	0,0115	0,0291	0,0591	0,1009	0,1489	0,1934	0,2225	0,2270	0,2039	0,1585	0,1032	0,0532	0,0193	0,0038	0,0002
	8	0,0000	0,0000	0,0001	0,0005	0,0024	0,0078	0,0199	0,0420	0,0762	0,1208	0,1700	0,2128	0,2367	0,2311	0,1936	0,1329	0,0683	0,0213	0,0021
	9	0,0000	0,0000	0,0000	0,0001	0,0004	0,0015	0,0048	0,0125	0,0277	0,0537	0,0923	0,1419	0,1954	0,2397	0,2581	0,2362	0,1720	0,0852	0,0173
	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0008	0,0025	0,0068	0,0161	0,0339	0,0639	0,1088	0,1678	0,2323	0,2835	0,2924	0,2301	0,0988
	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010	0,0029	0,0075	0,0174	0,0368	0,0712	0,1267	0,2062	0,3012	0,3766	0,3413
12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0008	0,0022	0,0057	0,0138	0,0317	0,0687	0,1422	0,2824	0,5404	
13	0	0,5133	0,2542	0,1209	0,0550	0,0238	0,0097	0,0037	0,0013	0,0004	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	
	1	0,3512	0,3672	0,2774	0,1787	0,1029	0,0540	0,0259	0,0113	0,0045	0,0016	0,0005	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	
	2	0,1109	0,2448	0,2937	0,2680	0,2059	0,1388	0,0836	0,0453	0,0220	0,0095	0,0036	0,0012	0,0003	0,0001	0,0000	0,0000	0,0000	0,0000	
	3	0,0214	0,0997	0,1900	0,2457	0,2517	0,2181	0,1651	0,1107	0,0660	0,0349	0,0162	0,0065	0,0022	0,0006	0,0001	0,0000	0,0000	0,0000	0,0000
	4	0,0028	0,0277	0,0838	0,1535	0,2097	0,2337	0,2222	0,1845	0,1350	0,0873	0,0495	0,0243	0,0101	0,0034	0,0009	0,0001	0,0000	0,0000	0,0000
	5	0,0003	0,0055	0,0266	0,0691	0,1258	0,1803	0,2154	0,2214	0,1989	0,1571	0,1089	0,0656	0,0336	0,0142	0,0047	0,0011	0,0001	0,0000	0,0000
	6	0,0000	0,0008	0,0063	0,0230	0,0559	0,1030	0,1546	0,1968	0,2169	0,2095	0,1775	0,1312	0,0833	0,0442	0,0186	0,0058	0,0011	0,0001	0,0000
	7	0,0000	0,0001	0,0011	0,0058	0,0186	0,0442	0,0833	0,1312	0,1775	0,2095	0,2169	0,1968	0,1546	0,1030	0,0559	0,0230	0,0063	0,0008	0,0000
8	0,0000	0,0000	0,0001	0,0011	0,0047	0,0142	0,0336	0,0656	0,1089	0,1571	0,1989	0,2214	0,2154	0,1803	0,1258	0,0691	0,0266	0,0055	0,0003	

Tabela 1 (Continuação).

<i>n</i>	<i>x</i>	<i>p</i>																		
		0,05	0,1	0,15	0,2	0,25	0,3	0,35	0,4	0,45	0,5	0,55	0,6	0,65	0,7	0,75	0,8	0,85	0,9	0,95
9	9	0,0000	0,0000	0,0000	0,0001	0,0009	0,0034	0,0101	0,0243	0,0495	0,0873	0,1350	0,1845	0,2222	0,2337	0,2097	0,1535	0,0838	0,0277	0,0028
	10	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006	0,0022	0,0065	0,0162	0,0349	0,0660	0,1107	0,1651	0,2181	0,2517	0,2457	0,1900	0,0997	0,0214
	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0012	0,0036	0,0095	0,0220	0,0453	0,0836	0,1388	0,2059	0,2680	0,2937	0,2448	0,1109
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0016	0,0045	0,0113	0,0259	0,0540	0,1029	0,1787	0,2774	0,3672	0,3512
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0013	0,0037	0,0097	0,0238	0,0550	0,1209	0,2542	0,5133
14	0	0,4877	0,2288	0,1028	0,0440	0,0178	0,0068	0,0024	0,0008	0,0002	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	1	0,3593	0,3559	0,2539	0,1539	0,0832	0,0407	0,0181	0,0073	0,0027	0,0009	0,0002	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	2	0,1229	0,2570	0,2912	0,2501	0,1802	0,1134	0,0634	0,0317	0,0141	0,0056	0,0019	0,0005	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	3	0,0259	0,1142	0,2056	0,2501	0,2402	0,1943	0,1366	0,0845	0,0462	0,0222	0,0093	0,0033	0,0010	0,0002	0,0000	0,0000	0,0000	0,0000	0,0000
	4	0,0037	0,0349	0,0998	0,1720	0,2202	0,2290	0,2022	0,1549	0,1040	0,0611	0,0312	0,0136	0,0049	0,0014	0,0003	0,0000	0,0000	0,0000	0,0000
	5	0,0004	0,0078	0,0352	0,0860	0,1468	0,1963	0,2178	0,2066	0,1701	0,1222	0,0762	0,0408	0,0183	0,0066	0,0018	0,0003	0,0000	0,0000	0,0000
	6	0,0000	0,0013	0,0093	0,0322	0,0734	0,1262	0,1759	0,2066	0,2088	0,1833	0,1398	0,0918	0,0510	0,0232	0,0082	0,0020	0,0003	0,0000	0,0000
	7	0,0000	0,0002	0,0019	0,0092	0,0280	0,0618	0,1082	0,1574	0,1952	0,2095	0,1952	0,1574	0,1082	0,0618	0,0280	0,0092	0,0019	0,0002	0,0000
	8	0,0000	0,0000	0,0003	0,0020	0,0082	0,0232	0,0510	0,0918	0,1398	0,1833	0,2088	0,2066	0,1759	0,1262	0,0734	0,0322	0,0093	0,0013	0,0000
	9	0,0000	0,0000	0,0000	0,0003	0,0018	0,0066	0,0183	0,0408	0,0762	0,1222	0,1701	0,2066	0,2178	0,1963	0,1468	0,0860	0,0352	0,0078	0,0004
	10	0,0000	0,0000	0,0000	0,0000	0,0003	0,0014	0,0049	0,0136	0,0312	0,0611	0,1040	0,1549	0,2022	0,2290	0,2202	0,1720	0,0998	0,0349	0,0037
	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0010	0,0033	0,0093	0,0222	0,0462	0,0845	0,1366	0,1943	0,2402	0,2501	0,2056	0,1142	0,0259
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0019	0,0056	0,0141	0,0317	0,0634	0,1134	0,1802	0,2501	0,2912	0,2570	0,1229
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0009	0,0027	0,0073	0,0181	0,0407	0,0832	0,1539	0,2539	0,3559	0,3593
14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0008	0,0024	0,0068	0,0178	0,0440	0,1028	0,2288	0,4877	

Tabela 1 (Continuação).

<i>n</i>	<i>x</i>	<i>P</i>																		
		0,05	0,1	0,15	0,2	0,25	0,3	0,35	0,4	0,45	0,5	0,55	0,6	0,65	0,7	0,75	0,8	0,85	0,9	0,95
15	0	0,4633	0,2059	0,0874	0,0352	0,0134	0,0047	0,0016	0,0005	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	1	0,3658	0,3432	0,2312	0,1319	0,0668	0,0305	0,0126	0,0047	0,0016	0,0005	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	2	0,1348	0,2669	0,2856	0,2309	0,1559	0,0916	0,0476	0,0219	0,0090	0,0032	0,0010	0,0003	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	3	0,0307	0,1285	0,2184	0,2501	0,2252	0,1700	0,1110	0,0634	0,0318	0,0139	0,0052	0,0016	0,0004	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000
	4	0,0049	0,0428	0,1156	0,1876	0,2252	0,2186	0,1792	0,1268	0,0780	0,0417	0,0191	0,0074	0,0024	0,0006	0,0001	0,0000	0,0000	0,0000	0,0000
	5	0,0006	0,0105	0,0449	0,1032	0,1651	0,2061	0,2123	0,1859	0,1404	0,0916	0,0515	0,0245	0,0096	0,0030	0,0007	0,0001	0,0000	0,0000	0,0000
	6	0,0000	0,0019	0,0132	0,0430	0,0917	0,1472	0,1906	0,2066	0,1914	0,1527	0,1048	0,0612	0,0298	0,0116	0,0034	0,0007	0,0001	0,0000	0,0000
	7	0,0000	0,0003	0,0030	0,0138	0,0393	0,0811	0,1319	0,1771	0,2013	0,1964	0,1647	0,1181	0,0710	0,0348	0,0131	0,0035	0,0005	0,0000	0,0000
	8	0,0000	0,0000	0,0005	0,0035	0,0131	0,0348	0,0710	0,1181	0,1647	0,1964	0,2013	0,1771	0,1319	0,0811	0,0393	0,0138	0,0030	0,0003	0,0000
	9	0,0000	0,0000	0,0001	0,0007	0,0034	0,0116	0,0298	0,0612	0,1048	0,1527	0,1914	0,2066	0,1906	0,1472	0,0917	0,0430	0,0132	0,0019	0,0000
	10	0,0000	0,0000	0,0000	0,0001	0,0007	0,0030	0,0096	0,0245	0,0515	0,0916	0,1404	0,1859	0,2123	0,2061	0,1651	0,1032	0,0449	0,0105	0,0006
	11	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006	0,0024	0,0074	0,0191	0,0417	0,0780	0,1268	0,1792	0,2186	0,2252	0,1876	0,1156	0,0428	0,0049
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0016	0,0052	0,0139	0,0318	0,0634	0,1110	0,1700	0,2252	0,2501	0,2184	0,1285	0,0307
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010	0,0032	0,0090	0,0219	0,0476	0,0916	0,1559	0,2309	0,2856	0,2669	0,1348
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0016	0,0047	0,0126	0,0305	0,0668	0,1319	0,2312	0,3432	0,3658
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0016	0,0047	0,0134	0,0352	0,0874	0,2059	0,4633

