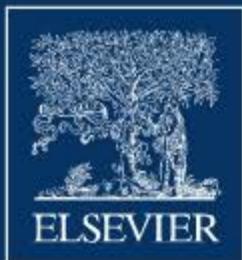


S O N I A V I E I R A

INTRODUÇÃO À
Bioestatística



4^a EDIÇÃO

AVISO LEGAL

Caso esta Obra na versão impressa possua quaisquer materiais complementares, tais como: CDs e/ou DVDs ou recursos on-line, estes serão disponibilizados na versão adquirida a partir da Biblioteca Digital através do ícone "Recursos Extras" dentro da própria Biblioteca Digital.

Introdução à Bioestatística

4^a EDIÇÃO

Sonia Vieira

Professora Titular de Bioestatística da Unicamp



ELSEVIER

© 2008 Elsevier Editora Ltda.

Todos os direitos reservados e protegidos pela Lei 9.610 de 19/02/1998.

Nenhuma parte deste livro, sem autorização prévia por escrito da editora, poderá ser reproduzida ou transmitida sejam quais forem os meios empregados: eletrônicos, mecânicos, fotográficos, gravação ou quaisquer outros.

Capa

Folio Design

Editoração Eletrônica

Rosane Guedes

Elsevier Editora Ltda.

Rua Sete de Setembro, 111 - 16º andar
20050-006 - Centro - Rio de Janeiro - RJ - Brasil
Telefone: (21) 3970-9300 - Fax: (21) 2507-1991
E-mail: info@elsevier.com.br

Escritório São Paulo

Rua Quintana, 753 - 8º andar
04569-011 - Brooklin - São Paulo - SP - Brasil
Telefone: (11) 5105-8555

Conheça nosso catálogo completo: cadastre-se em www.elsevier.com.br para ter acesso a conteúdos e serviços exclusivos e receber informações sobre nossos lançamentos e promoções.

NOTA

O conhecimento médico está em permanente mudança. Os cuidados normais de segurança devem ser seguidos, mas, como as novas pesquisas e a experiência clínica ampliam nosso conhecimento, alterações no tratamento e terapia à base de drogas podem ser necessárias ou apropriadas. Os leitores são aconselhados a checar informações mais atuais dos produtos, fornecidas pelos fabricantes de cada droga a ser administrada, para verificar a dose recomendada, o método e a duração da administração e as contraindicações. É responsabilidade do médico, com base na experiência e contando com o conhecimento do paciente, determinar as dosagens e o melhor tratamento para cada um individualmente. Nem o editor nem o autor assumem qualquer responsabilidade por eventual dano ou perda a pessoas ou a propriedade originada por esta publicação.

O Editor

ISBN: 978-85-352-5012-1

CIP-BRASIL. CATALOGAÇÃO-NA-FONTE
SINDICATO NACIONAL DOS EDITORES DE LIVROS, RJ

V718i

Vieira, Sonia, 1942-

Introdução à bioestatística [recurso eletrônico] / Sonia Vieira. - Rio de Janeiro : Elsevier, 2011.
345 p., recurso digital : il. ;

Formato: Flash

Requisitos do sistema: Adobe Flash Player

Modo de acesso: Word Wide Web

Apêndice

Inclui bibliografia e Índice

ISBN 978-85-352-5012-1 (recurso eletrônico)

1. Bioestatística. 2. Livros eletrônicos. I. Título.

11.7080.

CDD: 570.151.95

CDU: 57.087.1

Prefácio

Bioestatística é a Estatística aplicada às ciências da saúde. Profissionais e alunos dessas áreas querem aprender técnicas estatísticas porque elas são muito usadas na pesquisa, como bem mostra a literatura especializada. Mas Estatística é ciência complexa, que não se aprende com a simples busca de um termo na Internet. É difícil aprender Estatística? Sim e não. Aprender a fazer cálculos estatísticos usando programas de computador não é difícil, embora exija tempo, interesse e atenção. Entretanto, a condução e a avaliação de uma pesquisa dependem, em boa parte, do conhecimento do pesquisador sobre as potencialidades e as limitações das técnicas utilizadas. E entre o cálculo e a interpretação do resultado há um caminho a percorrer.

Este livro foi, então, escrito e reescrito muitas vezes, na tentativa de facilitar a aprendizagem. Buscamos explicar sempre a indicação e as restrições das técnicas ensinadas. Os conceitos são transmitidos mais pela intuição do que por demonstração, os exemplos são simples e das áreas da saúde e os exercícios exigem pouco trabalho de cálculo. É grande a quantidade de exemplos e o número de exercícios mais do que dobrou em relação à edição anterior, para bem ilustrar as técnicas aprendidas.

A leitura do texto exige os conhecimentos de matemática que são exigidos em exames vestibulares. De qualquer modo, as seções que envolvem maior aptidão para a matemática foram assinaladas com asterisco. Tais seções podem ser evitadas sem prejuízo do entendimento das subsequentes. Os cálculos podem ser feitos à mão ou com calculadora. Alunos de cursos avançados de Estatística usam, rotineiramente, um computador, mas acreditamos que é preciso manusear fórmulas para entender os conceitos básicos de Estatística. Não há como ter completa segurança na discussão de uma média aritmética, por exemplo, sem nunca ter usado papel e lápis para calcular esse tipo de estatística. Assim, sem despendar muito tempo com cálculos e demonstrações — o estudante adquire, neste livro, conhecimentos suficientes para tornar-se usuário competente das técnicas estatísticas mais comuns.