Tabela 2 *Distribuição acumulada de Poisson: $P(X \leq x) = \sum_{j=0}^x \frac{\lambda^j e^{-\lambda}}{j!}$.*

x	λ									
	0,05	0,1	0,15	0,2	0,25	0,3	0,35	0,4	0,45	0,5
0	0,9512	0,9048	0,8607	0,8187	0,7788	0,7408	0,7047	0,6703	0,6376	0,6065
1	0,9988	0,9953	0,9898	0,9825	0,9735	0,9631	0,9513	0,9384	0,9246	0,9098
2	1,0000	0,9998	0,9995	0,9989	0,9978	0,9964	0,9945	0,9921	0,9891	0,9856
3	1,0000	1,0000	1,0000	0,9999	0,9999	0,9997	0,9995	0,9992	0,9988	0,9982
4	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9999	0,9998
5	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

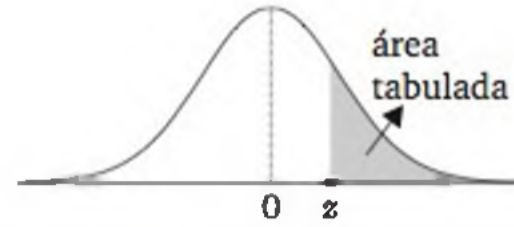
x	λ									
	0,55	0,6	0,65	0,7	0,75	0,8	0,85	0,9	0,95	
0	0,5769	0,5488	0,5220	0,4966	0,4724	0,4493	0,4274	0,4066	0,3867	
1	0,8943	0,8781	0,8614	0,8442	0,8266	0,8088	0,7907	0,7725	0,7541	
2	0,9815	0,9769	0,9717	0,9659	0,9595	0,9526	0,9451	0,9371	0,9287	
3	0,9975	0,9966	0,9956	0,9942	0,9927	0,9909	0,9889	0,9865	0,9839	
4	0,9997	0,9996	0,9994	0,9992	0,9989	0,9986	0,9982	0,9977	0,9971	
5	1,0000	1,0000	0,9999	0,9999	0,9999	0,9998	0,9997	0,9997	0,9995	
6	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	

x	λ									
	1	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9
0	0,3679	0,3329	0,3012	0,2725	0,2466	0,2231	0,2019	0,1827	0,1653	0,1496
1	0,7358	0,6990	0,6626	0,6268	0,5918	0,5578	0,5249	0,4932	0,4628	0,4337
2	0,9197	0,9004	0,8795	0,8571	0,8335	0,8088	0,7834	0,7572	0,7306	0,7037
3	0,9810	0,9743	0,9662	0,9569	0,9463	0,9344	0,9212	0,9068	0,8913	0,8747
4	0,9963	0,9946	0,9923	0,9893	0,9857	0,9814	0,9763	0,9704	0,9636	0,9559
5	0,9994	0,9990	0,9985	0,9978	0,9968	0,9955	0,9940	0,9920	0,9896	0,9868
6	0,9999	0,9999	0,9997	0,9996	0,9994	0,9991	0,9987	0,9981	0,9974	0,9966
7	1,0000	1,0000	1,0000	0,9999	0,9999	0,9998	0,9997	0,9996	0,9994	0,9992
8	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9999	0,9998

Tabela 2 (Continuação).

x	λ									
	2	2,5	3	4	5	6	7	8	9	10
0	0,1353	0,0821	0,0498	0,0183	0,0067	0,0025	0,0009	0,0003	0,0001	0,0000
1	0,4060	0,2873	0,1991	0,0916	0,0404	0,0174	0,0073	0,0030	0,0012	0,0005
2	0,6767	0,5438	0,4232	0,2381	0,1247	0,0620	0,0296	0,0138	0,0062	0,0028
3	0,8571	0,7576	0,6472	0,4335	0,2650	0,1512	0,0818	0,0424	0,0212	0,0103
4	0,9473	0,8912	0,8153	0,6288	0,4405	0,2851	0,1730	0,0996	0,0550	0,0293
5	0,9834	0,9580	0,9161	0,7851	0,6160	0,4457	0,3007	0,1912	0,1157	0,0671
6	0,9955	0,9858	0,9665	0,8893	0,7622	0,6063	0,4497	0,3134	0,2068	0,1301
7	0,9989	0,9958	0,9881	0,9489	0,8666	0,7440	0,5987	0,4530	0,3239	0,2202
8	0,9998	0,9989	0,9962	0,9786	0,9319	0,8472	0,7291	0,5925	0,4557	0,3328
9	1,0000	0,9997	0,9989	0,9919	0,9682	0,9161	0,8305	0,7166	0,5874	0,4579
10	1,0000	0,9999	0,9997	0,9972	0,9863	0,9574	0,9015	0,8159	0,7060	0,5830
11	1,0000	1,0000	0,9999	0,9991	0,9945	0,9799	0,9467	0,8881	0,8030	0,6968
12	1,0000	1,0000	1,0000	0,9997	0,9980	0,9912	0,9730	0,9362	0,8758	0,7916
13	1,0000	1,0000	1,0000	0,9999	0,9993	0,9964	0,9872	0,9658	0,9261	0,8645
14	1,0000	1,0000	1,0000	1,0000	0,9998	0,9986	0,9943	0,9827	0,9585	0,9165
15	1,0000	1,0000	1,0000	1,0000	0,9999	0,9995	0,9976	0,9918	0,9780	0,9513
16	1,0000	1,0000	1,0000	1,0000	1,0000	0,9998	0,9990	0,9963	0,9889	0,9730
17	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9996	0,9984	0,9947	0,9857
18	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9993	0,9976	0,9928
19	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9997	0,9989	0,9965
20	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9996	0,9984

Tabela 3 *Distribuição normal padrão.*

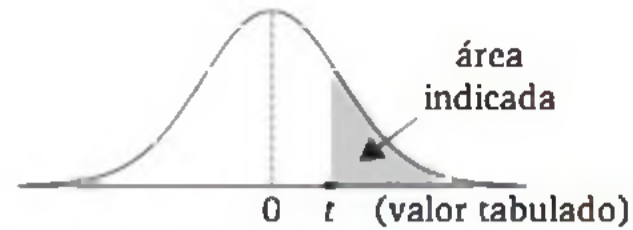


z	Segunda decimal de z									
	0	1	2	3	4	5	6	7	8	9
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2842	0,2810	0,2776
0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0722	0,0708	0,0694	0,0681
1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
1,8	0,0359	0,0352	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064

Tabela 3 (Continuação).

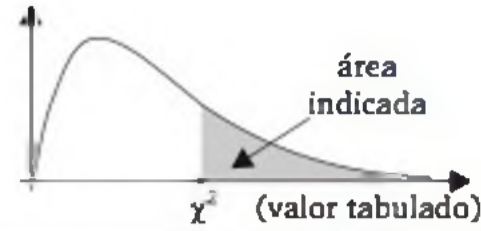
z	Segunda decimal de z									
	0	1	2	3	4	5	6	7	8	9
2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
2,9	0,0019	0,0018	0,0017	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
3,0	0,00135									
3,5	0,000 233									
4,0	0,000 031 7									
4,5	0,000 003 40									
5,0	0,000 000 287									

Tabela 4 Distribuição t de Student.



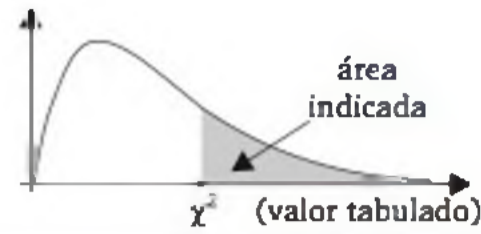
gl	Área na cauda superior								
	0,25	0,10	0,05	0,025	0,01	0,005	0,0025	0,001	0,0005
1	1,000	3,078	6,314	12,71	31,82	63,66	127,3	318,3	636,6
2	0,816	1,886	2,920	4,303	6,965	9,925	14,09	22,33	31,60
3	0,765	1,638	2,353	3,182	4,541	5,841	7,453	10,21	12,92
4	0,741	1,533	2,132	2,776	3,747	4,604	5,598	7,173	8,610
5	0,727	1,476	2,015	2,571	3,365	4,032	4,773	5,894	6,869
6	0,718	1,440	1,943	2,447	3,143	3,707	4,317	5,208	5,959
7	0,711	1,415	1,895	2,365	2,998	3,499	4,029	4,785	5,408
8	0,706	1,397	1,860	2,306	2,896	3,355	3,833	4,501	5,041
9	0,703	1,383	1,833	2,262	2,821	3,250	3,690	4,297	4,781
10	0,700	1,372	1,812	2,228	2,764	3,169	3,581	4,144	4,587
11	0,697	1,363	1,796	2,201	2,718	3,106	3,497	4,025	4,437
12	0,695	1,356	1,782	2,179	2,681	3,055	3,428	3,930	4,318
13	0,694	1,350	1,771	2,160	2,650	3,012	3,372	3,852	4,221
14	0,692	1,345	1,761	2,145	2,624	2,977	3,326	3,787	4,140
15	0,691	1,341	1,753	2,131	2,602	2,947	3,286	3,733	4,073
16	0,690	1,337	1,746	2,120	2,583	2,921	3,252	3,686	4,015
17	0,689	1,333	1,740	2,110	2,567	2,898	3,222	3,646	3,965
18	0,688	1,330	1,734	2,101	2,552	2,878	3,197	3,610	3,922
19	0,688	1,328	1,729	2,093	2,539	2,861	3,174	3,579	3,883
20	0,687	1,325	1,725	2,086	2,528	2,845	3,153	3,552	3,850
21	0,686	1,323	1,721	2,080	2,518	2,831	3,135	3,527	3,819
22	0,686	1,321	1,717	2,074	2,508	2,819	3,119	3,505	3,792
23	0,685	1,319	1,714	2,069	2,500	2,807	3,104	3,485	3,768
24	0,685	1,318	1,711	2,064	2,492	2,797	3,091	3,467	3,745
25	0,684	1,316	1,708	2,060	2,485	2,787	3,078	3,450	3,725
26	0,684	1,315	1,706	2,056	2,479	2,779	3,067	3,435	3,707
27	0,684	1,314	1,703	2,052	2,473	2,771	3,057	3,421	3,689
28	0,683	1,313	1,701	2,048	2,467	2,763	3,047	3,408	3,674
29	0,683	1,311	1,699	2,045	2,462	2,756	3,038	3,396	3,660
30	0,683	1,310	1,697	2,042	2,457	2,750	3,030	3,385	3,646
35	0,682	1,306	1,690	2,030	2,438	2,724	2,996	3,340	3,591
40	0,681	1,303	1,684	2,021	2,423	2,704	2,971	3,307	3,551
45	0,680	1,301	1,679	2,014	2,412	2,690	2,952	3,281	3,520
50	0,679	1,299	1,676	2,009	2,403	2,678	2,937	3,261	3,496
z	0,674	1,282	1,645	1,960	2,326	2,576	2,807	3,090	3,291

Tabela 5 Distribuição qui-quadrado.



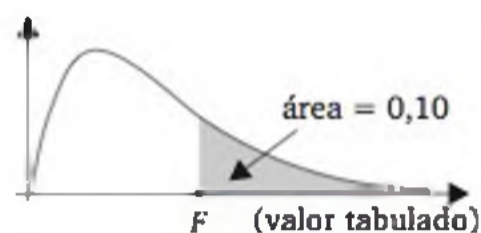
gl	Área na cauda superior								
	0,999	0,9975	0,995	0,99	0,975	0,95	0,9	0,75	
1	0,00	0,00	0,00	0,00	0,00	0,00	0,02	0,10	
2	0,00	0,01	0,01	0,02	0,05	0,10	0,21	0,58	
3	0,02	0,04	0,07	0,11	0,22	0,35	0,58	1,21	
4	0,09	0,14	0,21	0,30	0,48	0,71	1,06	1,92	
5	0,21	0,31	0,41	0,55	0,83	1,15	1,61	2,67	
6	0,38	0,53	0,68	0,87	1,24	1,64	2,20	3,45	
7	0,60	0,79	0,99	1,24	1,69	2,17	2,83	4,25	
8	0,86	1,10	1,34	1,65	2,18	2,73	3,49	5,07	
9	1,15	1,45	1,73	2,09	2,70	3,33	4,17	5,90	
10	1,48	1,83	2,16	2,56	3,25	3,94	4,87	6,74	
11	1,83	2,23	2,60	3,05	3,82	4,57	5,58	7,58	
12	2,21	2,66	3,07	3,57	4,40	5,23	6,30	8,44	
13	2,62	3,11	3,57	4,11	5,01	5,89	7,04	9,30	
14	3,04	3,58	4,07	4,66	5,63	6,57	7,79	10,17	
15	3,48	4,07	4,60	5,23	6,26	7,26	8,55	11,04	
16	3,94	4,57	5,14	5,81	6,91	7,96	9,31	11,91	
17	4,42	5,09	5,70	6,41	7,56	8,67	10,09	12,79	
18	4,90	5,62	6,26	7,01	8,23	9,39	10,86	13,68	
19	5,41	6,17	6,84	7,63	8,91	10,12	11,65	14,56	
20	5,92	6,72	7,43	8,26	9,59	10,85	12,44	15,45	
21	6,45	7,29	8,03	8,90	10,28	11,59	13,24	16,34	
22	6,98	7,86	8,64	9,54	10,98	12,34	14,04	17,24	
23	7,53	8,45	9,26	10,20	11,69	13,09	14,85	18,14	
24	8,08	9,04	9,89	10,86	12,40	13,85	15,66	19,04	
25	8,65	9,65	10,52	11,52	13,12	14,61	16,47	19,94	
26	9,22	10,26	11,16	12,20	13,84	15,38	17,29	20,84	
27	9,80	10,87	11,81	12,88	14,57	16,15	18,11	21,75	
28	10,39	11,50	12,46	13,56	15,31	16,93	18,94	22,66	
29	10,99	12,13	13,12	14,26	16,05	17,71	19,77	23,57	
30	11,59	12,76	13,79	14,95	16,79	18,49	20,60	24,48	
35	14,69	16,03	17,19	18,51	20,57	22,47	24,80	29,05	
40	17,92	19,42	20,71	22,16	24,43	26,51	29,05	33,66	
45	21,25	22,90	24,31	25,90	28,37	30,61	33,35	38,29	
50	24,67	26,46	27,99	29,71	32,36	34,76	37,69	42,94	
100	61,92	64,86	67,33	70,06	74,22	77,93	82,36	90,13	

Tabela 5 (Continuação).



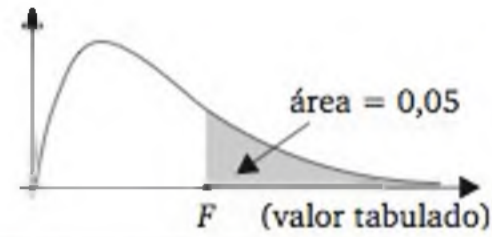
gl	Área na cauda superior							
	0,25	0,10	0,05	0,025	0,01	0,005	0,0025	0,001
1	1,32	2,71	3,84	5,02	6,63	7,88	9,14	10,83
2	2,77	4,61	5,99	7,38	9,21	10,60	11,98	13,82
3	4,11	6,25	7,81	9,35	11,34	12,84	14,32	16,27
4	5,39	7,78	9,49	11,14	13,28	14,86	16,42	18,47
5	6,63	9,24	11,07	12,83	15,09	16,75	18,39	20,51
6	7,84	10,64	12,59	14,45	16,81	18,55	20,25	22,46
7	9,04	12,02	14,07	16,01	18,48	20,28	22,04	24,32
8	10,22	13,36	15,51	17,53	20,09	21,95	23,77	26,12
9	11,39	14,68	16,92	19,02	21,67	23,59	25,46	27,88
10	12,55	15,99	18,31	20,48	23,21	25,19	27,11	29,59
11	13,70	17,28	19,68	21,92	24,73	26,76	28,73	31,26
12	14,85	18,55	21,03	23,34	26,22	28,30	30,32	32,91
13	15,98	19,81	22,36	24,74	27,69	29,82	31,88	34,53
14	17,12	21,06	23,68	26,12	29,14	31,32	33,43	36,12
15	18,25	22,31	25,00	27,49	30,58	32,80	34,95	37,70
16	19,37	23,54	26,30	28,85	32,00	34,27	36,46	39,25
17	20,49	24,77	27,59	30,19	33,41	35,72	37,95	40,79
18	21,60	25,99	28,87	31,53	34,81	37,16	39,42	42,31
19	22,72	27,20	30,14	32,85	36,19	38,58	40,88	43,82
20	23,83	28,41	31,41	34,17	37,57	40,00	42,34	45,31
21	24,93	29,62	32,67	35,48	38,93	41,40	43,77	46,80
22	26,04	30,81	33,92	36,78	40,29	42,80	45,20	48,27
23	27,14	32,01	35,17	38,08	41,64	44,18	46,62	49,73
24	28,24	33,20	36,42	39,36	42,98	45,56	48,03	51,18
25	29,34	34,38	37,65	40,65	44,31	46,93	49,44	52,62
26	30,43	35,56	38,89	41,92	45,64	48,29	50,83	54,05
27	31,53	36,74	40,11	43,19	46,96	49,65	52,22	55,48
28	32,62	37,92	41,34	44,46	48,28	50,99	53,59	56,89
29	33,71	39,09	42,56	45,72	49,59	52,34	54,97	58,30
30	34,80	40,26	43,77	46,98	50,89	53,67	56,33	59,70
35	40,22	46,06	49,80	53,20	57,34	60,27	63,08	66,62
40	45,62	51,81	55,76	59,34	63,69	66,77	69,70	73,40
45	50,98	57,51	61,66	65,41	69,96	73,17	76,22	80,08
50	56,33	63,17	67,50	71,42	76,15	79,49	82,66	86,66
100	109,1	118,5	124,3	129,6	135,8	140,2	144,3	149,4