Uma consequência importante de aprender Estatística — mais importante do que possa parecer à primeira vista — é a familiarização com o jargão próprio da área. Alguns termos do vocabulário comum têm significado técnico e específico, quando usados em Estatística. É claro que o conhecimento do significado comum ajuda, mas pode conduzir à interpretação errada quando substitui o significado técnico.

Essa 4^a edição de *Introdução à Bioestatística*, totalmente revista e ampliada, só foi possível porque o livro encontrou aceitação no meio acadêmico. Agradecemos,

pois, a todos aqueles que prestigiaram nosso trabalho, mas principalmente aos alunos, que nos ensinaram a ensinar. Importante, porém, é o fato de esse livro ter tido a revisão competente e altamente especializada de Martha Maria Mischan. Ronaldo Wada fez alguns dos vários gráficos, Márcio Vieira Hoffmann fez uma leitura crítica dos originais e William Saad Hossne escreveu a 4^a capa. Mas há também que agradecer ao Centro de Pós-Graduação São Leopoldo Mandic, pela oportunidade de trabalho.

A autora

Sumário

CAPÍTULO 1 NOÇÕES SOBRE AMOSTRAGEM 1

- 1.1 – O que é Estatística 3
- 1.2 – O que é população e o que é amostra? 4
- 1.3 – Por que se usam amostras? 4
- 1.4. – Como se obtém uma amostra? 5
 - 1.4.1 – Amostra aleatória ou probabilística 5
 - 1.4.2 – Amostra semiprobabilística 6
 - 1.4.2.1 – Amostra sistemática 7
 - 1.4.2.2 – Amostra por conglomerados 7
 - 1.4.2.3 – Amostra por quotas 8
 - 1.4.3 – Amostra não probabilística ou de conveniência 9
 - 1.4.4 – Avaliação das técnicas de amostragem 9
- 1.5 – Estatísticas e parâmetros 10
- 1.6 – Com quantas unidades se compõe uma amostra? 11
- 1.7 – A questão da representatividade 13
- 1.8 – Exercícios resolvidos 14
- 1.9 – Exercícios propostos 17

CAPÍTULO 2 APRESENTAÇÃO DE DADOS EM TABELAS 21

- 2.1 – Dados e variáveis 23
- 2.2 – Apuração de dados 24
- 2.3 – Componentes das tabelas 26
- 2.4 – Apresentação de dados qualitativos 28
- 2.5 – Tabelas de contingência 30
- 2.6 – Apresentação de dados numéricos 31
- 2.7 – Exercícios resolvidos 38
- 2.8 – Exercícios propostos 41

CAPÍTULO 3 APRESENTAÇÃO DE DADOS EM GRÁFICOS 47

- 3.1 – Apresentação de dados qualitativos 49
 - 3.1.1 – Gráficos de Barras 49
 - 3.1.2 – Gráfico de setores 54
- 3.2 – Apresentação de dados numéricos 56
 - 3.2.1 – Diagrama de linhas 56
 - 3.2.2 – Gráfico de pontos 57
 - 3.2.3 – Histograma 57
 - 3.2.4 – Polígono de freqüências 58
- 3.3 – Observações 59
- 3.4 – Exercícios resolvidos 60
- 3.5 – Exercícios propostos 62

CAPÍTULO 4 MEDIDAS DE TENDÊNCIA CENTRAL 65

- 4.1 – Símbolos matemáticos 67
- 4.2 – Média da amostra 68
- 4.3 – Mediana da amostra 74
- 4.4 – Moda da amostra 75
- 4.5 – Exercícios resolvidos 77
- 4.6 – Exercícios propostos 80

CAPÍTULO 5 MEDIDAS DE DISPERSÃO PARA UMA AMOSTRA 85

- 5.1 – Mínimo, máximo e amplitude 87
- 5.2 – Quartil 89
 - 5.2.1 – Diagrama de caixa (*Box plot*) 91
- 5.3 – Desvio padrão da amostra 93
 - 5.3.1 – Introduzindo a variância 93
 - 5.3.2 – Definindo o desvio padrão 95
 - 5.3.3 – Uma fórmula prática para calcular a variância 97
- 5.4 – Coeficiente de variação 98
- 5.5 – Exercícios resolvidos 99
- 5.6 – Exercícios propostos 104

CAPÍTULO 6 NOÇÕES SOBRE CORRELAÇÃO 107

- 6.1 – Diagrama de dispersão 109
- 6.2 – Coeficiente de correlação 115
- 6.3 – Pressuposições 119
- 6.4 – Cuidados na interpretação do coeficiente de correlação 119
- 6.5 – Exercícios resolvidos 120
- 6.6 – Exercícios propostos 124