Tabela 6 *Distribuição F de Snedecor.*
 $\alpha = 0,10$

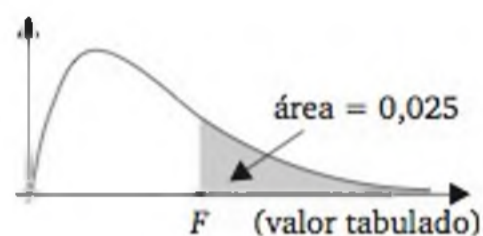


gl denom.	Graus de liberdade no numerador									
	1	2	3	4	5	6	7	8	9	10
1	39,86	49,50	53,59	55,83	57,24	58,20	58,91	59,44	59,86	60,19
2	8,53	9,00	9,16	9,24	9,29	9,33	9,35	9,37	9,38	9,39
3	5,54	5,46	5,39	5,34	5,31	5,28	5,27	5,25	5,24	5,23
4	4,54	4,32	4,19	4,11	4,05	4,01	3,98	3,95	3,94	3,92
5	4,06	3,78	3,62	3,52	3,45	3,40	3,37	3,34	3,32	3,30
6	3,78	3,46	3,29	3,18	3,11	3,05	3,01	2,98	2,96	2,94
7	3,59	3,26	3,07	2,96	2,88	2,83	2,78	2,75	2,72	2,70
8	3,46	3,11	2,92	2,81	2,73	2,67	2,62	2,59	2,56	2,54
9	3,36	3,01	2,81	2,69	2,61	2,55	2,51	2,47	2,44	2,42
10	3,29	2,92	2,73	2,61	2,52	2,46	2,41	2,38	2,35	2,32
11	3,23	2,86	2,66	2,54	2,45	2,39	2,34	2,30	2,27	2,25
12	3,18	2,81	2,61	2,48	2,39	2,33	2,28	2,24	2,21	2,19
13	3,14	2,76	2,56	2,43	2,35	2,28	2,23	2,20	2,16	2,14
14	3,10	2,73	2,52	2,39	2,31	2,24	2,19	2,15	2,12	2,10
15	3,07	2,70	2,49	2,36	2,27	2,21	2,16	2,12	2,09	2,06
16	3,05	2,67	2,46	2,33	2,24	2,18	2,13	2,09	2,06	2,03
17	3,03	2,64	2,44	2,31	2,22	2,15	2,10	2,06	2,03	2,00
18	3,01	2,62	2,42	2,29	2,20	2,13	2,08	2,04	2,00	1,98
19	2,99	2,61	2,40	2,27	2,18	2,11	2,06	2,02	1,98	1,96
20	2,97	2,59	2,38	2,25	2,16	2,09	2,04	2,00	1,96	1,94
21	2,96	2,57	2,36	2,23	2,14	2,08	2,02	1,98	1,95	1,92
22	2,95	2,56	2,35	2,22	2,13	2,06	2,01	1,97	1,93	1,90
23	2,94	2,55	2,34	2,21	2,11	2,05	1,99	1,95	1,92	1,89
24	2,93	2,54	2,33	2,19	2,10	2,04	1,98	1,94	1,91	1,88
25	2,92	2,53	2,32	2,18	2,09	2,02	1,97	1,93	1,89	1,87
26	2,91	2,52	2,31	2,17	2,08	2,01	1,96	1,92	1,88	1,86
27	2,90	2,51	2,30	2,17	2,07	2,00	1,95	1,91	1,87	1,85
28	2,89	2,50	2,29	2,16	2,06	2,00	1,94	1,90	1,87	1,84
29	2,89	2,50	2,28	2,15	2,06	1,99	1,93	1,89	1,86	1,83
30	2,88	2,49	2,28	2,14	2,05	1,98	1,93	1,88	1,85	1,82
35	2,85	2,46	2,25	2,11	2,02	1,95	1,90	1,85	1,82	1,79
40	2,84	2,44	2,23	2,09	2,00	1,93	1,87	1,83	1,79	1,76
45	2,82	2,42	2,21	2,07	1,98	1,91	1,85	1,81	1,77	1,74
50	2,81	2,41	2,20	2,06	1,97	1,90	1,84	1,80	1,76	1,73
100	2,76	2,36	2,14	2,00	1,91	1,83	1,78	1,73	1,69	1,66

Tabela 6 (Continuação).
 $\alpha = 0,05$

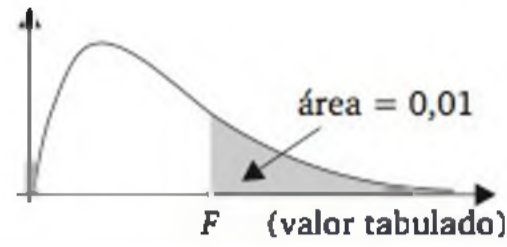


gl denom.	Graus de liberdade no numerador									
	1	2	3	4	5	6	7	8	9	10
1	161,45	199,50	215,71	224,58	230,16	233,99	236,77	238,88	240,54	241,88
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16
35	4,12	3,27	2,87	2,64	2,49	2,37	2,29	2,22	2,16	2,11
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08
45	4,06	3,20	2,81	2,58	2,42	2,31	2,22	2,15	2,10	2,05
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,03
100	3,94	3,09	2,70	2,46	2,31	2,19	2,10	2,03	1,97	1,93

Tabela 6 (Continuação).
 $\alpha = 0,025$ 

gl denom.	Graus de liberdade no numerador									
	1	2	3	4	5	6	7	8	9	10
1	647,79	799,48	864,15	899,60	921,83	937,11	948,20	956,64	963,28	968,63
2	38,51	39,00	39,17	39,25	39,30	39,33	39,36	39,37	39,39	39,40
3	17,44	16,04	15,44	15,10	14,88	14,73	14,62	14,54	14,47	14,42
4	12,22	10,65	9,98	9,60	9,36	9,20	9,07	8,98	8,90	8,84
5	10,01	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68	6,62
6	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52	5,46
7	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82	4,76
8	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36	4,30
9	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03	3,96
10	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78	3,72
11	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,59	3,53
12	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44	3,37
13	6,41	4,97	4,35	4,00	3,77	3,60	3,48	3,39	3,31	3,25
14	6,30	4,86	4,24	3,89	3,66	3,50	3,38	3,29	3,21	3,15
15	6,20	4,77	4,15	3,80	3,58	3,41	3,29	3,20	3,12	3,06
16	6,12	4,69	4,08	3,73	3,50	3,34	3,22	3,12	3,05	2,99
17	6,04	4,62	4,01	3,66	3,44	3,28	3,16	3,06	2,98	2,92
18	5,98	4,56	3,95	3,61	3,38	3,22	3,10	3,01	2,93	2,87
19	5,92	4,51	3,90	3,56	3,33	3,17	3,05	2,96	2,88	2,82
20	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84	2,77
21	5,83	4,42	3,82	3,48	3,25	3,09	2,97	2,87	2,80	2,73
22	5,79	4,38	3,78	3,44	3,22	3,05	2,93	2,84	2,76	2,70
23	5,75	4,35	3,75	3,41	3,18	3,02	2,90	2,81	2,73	2,67
24	5,72	4,32	3,72	3,38	3,15	2,99	2,87	2,78	2,70	2,64
25	5,69	4,29	3,69	3,35	3,13	2,97	2,85	2,75	2,68	2,61
26	5,66	4,27	3,67	3,33	3,10	2,94	2,82	2,73	2,65	2,59
27	5,63	4,24	3,65	3,31	3,08	2,92	2,80	2,71	2,63	2,57
28	5,61	4,22	3,63	3,29	3,06	2,90	2,78	2,69	2,61	2,55
29	5,59	4,20	3,61	3,27	3,04	2,88	2,76	2,67	2,59	2,53
30	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57	2,51
35	5,48	4,11	3,52	3,18	2,96	2,80	2,68	2,58	2,50	2,44
40	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,45	2,39
45	5,38	4,01	3,42	3,09	2,86	2,70	2,58	2,49	2,41	2,35
50	5,34	3,97	3,39	3,05	2,83	2,67	2,55	2,46	2,38	2,32
100	5,18	3,83	3,25	2,92	2,70	2,54	2,42	2,32	2,24	2,18

Tabela 6 (Continuação).
 $\alpha = 0,01$



gl denom.	Graus de liberdade no numerador									
	1	2	3	4	5	6	7	8	9	10
1	4.052,2	4.999,3	5.403,5	5.624,3	5.764,0	5.859,0	5.928,3	5.981,0	6.022,4	6.055,9
2	98,50	99,00	99,16	99,25	99,30	99,33	99,36	99,38	99,39	99,40
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,34	7,23
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69
17	8,40	6,11	5,19	4,67	4,34	4,10	3,93	3,79	3,68	3,59
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98
35	7,42	5,27	4,40	3,91	3,59	3,37	3,20	3,07	2,96	2,88
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80
45	7,23	5,11	4,25	3,77	3,45	3,23	3,07	2,94	2,83	2,74
50	7,17	5,06	4,20	3,72	3,41	3,19	3,02	2,89	2,78	2,70
100	6,90	4,82	3,98	3,51	3,21	2,99	2,82	2,69	2,59	2,50

Tabela 7 Valores críticos (d_c) para os testes de Kolmogorov-Smirnov e Lilliefors.

Kolmogorov-Smirnov: parâmetros conhecidos			Lilliefors: parâmetros estimados		
n	$\alpha = 5\%$	$\alpha = 1\%$	n	$\alpha = 5\%$	$\alpha = 1\%$
1	0,975	0,995	4	0,381	0,417
2	0,842	0,929	5	0,337	0,405
3	0,708	0,829	6	0,319	0,364
4	0,624	0,734	7	0,300	0,348
5	0,563	0,669	8	0,285	0,331
6	0,519	0,617	9	0,271	0,311
7	0,483	0,576	10	0,258	0,294
8	0,454	0,542	11	0,249	0,284
9	0,430	0,513	12	0,242	0,275
10	0,409	0,490	13	0,234	0,268
11	0,391	0,468	14	0,227	0,261
12	0,375	0,449	15	0,220	0,257
13	0,361	0,432	16	0,213	0,250
14	0,349	0,418	17	0,206	0,245
15	0,338	0,404	18	0,200	0,239
16	0,327	0,392	19	0,179	0,235
17	0,318	0,381	20	0,190	0,231
18	0,309	0,371	25	0,173	0,200
19	0,301	0,361	30	0,161	0,187
20	0,294	0,352	$n > 30$ $d_c = 0,886/\sqrt{n}$ $d_c = 1,031/\sqrt{n}$		
25	0,264	0,317			
30	0,242	0,290			
35	0,224	0,269			
40	0,210	0,252			
45	0,198	0,238			
50	0,188	0,227			
$n > 50$ $d_c = 1,36/\sqrt{n}$ $d_c = 1,63/\sqrt{n}$					

Tabela 8 *Escores limites para a estatística S_+ do teste de sinais por postos (Wilcoxon).*

<i>n</i>	<i>P(S₊ ≤ s_c) aproximadamente igual a</i>																
	0,005	0,01	0,025	0,05	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	0,95	0,975	0,99	0,995
4	0	0	0	0	1	3	3	4	5	6	7	7	9	10	10	10	10
5	0	0	0	1	3	4	5	6	7.5	9	10	11	12	14	15	15	15
6	0	0	1	3	4	6	8	9	10.5	12	13	15	17	18	20	21	21
7	0	1	3	4	6	9	11	12	14	16	17	19	22	24	25	27	28
8	1	2	4	6	9	12	14	16	18	20	22	24	27	30	32	34	35
9	2	4	6	9	11	15	18	20	22.5	25	27	30	34	36	39	41	43
10	4	6	9	11	15	19	22	25	27.5	30	33	36	40	44	46	49	51
11	6	8	11	14	18	23	27	30	33	36	39	43	48	52	55	58	60
12	8	10	14	18	22	28	32	36	39	42	46	50	56	60	64	68	70
13	10	13	18	22	27	33	38	42	45.5	49	53	58	64	69	73	78	81
14	13	16	22	26	32	39	44	48	52.5	57	61	66	73	79	83	89	92
15	16	20	26	31	37	45	51	55	60	65	69	75	83	89	94	100	104
16	20	24	30	36	43	51	58	63	68	73	78	85	93	100	106	112	116
17	24	28	35	42	49	58	65	71	76.5	82	88	95	104	111	118	125	129
18	28	33	41	48	56	66	73	80	85.5	91	98	105	115	123	130	138	143
19	33	38	47	54	63	74	82	89	95	101	108	116	127	136	143	152	157
20	38	44	53	61	70	82	91	98	105	112	119	128	140	149	157	166	172

Tabela 9 *Escores limites para a estatística U do teste de Mann-Whitney.*

$n_1 = 2$	$P(U \leq u_c)$ aproximadamente igual a											
n_2	0,001	0,005	0,01	0,025	0,05	0,1	0,9	0,95	0,975	0,99	0,995	0,999
2	0	0	0	0	0	0	4	4	4	4	4	4
3	0	0	0	0	0	1	5	6	6	6	6	6
4	0	0	0	0	0	1	7	8	8	8	8	8
5	0	0	0	0	1	2	8	9	10	10	10	10
6	0	0	0	0	1	2	10	11	12	12	12	12
7	0	0	0	0	1	2	12	13	14	14	14	14
8	0	0	0	1	2	3	13	14	15	16	16	16
9	0	0	0	1	2	3	15	16	17	18	18	18
10	0	0	0	1	2	4	16	18	19	20	20	20
11	0	0	0	1	2	4	18	20	21	22	22	22
12	0	0	0	2	3	5	19	21	22	24	24	24
13	0	0	1	2	3	5	21	23	24	25	26	26
14	0	0	1	2	4	5	23	24	26	27	28	28
15	0	0	1	2	4	6	24	26	28	29	30	30
16	0	0	1	2	4	6	26	29	30	31	32	32
17	0	0	1	3	4	7	27	30	31	33	34	34
18	0	0	1	3	5	7	29	31	33	35	36	36
19	0	1	2	3	5	8	30	33	36	36	37	38
20	0	1	2	3	5	8	32	35	37	38	39	40

Tabela 9 (Continuação).

$n_1 = 3$	$P(U \leq u_c)$ aproximadamente igual a											
n_2	0,001	0,005	0,01	0,025	0,05	0,1	0,9	0,95	0,975	0,99	0,995	0,999
2	0	0	0	0	0	1	5	6	6	6	6	6
3	0	0	0	0	1	2	7	8	9	9	9	9
4	0	0	0	0	1	2	10	11	12	12	12	12
5	0	0	0	1	2	3	12	13	14	15	15	15
6	0	0	0	2	3	4	14	15	16	18	18	18
7	0	0	1	2	3	5	16	18	19	20	21	21
8	0	0	1	3	4	6	18	20	21	23	24	24
9	0	0	2	3	5	6	21	22	24	25	26	27
10	0	1	2	4	5	7	23	25	26	28	29	30
11	0	1	2	4	6	8	25	27	29	31	32	33
12	0	2	3	5	6	9	27	30	31	33	34	36
13	0	2	3	5	7	10	29	32	34	36	37	39
14	0	2	3	6	8	11	31	34	36	39	40	42
15	0	3	4	6	8	11	34	37	39	41	42	45
16	0	3	4	7	9	12	36	39	41	44	45	48
17	1	3	5	7	10	13	38	41	44	46	48	50
18	1	3	5	8	10	14	40	44	46	49	51	53
19	1	4	5	8	11	15	42	46	49	52	53	55
20	1	4	6	9	12	16	44	48	51	54	56	59

Tabela 9 (Continuação).

$n_1 = 4$	$P(U \leq u_c)$ aproximadamente igual a											
n_2	0,001	0,005	0,01	0,025	0,05	0,1	0,9	0,95	0,975	0,99	0,995	0,999
2	0	0	0	0	0	1	7	8	8	8	8	8
3	0	0	0	0	1	2	10	11	12	12	12	12
4	0	0	0	1	2	4	12	14	15	16	16	16
5	0	0	1	2	3	5	15	17	18	19	20	20
6	0	1	2	3	4	6	18	20	21	22	23	24
7	0	1	2	4	5	7	21	23	24	26	27	28
8	0	2	3	5	6	8	24	26	27	29	30	32
9	0	2	4	5	7	10	26	29	31	32	34	36
10	1	3	4	6	8	11	29	32	34	36	37	39
11	1	3	5	7	9	12	32	35	37	39	41	43
12	1	4	6	8	10	13	35	38	40	42	44	47
13	2	4	6	9	11	14	38	41	43	46	48	50
14	2	5	7	10	12	16	40	44	46	49	51	54
15	2	6	8	11	13	17	43	47	49	52	54	58
16	3	6	8	12	15	18	46	49	52	56	58	61
17	3	7	9	12	16	19	49	52	56	59	61	65
18	4	7	10	13	17	21	51	55	59	62	65	68
19	4	8	10	14	18	22	54	58	62	66	68	72
20	4	9	11	15	19	23	57	61	65	69	71	76

Tabela 9 (Continuação).

$n_1 = 5$	$P(U \leq u_c)$ aproximadamente igual a											
n_2	0,001	0,005	0,01	0,025	0,05	0,1	0,9	0,95	0,975	0,99	0,995	0,999
2	0	0	0	0	1	2	8	9	10	10	10	10
3	0	0	0	1	2	3	12	13	14	15	15	15
4	0	0	1	2	3	5	15	17	18	19	20	20
5	0	1	2	3	5	6	19	20	22	23	24	25
6	0	2	3	4	6	8	22	24	26	27	28	30
7	0	2	4	6	7	9	26	28	29	31	33	35
8	1	3	5	7	8	11	29	31	33	35	37	39
9	2	4	6	8	10	13	32	35	37	39	41	43
10	2	5	7	9	12	14	36	38	41	43	45	48
11	3	6	8	10	13	16	39	42	45	47	49	52
12	3	7	9	12	14	18	42	46	48	51	53	57
13	4	8	10	13	16	19	46	49	52	55	57	61
14	4	8	11	14	17	21	49	53	56	59	62	66
15	5	9	12	15	19	23	52	56	60	63	66	70
16	6	10	13	16	20	24	56	60	64	67	70	74
17	6	11	14	18	21	26	59	64	67	71	74	79
18	7	12	15	19	23	28	62	67	71	75	78	83
19	8	13	16	20	24	29	66	71	75	79	82	87
20	8	14	17	21	26	31	69	74	79	83	86	92

Tabela 9 (Continuação).