CAPÍTULO 7 NOÇÕES SOBRE REGRESSÃO 131

- 7.1 – Gráfico de linhas 133
- 7.2 – Reta de regressão 135
- 7.3 – Escolha da variável explanatória 142
- 7.4 – Coeficiente de determinação 143
- 7.5 – Uma pressuposição básica 145
- 7.6 – Outros tipos de regressão 147
- 7.7 – Exercícios resolvidos 151
- 7.8 – Exercícios propostos 155

CAPÍTULO 8 NOÇÕES SOBRE PROBABILIDADE 161

- 8.1 – Definição clássica de probabilidade 163
- 8.2 – Freqüência relativa como estimativa de probabilidade 164
- 8.3 – Eventos mutuamente exclusivos e eventos independentes 166
 - 8.3.1. – Eventos mutuamente exclusivos 166
 - 8.3.2 – Eventos independentes 166
 - 8.3.2.1 – Conjuntos 166
 - 8.3.2.2 – Condição de independência 167
 - 8.3.2.3 – Diferença nos conceitos 170
- 8.4 – Probabilidade condicional 170
 - *8.5 – Teorema da soma ou a regra do “ou” 173
 - *8.6 – Teorema do produto ou a regra do “e” 174
- 8.7 – Exercícios resolvidos 176
- 8.8 – Exercícios propostos 180

CAPÍTULO 9 DISTRIBUIÇÃO BINOMIAL 183

- 9.1 – Variável aleatória 185
 - 9.1.1 – Variável aleatória binária 186
 - 9.1.2 – Variável aleatória binomial 186
- 9.2 – Distribuição de probabilidades 187
- 9.3 – Distribuição binomial 189
 - 9.3.1 – Caracterização da distribuição binomial 192
 - *9.3.2 – Função de distribuição na distribuição binomial 192
 - *9.3.3 – Média e variância na distribuição binomial 194
- 9.4 – Revisão sobre análise combinatória 195
- 9.5 – Exercícios resolvidos 195
- 9.6 – Exercícios propostos 202

CAPÍTULO 10 DISTRIBUIÇÃO NORMAL 205

- 10.1 – Características da distribuição normal 209
- *10.2 – Distribuição normal reduzida 213
- *10.3 – Probabilidades na distribuição normal 216
- 10.4 – Usos da distribuição normal 219
- 10.5 – Exercícios resolvidos 221
- 10.6 – Exercícios propostos 224

CAPÍTULO 11 INTERVALO DE CONFIANÇA 227

- 11.1 – Intervalo de confiança para uma proporção 230
 - 11.1.1 – Cálculo do intervalo de confiança para uma proporção
 - 11.1.2 – Pressuposições 231
 - 11.1.3 – A margem do erro 232
- 11.2 – Intervalos de confiança para uma média 233
 - 11.2.1 – Erro padrão da média 233
 - 11.2.2 – Cálculo do intervalo de confiança para uma média 236
- 11.3 – Cuidados na interpretação dos intervalos de confiança 237
- 11.4 – Pequenas amostras 237
- 11.5 – Exercícios resolvidos 240
- 11.6 – Exercícios propostos 242

CAPÍTULO 12 TESTE DE QUI-QUADRADO 245

12.1 – Teste de χ^2 de Pearson para aderência 252

 12.1.1 – Resumo do procedimento 255

12.2 – Tabelas 2 x 2 256

 12.2.1 – Teste de χ^2 para independência 256

 12.2.2 – Usos e restrições do teste de χ^2 258

 12.2.3 – Medida de associação 259

12.3 – Exercícios resolvidos 260

12.4 – Exercícios propostos 265

CAPÍTULO 13 TESTE t DE STUDENT 269

13.1 – O teste t nos estudos com dados pareados 272

 13.1.1 – Testes unilaterais e testes bilaterais 276

13.2 – O teste t na comparação de dois grupos independentes 279

 13.2.1 – O caso das variâncias desiguais 281

13.3 – O teste t para o coeficiente de correlação 285

13.4 – Exercícios resolvidos 286

13.5 – Exercícios propostos 290

Respostas aos Exercícios Propostos 295

Tabelas 325

Sugestões para leitura 341

Índice Remissivo 343

(página deixada intencionalmente em branco)

Noções sobre Amostragem

1

(página deixada intencionalmente em branco)

Grande parte das pessoas que conhecemos já ouviu falar de prévias eleitorais, de censo, de pesquisa de opinião. A maioria das pessoas que conhecemos já respondeu perguntas sobre a qualidade dos serviços de um bar ou de uma lanchonete, já assistiu no rádio ou na televisão programas em que pedem para o ouvinte ou telespectador votar em um cantor ou em uma música, ou dar opinião sobre determinado assunto por telefone ou por e-mail.

O uso tão difundido de *levantamento de dados* — que no Brasil chama- mos popularmente de “pesquisa” — faz pensar que esse é um trabalho fácil. Por conta disso, ao ler um relatório de pesquisa no jornal da cidade, muita gente se acha capaz de fazê-lo, e até melhor, pois entende que, para levantar dados, basta fazer perguntas e depois contar as respostas. Mas não é bem assim. Um bom *levantamento de dados* exige conhecimentos de Estatística.