$n_1 = 6$	$P(U \leq u_c)$ aproximadamente igual a											
n_2	0,001	0,005	0,01	0,025	0,05	0,1	0,9	0,95	0,975	0,99	0,995	0,999
2	0	0	0	0	1	2	10	11	12	12	12	12
3	0	0	0	2	3	4	14	15	16	18	18	18
4	0	1	2	3	4	6	18	20	21	22	23	24
5	0	2	3	4	6	8	22	24	26	27	28	30
6	0	3	4	6	8	10	26	28	30	32	33	36
7	0	4	5	7	9	12	30	33	35	37	38	42
8	2	5	7	9	11	14	34	37	39	41	43	46
9	3	6	8	11	13	16	38	41	43	46	48	51
10	4	7	9	12	15	18	42	45	48	51	53	56
11	5	8	10	14	17	20	46	49	52	56	58	61
12	5	10	12	15	18	22	50	54	57	60	62	67
13	6	11	13	17	20	24	54	58	61	65	67	72
14	7	12	14	18	22	26	58	62	66	70	72	77
15	8	13	16	20	24	28	62	66	70	74	77	82
16	9	14	17	22	26	30	66	70	74	79	82	87
17	10	16	19	23	27	32	70	75	79	83	86	92
18	11	17	20	25	29	35	73	79	83	88	91	97
19	12	18	21	26	31	37	77	83	88	93	96	102
20	13	19	23	28	33	39	81	87	92	97	101	107

Tabela 9 (Continuação).

$n_1 = 7$	$P(U \leq u_c)$ aproximadamente igual a											
n_2	0,001	0,005	0,01	0,025	0,05	0,1	0,9	0,95	0,975	0,99	0,995	0,999
2	0	0	0	0	1	2	12	13	14	14	14	14
3	0	0	1	2	3	5	16	18	19	20	21	21
4	0	1	2	4	5	7	21	23	24	26	27	28
5	0	2	4	6	7	9	26	28	29	31	33	35
6	1	4	5	7	9	12	30	33	35	37	38	41
7	2	5	7	9	12	14	35	37	40	42	44	47
8	3	7	8	11	14	17	39	42	45	48	49	53
9	4	8	10	13	16	19	44	47	50	53	55	59
10	6	10	12	15	18	22	48	52	55	58	60	64
11	7	11	13	17	20	24	53	57	60	64	66	70
12	8	13	15	19	22	27	57	62	65	69	71	76
13	9	14	17	21	25	29	62	66	70	74	77	82
14	10	16	18	23	27	32	66	71	75	80	82	88
15	11	17	20	25	29	34	71	76	80	85	88	94
16	12	19	22	27	31	37	75	81	85	90	93	100
17	14	20	24	29	34	39	80	85	90	95	99	105
18	15	22	25	31	36	42	84	90	95	101	104	111
19	16	23	27	33	38	44	89	95	100	106	110	117
20	17	25	29	35	40	47	93	100	105	111	115	123

Tabela 9 (Continuação).

$n_1 = 8$	$P(U \leq u_c)$ aproximadamente igual a											
n_2	0,001	0,005	0,01	0,025	0,05	0,1	0,9	0,95	0,975	0,99	0,995	0,999
2	0	0	0	1	2	3	13	14	15	16	16	16
3	0	0	1	3	4	6	18	20	21	23	24	24
4	0	2	3	5	6	8	24	26	27	29	30	32
5	1	3	5	7	9	11	29	31	33	35	37	39
6	2	5	7	9	11	14	34	37	39	41	43	46
7	3	7	8	11	14	17	39	42	45	48	49	53
8	5	8	10	14	16	20	44	48	50	54	56	59
9	6	10	12	16	19	23	49	53	56	60	62	66
10	7	12	14	18	21	25	55	59	62	66	68	73
11	9	14	16	20	24	28	60	64	68	72	74	79
12	10	16	18	23	27	31	65	69	73	78	80	86
13	12	18	21	25	29	34	70	75	79	83	86	92
14	13	19	23	27	32	37	75	80	85	89	93	99
15	15	21	25	30	34	40	80	86	90	95	99	105
16	16	23	27	32	37	43	85	91	96	101	105	112
17	18	25	29	35	40	46	90	96	101	107	111	118
18	19	27	31	37	42	49	95	102	107	113	117	125
19	21	29	33	39	45	52	100	107	113	119	123	131
20	22	31	35	42	48	55	105	112	118	125	129	138

Tabela 9 (Continuação).

$n_1 = 9$	$P(U \leq u_c)$ aproximadamente igual a											
n_2	0,001	0,005	0,01	0,025	0,05	0,1	0,9	0,95	0,975	0,99	0,995	0,999
2	0	0	0	1	2	3	15	16	17	18	18	18
3	0	1	2	3	5	6	21	22	24	25	26	27
4	0	2	4	5	7	10	26	29	31	32	34	36
5	2	4	6	8	10	13	32	35	37	39	41	43
6	3	6	8	11	13	16	38	41	43	46	48	51
7	4	8	10	13	16	19	44	47	50	53	55	59
8	6	10	12	16	19	23	49	53	56	60	62	66
9	8	12	15	18	22	26	55	59	63	66	69	73
10	9	14	17	21	25	29	61	65	69	73	76	81
11	11	17	19	24	28	32	67	71	75	80	82	88
12	13	19	22	27	31	36	72	77	81	86	89	95
13	15	21	24	29	34	39	78	83	88	93	96	102
14	16	23	27	32	37	42	84	89	94	99	103	110
15	18	25	29	35	40	46	89	95	100	106	110	117
16	20	28	32	38	43	49	95	101	106	112	116	124
17	22	30	34	40	46	53	100	107	113	119	123	131
18	24	32	37	43	49	56	106	113	119	125	130	138
19	26	34	39	46	52	59	112	119	125	132	137	145
20	27	37	41	49	55	63	117	125	131	139	143	153

Tabela 9 (Continuação).

$n_1 = 10$	$P(U \leq u_c)$ aproximadamente igual a												
	n_2	0,001	0,005	0,01	0,025	0,05	0,1	0,9	0,95	0,975	0,99	0,995	0,999
2	0	0	0	1	2	4	16	18	19	20	20	20	20
3	0	1	2	4	5	7	23	25	26	28	29	29	30
4	1	3	4	6	8	11	29	32	34	36	37	37	39
5	2	5	7	9	12	14	36	38	41	43	45	45	48
6	4	7	9	12	15	18	42	45	48	51	53	53	56
7	6	10	12	15	18	22	48	52	55	58	60	60	64
8	7	12	14	18	21	25	55	59	62	66	68	68	73
9	9	14	17	21	25	29	61	65	69	73	76	76	81
10	11	17	20	24	28	33	67	72	76	80	83	83	89
11	13	19	23	27	32	37	73	78	83	87	91	91	97
12	15	22	25	30	35	40	80	85	90	95	98	98	105
13	18	25	28	34	38	44	86	92	96	102	105	105	112
14	20	27	31	37	42	48	92	98	103	109	113	113	120
15	22	30	34	40	45	52	98	105	110	116	120	120	128
16	24	32	37	43	49	55	105	111	117	123	128	128	136
17	26	35	39	46	52	59	111	118	124	131	135	135	144
18	28	38	42	49	56	63	117	124	131	138	142	142	152
19	30	40	45	53	59	67	123	131	137	145	150	150	160
20	33	43	48	56	63	71	129	137	144	152	157	157	167

Tabela 9 (Continuação).

$n_1 = 11$	$P(U \leq u_c)$ aproximadamente igual a											
n_2	0,001	0,005	0,01	0,025	0,05	0,1	0,9	0,95	0,975	0,99	0,995	0,999
2	0	0	0	1	2	4	18	20	21	22	22	22
3	0	1	2	4	6	8	25	27	29	31	32	33
4	1	3	5	7	9	12	32	35	37	39	41	43
5	3	6	8	10	13	16	39	42	45	47	49	52
6	5	8	10	14	17	20	46	49	52	56	58	61
7	7	11	13	17	20	24	53	57	60	64	66	70
8	9	14	16	20	24	28	60	64	68	72	74	79
9	11	17	19	24	28	32	67	71	75	80	82	88
10	13	19	23	27	32	37	73	78	83	87	91	97
11	16	22	26	31	35	41	80	86	90	95	99	105
12	18	25	29	34	39	45	87	93	98	103	107	114
13	21	28	32	38	43	49	94	100	105	111	115	122
14	23	31	35	41	47	53	101	107	113	119	123	131
15	25	34	38	45	51	58	107	114	120	127	131	140
16	28	37	42	48	55	62	114	121	128	134	139	148
17	30	40	45	52	58	66	121	129	135	142	147	157
18	33	43	48	56	62	70	128	136	142	150	155	165
19	35	46	51	59	66	74	135	143	150	158	163	174
20	38	49	54	63	70	79	141	150	157	166	171	182

Tabela 9 (Continuação).

$n_1 = 12$	$P(U \leq u_c)$ aproximadamente igual a												
	n_2	0,001	0,005	0,01	0,025	0,05	0,1	0,9	0,95	0,975	0,99	0,995	0,999
2	0	0	0	2	3	5	19	21	22	24	24	24	24
3	0	2	3	5	6	9	27	30	31	33	34	34	36
4	1	4	6	8	10	13	35	38	40	42	44	44	47
5	3	7	9	12	14	18	42	46	48	51	53	53	57
6	5	10	12	15	18	22	50	54	57	60	62	62	67
7	8	13	15	19	22	27	57	62	65	69	71	71	76
8	10	16	18	23	27	31	65	69	73	78	80	80	86
9	13	19	22	27	31	36	72	77	81	86	89	89	95
10	15	22	25	30	35	40	80	85	90	95	98	98	105
11	18	25	29	34	39	45	87	93	98	103	107	107	114
12	21	28	32	38	43	50	94	101	106	112	116	116	123
13	24	32	36	42	48	54	102	108	114	120	124	124	132
14	26	35	39	46	52	59	109	116	122	129	133	133	142
15	29	38	43	50	56	64	116	124	130	137	142	142	151
16	32	42	47	54	61	68	124	131	138	145	150	150	160
17	35	45	50	58	65	73	131	139	146	154	159	159	169
18	38	48	54	62	69	78	138	147	154	162	168	168	178
19	41	52	57	66	73	82	146	155	162	171	176	176	187
20	43	55	61	70	78	87	153	162	170	179	185	185	197

Tabela 9 (Continuação).

$n_1 = 13$	$P(U \leq u_c)$ aproximadamente igual a											
n_2	0,001	0,005	0,01	0,025	0,05	0,1	0,9	0,95	0,975	0,99	0,995	0,999
2	0	0	1	2	3	5	21	23	24	25	26	26
3	0	2	3	5	7	10	29	32	34	36	37	39
4	2	4	6	9	11	14	38	41	43	46	48	50
5	4	8	10	13	16	19	46	49	52	55	57	61
6	6	11	13	17	20	24	54	58	61	65	67	72
7	9	14	17	21	25	29	62	66	70	74	77	82
8	12	18	21	25	29	34	70	75	79	83	86	92
9	15	21	24	29	34	39	78	83	88	93	96	102
10	18	25	28	34	38	44	86	92	96	102	105	112
11	21	28	32	38	43	49	94	100	105	111	115	122
12	24	32	36	42	48	54	102	108	114	120	124	132
13	27	35	40	46	52	59	110	117	123	129	134	142
14	30	39	44	51	57	64	118	125	131	138	143	152
15	33	43	48	55	62	69	126	133	140	147	152	162
16	36	46	52	60	66	75	133	142	148	156	162	172
17	39	50	56	64	71	80	141	150	157	165	171	182
18	43	54	60	68	76	85	149	158	166	174	180	191
19	46	58	64	73	81	90	157	166	174	183	189	201
20	49	61	68	77	85	95	165	175	183	192	199	211

Tabela 9 (Continuação).

$n_1 = 14$	$P(U \leq u_c)$ aproximadamente igual a											
n_2	0,001	0,005	0,01	0,025	0,05	0,1	0,9	0,95	0,975	0,99	0,995	0,999
2	0	0	1	2	4	5	23	24	26	27	28	28
3	0	2	3	6	8	11	31	34	36	39	40	42
4	2	5	7	10	12	16	40	44	46	49	51	54
5	4	8	11	14	17	21	49	53	56	59	62	66
6	7	12	14	18	22	26	58	62	66	70	72	77
7	10	16	18	23	27	32	66	71	75	80	82	88
8	13	19	23	27	32	37	75	80	85	89	93	99
9	16	23	27	32	37	42	84	89	94	99	103	110
10	20	27	31	37	42	48	92	98	103	109	113	120
11	23	31	35	41	47	53	101	107	113	119	123	131
12	26	35	39	46	52	59	109	116	122	129	133	142
13	30	39	44	51	57	64	118	125	131	138	143	152
14	33	43	48	56	62	70	126	134	140	148	153	163
15	37	57	52	60	67	75	135	143	150	158	163	173
16	40	51	57	65	72	81	143	152	159	167	173	184
17	44	55	61	70	78	86	152	160	168	177	183	194
18	47	59	66	75	83	92	160	169	177	186	193	205
19	51	64	70	79	88	98	168	178	187	196	202	215
20	55	68	74	84	93	103	177	187	196	206	212	225

Tabela 9 (Continuação).

$n_1 = 15$	$P(U \leq u_c)$ aproximadamente igual a											
n_2	0,001	0,005	0,01	0,025	0,05	0,1	0,9	0,95	0,975	0,99	0,995	0,999
2	0	0	1	2	4	6	24	26	28	29	30	30
3	0	3	4	6	8	11	34	37	39	41	42	45
4	2	6	8	11	13	17	43	47	49	52	54	58
5	5	9	12	15	19	23	52	56	60	63	66	70
6	8	13	16	20	24	28	62	66	70	74	77	82
7	11	17	20	25	29	34	71	76	80	85	88	94
8	15	21	25	30	34	40	80	86	90	95	99	105
9	18	25	29	35	40	46	89	95	100	106	110	117
10	22	30	34	40	45	52	98	105	110	116	120	128
11	25	34	38	45	51	58	107	114	120	127	131	140
12	29	38	43	50	56	64	116	124	130	137	142	151
13	33	43	48	55	62	69	126	133	140	147	152	162
14	37	47	52	60	67	75	135	143	150	158	163	173
15	41	52	57	65	73	81	144	152	160	168	173	184
16	44	56	62	71	78	87	153	162	169	178	184	196
17	48	61	67	76	84	93	162	171	179	188	194	207
18	52	65	71	81	89	99	171	181	189	199	205	218
19	56	70	76	86	95	105	180	190	199	209	215	229
20	60	74	81	91	101	111	189	199	209	219	226	240

Tabela 9 (Continuação).

$n_1 = 16$	$P(U \leq u_c)$ aproximadamente igual a											
	n_2	0,001	0,005	0,01	0,025	0,05	0,1	0,9	0,95	0,975	0,99	0,995
2	0	0	1	2	4	6	26	28	30	31	32	32
3	0	3	4	7	9	12	36	39	41	44	45	48
4	3	6	8	12	15	18	46	49	52	56	58	61
5	6	10	13	16	20	24	56	60	64	67	70	74
6	9	14	17	22	26	30	66	70	74	79	82	87
7	12	19	22	27	31	37	75	81	85	90	93	100
8	16	23	27	32	37	43	85	91	96	101	105	112
9	20	28	32	38	43	49	95	101	106	112	116	124
10	24	32	37	43	43	55	105	111	117	123	128	136
11	28	37	42	48	55	62	114	121	128	134	139	148
12	32	42	47	54	61	68	124	131	138	145	150	160
13	36	46	52	60	66	75	133	142	148	156	162	172
14	40	51	57	65	72	81	143	152	159	167	173	184
15	44	56	62	71	78	87	153	162	169	178	184	196
16	49	61	67	76	84	94	162	172	180	189	195	207
17	53	66	72	82	90	100	172	182	190	200	206	219
18	57	71	77	87	96	107	181	192	201	211	217	231
19	61	75	83	93	102	113	191	202	211	221	229	243
20	66	80	88	99	108	120	200	212	221	232	240	254

Tabela 9 (Continuação).

$n_1 = 17$	$P(U \leq u_c)$ aproximadamente igual a											
n_2	0,001	0,005	0,01	0,025	0,05	0,1	0,9	0,95	0,975	0,99	0,995	0,999
2	0	0	1	3	4	7	27	30	31	33	34	34
3	1	3	5	7	10	13	38	41	44	46	48	50
4	3	7	9	12	16	19	49	52	56	59	61	65
5	6	11	14	18	21	26	59	64	67	71	74	79
6	10	16	19	23	27	32	70	75	79	83	86	92
7	14	20	24	29	34	39	80	85	90	95	99	105
8	18	25	29	35	40	46	90	96	101	107	111	118
9	22	30	34	40	46	53	100	107	113	119	123	131
10	26	35	39	46	52	59	111	118	124	131	135	144
11	30	40	45	52	58	66	121	129	135	142	147	157
12	35	45	50	58	65	73	131	139	146	154	159	169
13	39	50	56	64	71	80	141	150	157	165	171	182
14	44	55	61	70	78	86	152	160	168	177	183	194
15	48	61	67	76	84	93	162	171	179	188	194	207
16	53	66	72	82	90	100	172	182	190	200	206	219
17	58	71	78	88	97	107	182	192	201	211	218	231
18	62	76	83	94	103	114	192	192	192	203	230	244
19	67	82	89	100	110	121	202	213	223	234	241	256
20	71	87	94	106	116	128	212	224	234	246	253	269

Tabela 9 (Continuação).

$n_1 = 18$	$P(U \leq u_c)$ aproximadamente igual a											
	n_2	0,001	0,005	0,01	0,025	0,05	0,1	0,9	0,95	0,975	0,99	0,995
2	0	0	1	3	5	7	29	31	33	35	36	36
3	1	3	5	8	10	14	40	44	46	49	51	53
4	4	7	10	13	17	21	51	55	59	62	65	68
5	7	12	15	19	23	28	62	67	71	75	78	83
6	11	17	20	25	29	35	73	79	83	88	91	97
7	15	22	25	31	36	42	84	90	95	101	104	111
8	19	27	31	37	42	49	95	102	107	113	117	125
9	24	32	37	43	49	56	106	113	119	125	130	138
10	28	38	42	49	56	63	117	124	131	138	142	152
11	33	43	48	56	62	70	128	136	142	150	155	165
12	38	48	54	62	69	78	138	147	154	162	168	178
13	43	54	60	68	76	85	149	158	166	174	180	191
14	47	59	66	75	83	92	160	169	177	186	193	205
15	52	65	71	81	89	99	171	181	189	199	205	218
16	57	71	77	87	96	107	181	192	210	211	217	231
17	62	76	83	94	103	114	192	203	212	223	230	244
18	67	82	89	100	110	121	203	214	224	235	242	257
19	72	88	95	107	117	129	213	225	235	247	254	270
20	77	93	101	113	124	136	224	236	247	259	267	283

Tabela 9 (Continuação).