1.1 – O QUE É ESTATÍSTICA?

Para muitas pessoas, a palavra Estatística lembra números. Elas têm razão em parte: a Estatística trata de números, mas trata, também, de outras coisas.

Estatística é a ciência que fornece os princípios e os métodos para coleta, organização, resumo, análise e interpretação de dados.

Dados corretamente coletados fornecem conhecimentos que não seriam obtidos por simples especulação. Mas nem sempre é possível levantar *todos* os dados. Um exemplo disso são as prévias eleitorais, que fornecem as estimativas da porcentagem de votos em cada candidato. As prévias são feitas regularmente e publicadas. Mas quem são as pessoas que os institutos de pesquisa devem entrevistar?

Se estivermos pensando em eleições presidenciais, a idéia seria entrevistar *todos* os portadores de título de eleitor do Brasil. Mas como as prévias eleitorais são feitas com freqüência, não é possível entrevistar todos os eleitores (incluindo você e eu) a cada 10 dias, por exemplo, para conhecer as intenções de voto de todos nós. Então as prévias eleitorais são feitas com *pequeno número* de eleitores: de 1.500 a 3.000. É o que chamamos de amostra.

1.2 – O QUE É POPULAÇÃO E O QUE É AMOSTRA?

População ou universo é o conjunto de unidades sobre o qual desejamos obter informação. *Amostra* é todo subconjunto de unidades retiradas de uma população para obter a informação desejada.

É importante entender que população é o termo que os estatísticos usam para descrever um grande conjunto de unidades que têm algo em comum. Na área de saúde, a população pode ser constituída por pacientes ou por animais, mas também pode ser constituída por radiografias, por prontuários, por necropsias, por contas hospitalares, por certidões de óbito.

A distinção entre os dados realmente coletados (amostra) e a vasta quantidade de dados que poderiam ser observados (população) é a chave para o bom entendimento da Estatística. O uso de amostras permite obter respostas razoáveis, com margem de erro conhecida. Considere a questão das prévias eleitorais. Os resultados — desde que obtidos de amostras representativas — são confiáveis. Na maioria das vezes, a predição do ganhador da eleição é correta.

O levantamento de dados de toda a população chama-se *censo*. A Fundação Instituto Brasileiro de Geografia e Estatística (IBGE) faz o *Censo Demográfico do Brasil* a cada 10 anos, por exigência da Constituição da República Federativa do Brasil. São coletadas informações sobre sexo, idade e nível de renda de todos os residentes no Brasil.

1.3 – POR QUE SE USAM AMOSTRAS?

As razões que levam os pesquisadores a trabalhar com amostras — e não com toda a população — são poucas, mas absolutamente relevantes.

- Custo e demora dos censos
- Populações muito grandes
- Impossibilidade física de examinar toda a população
- Comprovado valor científico das informações coletadas por meio de amostras

A primeira razão para estudar uma amostra, em lugar de toda a população, é a questão *do custo e da demora dos censos*. Por exemplo, qual é, em média, o peso ao nascer de nascidos vivos no Brasil em determinado ano? Avaliar toda a população pode ser impossível para o pesquisador, porque levaria muito tempo e seria muito caro.

Outra razão para estudar amostras é o fato de existirem *populações tão grandes* que estudá-las por inteiro seria impossível. Por exemplo, quantos

peixes tem o mar? Esse número é, em determinado momento, matematicamente finito, mas tão grande que pode ser considerado infinito para qualquer finalidade prática. Então, quem faz pesquisas sobre peixes do mar trabalha, necessariamente, com amostras.

Outras vezes é *impossível* estudar toda a população porque o estudo destrói as unidades. Uma empresa que fabrica fósforos e queira testar a qualidade do produto que fabrica não pode acender todos os fósforos que fabricou — mas apenas alguns deles.

O uso de amostras tem, ainda, outra razão: o estudo cuidadoso de uma amostra tem maior *valor científico* do que o estudo sumário de toda a população. Imagine, como exemplo, que um pesquisador queira estudar os hábitos de consumo de bebidas alcoólicas entre adolescentes de uma grande cidade. É melhor que o pesquisador faça a avaliação criteriosa de uma amostra — do que a avaliação sumária de toda a população de adolescentes da cidade.

1.4 – COMO SE OBTÉM UMA AMOSTRA?

Antes de obter uma amostra, é preciso definir os *critérios* que serão usados para selecionar as unidades que comporão essa amostra. De acordo com a técnica usada, tem-se um tipo de amostra. Serão definidas aqui:

- amostra aleatória, casual, ou probabilística;
- amostra semiprobabilística;
- amostra não-probabilística ou de conveniência.

1.4.1 – Amostra aleatória ou probabilística

A *amostra aleatória ou probabilística* é constituída por n unidades retiradas *ao acaso* da população. Em outras palavras, a amostra aleatória é obtida por sorteio. Logo, toda unidade da população tem probabilidade conhecida de pertencer à amostra.