$n_1 = 19$	$P(U \leq u_c)$ aproximadamente igual a											
n_2	0,001	0,005	0,01	0,025	0,05	0,1	0,9	0,95	0,975	0,99	0,995	0,999
2	0	1	2	3	5	8	30	33	35	36	37	38
3	1	4	5	8	11	15	42	46	49	52	53	56
4	4	8	10	14	18	22	54	58	62	66	68	72
5	8	13	16	20	24	29	66	71	75	79	82	87
6	12	18	21	26	31	37	77	83	88	93	96	102
7	16	23	27	33	38	44	89	95	100	106	110	117
8	21	29	33	39	45	52	100	107	113	119	123	131
9	26	34	39	46	52	59	112	119	125	132	137	145
10	30	40	45	53	59	67	123	131	137	145	150	160
11	35	46	51	59	66	74	135	143	150	158	163	174
12	41	52	57	66	73	82	146	155	162	171	176	187
13	46	58	64	73	81	90	157	166	174	183	189	201
14	51	64	70	79	88	98	168	178	187	196	202	215
15	56	70	76	86	95	105	180	190	199	209	215	229
16	61	75	83	93	102	113	191	202	211	221	229	243
17	67	82	89	100	110	121	202	213	223	234	241	256
18	72	88	95	107	117	129	213	225	235	247	254	270
19	78	94	102	114	124	136	225	237	247	259	267	283
20	83	100	108	120	131	144	236	249	260	272	280	297

Tabela 9 (Continuação).

$n_1 = 20$	$P(U \leq u_c)$ aproximadamente igual a											
	n_2	0,001	0,005	0,01	0,025	0,05	0,1	0,9	0,95	0,975	0,99	0,995
2	0	1	2	3	5	8	32	35	37	38	39	40
3	1	4	6	9	12	16	44	48	51	54	56	59
4	4	9	11	15	19	23	57	61	65	69	71	76
5	8	14	17	21	26	31	69	74	79	83	86	92
6	13	19	23	28	33	39	81	87	92	97	101	107
7	17	25	29	35	40	47	93	100	105	111	115	123
8	22	31	35	42	48	55	105	112	118	125	129	138
9	27	37	41	49	55	63	117	125	131	139	143	153
10	33	43	48	56	63	71	129	137	144	152	157	167
11	38	49	54	63	70	79	141	150	157	166	171	182
12	43	55	61	70	78	87	153	162	170	179	185	197
13	49	61	68	77	85	95	165	175	183	192	199	211
14	55	68	74	84	93	103	177	187	196	206	212	225
15	60	74	81	91	101	111	189	199	209	219	226	240
16	66	80	88	99	108	120	200	212	221	232	240	254
17	71	87	94	106	116	128	212	224	234	246	253	269
18	77	93	101	113	124	136	224	236	247	259	267	283
19	83	100	108	120	131	144	236	249	260	272	280	297
20	89	106	115	128	139	152	248	261	272	285	294	311

Tabela 10 *Valor absoluto mínimo para o coeficiente de correlação r de Pearson ser significativo.*

	Nível de significância, α, num teste unilateral					
	0,100	0,050	0,025	0,010	0,005	0,001
n	Nível de significância, α, num teste bilateral					
	0,200	0,100	0,050	0,020	0,010	0,002
5	0,687	0,805	0,878	0,934	0,959	0,986
6	0,608	0,729	0,811	0,882	0,917	0,963
7	0,551	0,669	0,754	0,833	0,875	0,935
8	0,507	0,621	0,707	0,789	0,834	0,905
9	0,472	0,582	0,666	0,750	0,798	0,875
10	0,443	0,549	0,632	0,715	0,765	0,847
11	0,419	0,521	0,602	0,685	0,735	0,820
12	0,398	0,497	0,576	0,658	0,708	0,795
13	0,380	0,476	0,553	0,634	0,684	0,772
14	0,365	0,458	0,532	0,612	0,661	0,750
15	0,351	0,441	0,514	0,592	0,641	0,730
16	0,338	0,426	0,497	0,574	0,623	0,711
17	0,327	0,412	0,482	0,558	0,606	0,694
18	0,317	0,400	0,468	0,543	0,590	0,678
19	0,308	0,389	0,456	0,529	0,575	0,662
20	0,299	0,378	0,444	0,516	0,561	0,648
21	0,291	0,369	0,433	0,503	0,549	0,635
22	0,284	0,360	0,423	0,492	0,537	0,622
23	0,277	0,352	0,413	0,482	0,526	0,610
24	0,271	0,344	0,404	0,472	0,515	0,599
25	0,265	0,337	0,396	0,462	0,505	0,588
26	0,260	0,330	0,388	0,453	0,496	0,578
27	0,255	0,323	0,381	0,445	0,487	0,568
28	0,250	0,317	0,374	0,437	0,479	0,559
29	0,245	0,311	0,367	0,430	0,471	0,550
30	0,241	0,306	0,361	0,423	0,463	0,541

Tabela 10 (Continuação).

n	Nível de significância, α, num teste bilateral					
	0,200	0,100	0,050	0,020	0,010	0,002
35	0,222	0,283	0,334	0,392	0,430	0,504
40	0,207	0,264	0,312	0,367	0,403	0,474
45	0,195	0,248	0,294	0,346	0,380	0,449
50	0,184	0,235	0,279	0,328	0,361	0,427
60	0,168	0,214	0,254	0,300	0,330	0,391
70	0,155	0,198	0,235	0,278	0,306	0,363
80	0,145	0,185	0,220	0,260	0,286	0,340
90	0,136	0,174	0,207	0,245	0,270	0,322
100	0,129	0,165	0,197	0,232	0,256	0,305

Nota: Tabela construída da estatística $t = r \cdot \sqrt{n-2} / \sqrt{1-r^2}$ que tem distribuição t de Student com $gl = n - 2$, sob as suposições de os dados terem distribuição normal e a correlação ser linear.



Bibliografia

BARBETTA, P. A. *Estatística aplicada às ciências sociais*. 5. ed. Florianópolis: Editora da UFSC, 2002.

BOX, G. E. P.; HUNTER, W. G.; HUNTER, J. S. *Statistics for experimenters*. New York: John Wiley, 1978.

BUSSAB, W. O.; MORETTIN, P. A. *Estatística básica*. 5. ed. São Paulo: Saraiva, 2002.

COCHRAN, W. G. *Sampling techniques*. 3. ed. New York: John Wiley, 1977.

CONOVER, W. J. *Practical nonparametric statistics*. 2. ed. New York: John Wiley, 1980.

FISHER, R. A. *The design of experiments*. 6. ed. Edimburgo: Oliver and Boyd, 1951.

JAIN, R. *The art of computer systems performance analysis: techniques for experimental design, measurement, simulation, and modeling*. New York: John Wiley, 1991.

MENDENHALL, N. *Probabilidade e estatística*. Rio de Janeiro: Campos, 1985. v. 1 e 2.

MAGALHÃES, A. N.; LIMA, A. C. P. *Noções de probabilidade e estatística*. 2. ed. São Paulo: IME-USP, 1999.

MONTGOMERY, D. C.; RUNGER, G. C. *Estatística aplicada e probabilidade para engenheiros*. Rio de Janeiro: LTC, 2003.

MONTGOMERY, D. C. *Design and analysis of experiments*. 4. ed. New York: John Wiley, 1997.

SPRENT, P. *Applied nonparametric statistical methods*. Bristol: Chapman and Hall, 1989.

STIGLER, S. M. *The history of statistics: the measurement of uncertainty before 1900*. Cambridge: Harvard University Press, 1986.

ESTATÍSTICA

para Cursos de

Engenharia e Informática

Este livro enfatiza a relação dos métodos estatísticos com o planejamento e desenvolvimento de pesquisas em Engenharia e Informática, o que diferencia da maioria dos livros-texto de Estatística, que tem uma preocupação excessiva com os aspectos matemáticos dos métodos estatísticos. Aqui, os métodos estatísticos são apresentados com muitas ilustrações, de forma simples e intuitiva. Os exemplos de motivação e exercícios são usualmente associados a problemas típicos de Engenharia e Informática, tornando o aprendizado mais agradável e motivador aos estudantes dessas áreas.

O texto mostra como a estatística é importante na vida do engenheiro ou do profissional em informática, apresenta alguns conceitos básicos e oferece visão geral dos planos amostrais e projetos de experimentos. Também desenvolve os tópicos clássicos da Estatística: análise exploratória de dados, modelos probabilísticos, teoria da estimação e testes de hipóteses. Os três últimos capítulos descrevem as análises estatísticas para comparação de tratamentos, procedimentos não paramétricos e análise de regressão, ilustrando com problemas reais de pesquisa.

Aplicação

Livro-texto para a disciplina *Estatística* dos cursos de graduação e pós-graduação em Engenharia (todas as áreas), Ciência da Computação e Sistemas de Informação. Material de consulta para pós-graduandos, pesquisadores, estatísticos, engenheiros e profissionais da Informática.

publicação atlas

www.EditoraAtlas.com.br