Para obter uma amostra aleatória, é preciso que a população seja conhecida e cada unidade esteja identificada por nome ou por número. Os elementos que constituirão a amostra são escolhidos por sorteio. Algumas pessoas acreditam que o sorteio por computador é mais “sério”, ou mais “exato”. Hoje em dia, é mais fácil. No entanto, o sorteio feito com papéis em uma caixa ou bolas em uma urna (usados em programas de televisão) ajuda entender as regras do procedimento aleatório.

Uma amostra aleatória pode ser:

- simples
- estratificada.

A amostra aleatória simples é obtida por sorteio de uma população constituída por *unidades homogêneas* para a variável que você quer estudar.

Exemplo 1.1: Uma amostra aleatória simples.

Imagine que você precisa obter uma amostra de 2% dos 500 pacientes de uma clínica para entrevistá-los sobre a qualidade de atendimento da secretaria. Qual seria o procedimento para obter uma amostra aleatória simples?

Solução

Para obter uma amostra aleatória de 2% dos 500 pacientes, você precisa sortear 10. Você pode fazer isso da maneira mais antiga e conhecida (e também a mais trabalhosa). Comece escrevendo o nome de todos os pacientes em pedaços de papel. Coloque todos os pedaços de papel em uma urna, misture bem e retire um nome. Repita o procedimento até ter os nomes dos 10 pacientes que comporão sua amostra.

A amostra aleatória estratificada é usada quando a população é constituída por *unidades heterogêneas* para a variável que se quer estudar. Nesse caso, as unidades da população devem ser identificadas; depois, as unidades similares devem ser reunidas em subgrupos chamados *estratos*. O sorteio é feito dentro de cada estrato.

Exemplo 1.2: Uma amostra estratificada

Imagine que você precisa obter uma amostra de 2% dos 500 pacientes de uma clínica para entrevistá-los sobre a qualidade de atendimento da secretaria. Você suspeita que homens sejam mais bem atendidos do que mulheres. Aproximadamente metade dos pacientes é do sexo masculino. Você quer obter dados dos dois sexos. Qual seria o procedimento?

Solução

Comece separando homens de mulheres. Você tem, então, dois estratos, um de homens, outro de mulheres. Depois você obtém uma amostra aleatória de cada sexo (ou cada estrato) e reúne os dados dos dois estratos numa só amostra aleatória estratificada.

1.4.2 – Amostra semiprobabilística

A amostra semiprobabilística é constituída por n unidades retiradas da população por procedimento parcialmente aleatório. Dentre as amostras semiprobabilísticas, temos:

- amostra sistemática;

- amostra por conglomerados;
- amostra por quotas.

1.4.2.1 – Amostra sistemática

A *amostra sistemática* é constituída por n unidades retiradas da população segundo um sistema preestabelecido. Por exemplo, se você quiser uma amostra constituída por $1/8$ da população, você sorteia um número que caia entre 1 e 8. Se for sorteado o número 3, por exemplo, a terceira unidade (número 3) será selecionada para a amostra. A partir daí, tome, *sistematicamente*, a terceira unidade de cada oito, em seqüência. No caso do exemplo, a primeira unidade é 3. Seguem, de oito em oito, as unidades de números: 11, 19, 27 etc.

Exemplo 1.3: Uma amostra sistemática

Imagine que você precisa obter uma amostra de 2% dos 500 pacientes de uma clínica para entrevistá-los sobre a qualidade de atendimento da secretaria. Como você obteria uma amostra sistemática?

Solução

Uma amostra de 2% dos 500 pacientes significa amostra de tamanho 10. Para obter a amostra, você pode dividir 500 por 10, e obter 50. Sorteie então um número entre 1 e 50, inclusive. Se sair o número 27, por exemplo, esse será o número do primeiro paciente que será incluído na amostra. Depois, a partir do número 27, conte 50 e chame esse paciente. Proceda dessa forma até completar a amostra de 10 pacientes.

1.4.2.2 – Amostra por conglomerados

A *amostra por conglomerados* é constituída por n unidades tomadas de alguns *conglomerados*. O conglomerado é um conjunto de unidades que estão agrupadas, qualquer que seja a razão. Um asilo é um conglomerado de idosos, uma universidade pública é um conglomerado de pessoas com bom nível socioeconômico, um serviço militar é um conglomerado de adultos jovens saudáveis. Como exemplo, imagine que um dentista quer levantar dados sobre a necessidade de aparelho ortodôntico em crianças de 12 anos. Ele pode sortear três escolas de primeiro grau (*conglomerados*) e examinar todas as crianças com 12 anos dessas escolas.

Exemplo 1.4: Uma amostra por conglomerados

Um professor de Educação Física quer estudar o efeito da terapia de reposição hormonal (uso de hormônios por mulheres depois da menopausa) sobre o desempenho nos exercícios. Como obteria uma amostra por conglomerados?

Solução

O professor de Educação Física pode sortear duas academias de ginástica da cidade e avaliar o desempenho das mulheres que freqüentam a academia e já tiveram a menopausa (tanto as que fazem como as que não fazem uso da terapia de reposição hormonal) para posterior comparação.

1.4.2.3 – Amostra por quotas

A *amostra por quotas* é constituída por n unidades retiradas da população segundo *quotas* estabelecidas de acordo com a distribuição desses elementos na população. A idéia de quota é semelhante à de estrato, com uma diferença básica: você seleciona a amostra por julgamento e depois confirma as características das unidades amostradas.

A *amostragem por quotas* não é aleatória, embora muitos pensem que é. A grande vantagem é ser relativamente barata. Por esta razão, é muito usada em levantamentos de opinião e pesquisas de mercado.

Exemplo 1.5: Uma amostra por quotas

Considere uma pesquisa sobre a preferência de modelo de carro. Como se faz uma amostra por quotas?

Solução

Você possivelmente irá entrevistar homens e mulheres com mais de 18 anos que vivem em uma metrópole (por exemplo, Curitiba), na proporção apresentada pelo censo demográfico em termos de sexo, idade e renda. Você então sai às ruas para trabalhar com a incumbência de entrevistar determinada *quota de pessoas*, com determinadas características.

Por exemplo, você pode ser incumbido de entrevistar 30 homens com “mais de 50 anos que recebam mais de seis e menos de 10 salários mínimos”. Então você deverá julgar, pela aparência da pessoa, se ela se enquadra nas características descritas — homem de mais de 50 anos que ganha entre seis e 10 salários mínimos. Se achar que viu a pessoa certa, deve fazer a abordagem e depois confirmar as características com perguntas. O número de pessoas em determinada quota depende do número delas na população.

1.4.3 – Amostra não-probabilística ou de conveniência

A amostra *não-probabilística* ou *de conveniência* é constituída por n unidades reunidas em uma amostra simplesmente porque o pesquisador tem fácil acesso a essas unidades. Assim, o professor que toma os alunos de sua classe como amostra de toda a escola está usando uma amostra de conveniência.

Exemplo 1.6: Uma amostra não-probabilística

Imagine que um nutricionista quer entrevistar 50 mães de crianças com idades de 3 e 4 anos para conhecer os hábitos alimentares dessas crianças. Como obteria essa amostra?

Solução

Se o nutricionista trabalha em uma escola, para obter a amostra de 50 mães de crianças de 3 e 4 anos, provavelmente procurará as mães de crianças matriculadas na escola em que trabalha.

1.4.4 – Avaliação das técnicas de amostragem

As *amostras aleatórias* exigem que o pesquisador tenha a listagem com todas as unidades da população, porque é dessa listagem que serão sorteadas as unidades que comporão a amostra. Essa exigência inviabiliza a tomada de amostras aleatórias em grande parte dos casos. Por exemplo, não é possível obter uma amostra aleatória de cariocas simplesmente porque não temos uma lista com o nome de todos os cariocas.

A *amostra sistemática* não exige que a população seja conhecida, mas é preciso que esteja organizada em filas, em arquivos, ou mesmo em ruas, como os domicílios de uma cidade. Por exemplo, para tomar uma amostra dos domicílios de uma cidade, parte-se de um ponto sorteado e toma-se, de tantos em tantos, um domicílio para a amostra.

A *amostra por conglomerados* exige livre acesso aos conglomerados, o que nem sempre se consegue. Um médico pode sortear cinco hospitais da cidade de São Paulo para entrevistar pacientes internados por problemas cardíacos, mas dificilmente conseguirá permissão da diretoria de todos esses cinco hospitais para fazer sua pesquisa.

A *amostra por quotas* exige algum conhecimento da população, mas as unidades não precisam estar numeradas ou identificadas. Se você quiser uma amostra de homens e de mulheres empregados de uma grande empresa, basta saber, por exemplo, a proporção de homens e mulheres na empresa, e amostrar na mesma proporção.

De qualquer forma, as amostras probabilísticas são preferíveis do ponto de vista do estatístico, mas, na prática, elas nem sempre são possíveis. Na área de saúde, o pesquisador trabalha, necessariamente, com unidades às quais tem *acesso*: ratos de um laboratório, universitários, pacientes em tratamento no ambulatório da universidade, crianças matriculadas em escolas. As amostras de conveniência não invalidam a pesquisa, mas precisam ser *muito bem descritas* porque representam apenas a população de indivíduos semelhantes àqueles incluídos na amostra.

Por essa razão, uma enfermeira que usar os dados de um hospital para estimar a probabilidade de morte por desidratação poderá generalizar seus achados apenas para pacientes internados por desidratação. Como são internados apenas os casos graves, é possível que a mortalidade entre pacientes internados seja maior do que entre pacientes não-internados — então não teria sentido generalizar os achados para todas os pacientes com desidratação.

1.5 – ESTATÍSTICAS E PARÂMETROS

Já sabemos a diferença entre amostra e população. Precisamos agora estabelecer distinção entre valores obtidos da amostra e valores obtidos da população.

A *estatística* resume uma característica da amostra; o *parâmetro* resume uma característica da população.

Quando você ouve no noticiário que, de acordo com a pesquisa de determinado instituto, 44% dos brasileiros aprovam determinada atitude do Presidente da República, você foi apresentado a uma *estatística*. Essa estatística resume o que as pessoas que compuseram a amostra (provavelmente 1.500 ou 2.000) pensam da atitude em questão. É um indicador ou uma *estimativa* do parâmetro correspondente — a porcentagem da população brasileira que aprovou a atitude.

Mas não existe garantia de que as estatísticas (estimativas obtidas com base nos dados da amostra) tenham valor igual, ou mesmo próximo do parâmetro (valor verdadeiro na população). No entanto, isto ocorrerá na maioria das vezes — desde que a amostra tenha sido obtida de acordo com a técnica correta e tenha sido bem dimensionada (o tamanho seja adequado).

1.6 – COM QUANTAS UNIDADES SE COMPÕE UMA AMOSTRA?

Do ponto de vista do estatístico, as amostras devem ser grandes para dar maior confiança às conclusões obtidas. Para entender as razões desse ponto de vista, imagine que em uma cidade existem dois hospitais¹. Em um deles nascem, em média, 120 bebês por dia e, no outro, nascem 12. A razão de meninos para meninas é, em média, 50% nos dois hospitais.

Em uma ocasião nasceu, em um dos hospitais, duas vezes mais meninos do que meninas. Em qual dos hospitais é mais provável que isso tenha ocorrido? Para o estatístico, a resposta é óbvia: é mais provável que o fato tenha ocorrido no hospital em que nasce menor número de crianças. A probabilidade de uma estimativa desviar-se muito do parâmetro (do valor verdadeiro) é maior quando a amostra é pequena.

A “qualidade” de uma estimativa depende, em muito, do número de unidades que compõe a amostra (tamanho da amostra). No entanto, *desde que a população seja muito maior do que a amostra*, a “qualidade” da estatística não depende do tamanho da população. De qualquer modo, as amostras não devem ser muito grandes, porque isso seria perda de recursos. Também não devem ser muito pequenas, porque o resultado do trabalho seria de pouca utilidade.

Como se determina o tamanho da amostra? Na prática, o tamanho da amostra é determinado mais por considerações reais ou imaginárias a respeito do custo de cada unidade amostrada do que por técnicas estatísticas. Se seu orçamento for curto, não tente enquadrar nele uma pesquisa ambiciosa. Mas o pesquisador precisa sempre levar em conta o que é usual na área. Então você tem aqui a regra de ouro para determinar o tamanho da amostra: veja o que se faz na sua área, consultando a literatura, mas verifique também o que seu orçamento permite fazer.

De qualquer forma, o tamanho da amostra pode ser determinado por critério estatístico². As fórmulas de cálculo são bem conhecidas. Mas a aplicação dessas fórmulas exige conhecimentos acima do nível deste livro. Será apresentada aqui apenas uma equação que dará idéia do problema.

Um exemplo ajuda muito³. Imagine que um antropólogo está estudando os habitantes de uma ilha isolada e que, entre outras coisas, quer determinar a porcentagem de pessoas dessa ilha com sangue tipo O. Quantas

¹Baseado em um exemplo de KAHNEMEN, D. e TVERSKY, A., "Judgement under uncertainty: heuristics and bias", *Science* 185, 27 de setembro de 1974.

²Ver, por exemplo: 1. COCHRAN, W. **Sampling techniques**. Nova York, Wiley, 1977; 2. LOHR, S. L. **Sampling: Design and analysis**. Pacific Grove, Brooks, 1999. 3. BOLFARINE, H. e BUSSAB, W. O. **Elementos de amostragem**. São Paulo, Edgard Blucher, 2005.

³O exemplo é de COCHRAN, W. opus cited, p. 72-73.

pessoas (tamanho da amostra) devem ser examinadas? O tamanho da amostra pode ser determinado por uma equação que, no entanto, não pode ser resolvida sem resposta para algumas questões.

A primeira questão é: qual é a *margem de erro* que o antropólogo admite em seus resultados? Vamos imaginar que ele diz ficar satisfeito com uma margem de erro de $\pm 5\%$, isto é, se 43% das pessoas da amostra tiverem sangue tipo 0, a verdadeira porcentagem de pessoas com sangue tipo 0 na ilha deverá estar entre 38 e 48, ou seja, no intervalo $43\% \pm 5\%$.

Neste ponto, convém avisar o antropólogo de que, como estará trabalhando com uma só amostra, existe a chance de ele, por azar, tomar uma amostra pouco representativa. O antropólogo então concorda em admitir a probabilidade de uma amostra errada em cada 20. Isto significa que ele terá probabilidade $\left(\frac{19}{20}\right) = 0,95$ de obter a verdadeira porcentagem de sangue tipo 0 dentro do intervalo calculado. Temos então o *nível de confiança*: 95%.

Mas é preciso saber, ainda, o valor que o antropólogo espera para a porcentagem de pessoas com sangue tipo 0 na ilha. Ele diz que, com base no que sabe de outras populações, é razoável esperar que essa porcentagem esteja entre 30% e 60%. Ótimo. Admitiremos, por simplicidade, que essa porcentagem seja 50%. Podemos agora aplicar a fórmula:

$$n = \frac{z^2 p(100 - p)}{d^2}$$

em que z é um valor dado em tabelas e associado ao nível de confiança, conforme veremos no Capítulo 11 deste livro. Aproximadamente, $z = 2$; logo $z^2 = 4$. A porcentagem de pessoas com sangue tipo 0 na ilha, segundo o antropólogo, deve ser, em porcentagem:

$$p = 50$$

Logo:

$$100 - p = 50$$

O valor d é a margem de erro. Em porcentagem:

$$d = 5$$

Logo:

$$d^2 = 25$$

Então o tamanho da amostra deve ser

$$n = \frac{4 \times 50 \times 50}{25} = 400$$

A equação dada aqui está simplificada e só vale se a população da ilha for tão grande que, para finalidade de estatísticas, possa ser considerada *infinita*. A equação também só pode ser aplicada se p estiver entre 30% e 70%.

Mas importante é saber que não basta ter em mãos uma fórmula, ou um programa de computador para estimar o tamanho de uma amostra. É preciso algum conhecimento prévio (estimativas preliminares de um ou mais parâmetros, obtidas de amostras piloto ou da literatura) e uma boa dose de bom senso.

1.7 – A QUESTÃO DA REPRESENTATIVIDADE

A amostra só traz informação sobre a *população da qual foi retirada*. Não tem sentido, por exemplo, estudar os hábitos de higiene de índios bolivianos e considerar que as informações “servem” para descrever os hábitos de higiene de moradores da periferia da cidade de São Paulo. Ainda, a amostra deve ter o *tamanho usual da área* em que a pesquisa se enquadra. Amostras demasiado pequenas não dão informação útil. Desconfie, também, de amostras muito grandes. Será que o pesquisador observou cada unidade amostrada com o devido cuidado?

As amostras podem ser *representativas* ou *não-representativas*. E não se pode julgar a qualidade da amostra pelos resultados obtidos. Se você jogar uma moeda 10 vezes, *podem* ocorrer 10 caras. Provável? Não. Possível? Sim.

Conclusões e decisões tomadas com base em amostras só têm sentido na medida em que as amostras representam a população. Para bem interpretar os dados e tirar conclusões adequadas, não basta olhar os números: é preciso entender como a amostra foi tomada e se não incidiram, no processo de amostragem, alguns fatores que poderiam trazer tendência aos dados.

Como você sabe se uma amostra é tendenciosa? Não há *fórmulas* de matemática ou estatística para dizer se a amostra é tendenciosa ou representativa da população. Você terá de ter bom senso e conhecimento na área. São, portanto, necessários muitos cuidados, porque os erros de amostragem podem ser sérios.

Tendência é a diferença entre a estimativa que se obteve na amostra e o parâmetro que se quer estimar.

Exemplo 1.7: Uma amostra tendenciosa.

Em 1988, Shere Hite⁴ levantou, por meio de questionários inseridos em revistas femininas americanas, dados sobre a sexualidade feminina. Estima-se que cerca de 100.000 mulheres foram colocadas em contato com o questionário, mas só 4.500 responderam. Mesmo assim, a amostra é grande. Você acha que essa amostra pode dar boa idéia do comportamento sexual das mulheres americanas daquela época?

⁴O exemplo é de SILVER, M. **Estatística para Administração**. São Paulo, Atlas, 2000.

Solução

O comportamento dos voluntários é diferente do comportamento dos não-voluntários. Então — embora seja difícil ou até impossível estudar o comportamento de pessoas que não respondem a um questionário — não se pode concluir que a amostra de respondentes representa toda a população (incluindo aqueles que não respondem). Conclusões baseadas em amostras de pessoas que, voluntariamente, destacam o encarte de uma revista, respondem ao questionário e o remetem pelo correio são tendenciosas. Não se pode fugir à conclusão de que o questionário foi respondido apenas por leitoras da revista e, entre elas, mulheres dispostas a falar sobre sua vida pessoal.

Finalmente, algumas pessoas dizem não acreditar em resultados obtidos de pesquisas porque elas próprias nunca foram chamadas para opinar. Se você é dos que não acreditam em pesquisas porque nunca foi entrevistado, então, por coerência, não tome um analgésico, não dirija um carro, não beba cerveja. Afinal, a qualidade desses produtos também é avaliada por amostragem, das quais possivelmente você também não participou. É verdade que ocorrem erros, é verdade que existem fraudes e é verdade que o improvável também acontece, mas daí a achar que não existem acertos vai uma enorme distância. O Brasil tem excelentes institutos de pesquisa.

1.8 – EXERCÍCIOS RESOLVIDOS

1.8.1 – Os prontuários dos pacientes de um hospital estão organizados em um arquivo, por ordem alfabética. Qual é a maneira mais rápida de amostrar 1/3 do total de prontuários?

Seleciona-se, para a amostra, um de cada três prontuários ordenados (por exemplo, o terceiro de cada três).

1.8.2 – Um pesquisador tem 10 gaiolas, cada uma com seis ratos. Como o pesquisador pode selecionar 10 ratos para uma amostra?

O pesquisador pode usar a técnica de amostragem aleatória estratificada, isto é, sortear um rato de cada gaiola para compor a amostra.

1.8.3 – Para levantar dados sobre o número de filhos por mulher, em uma comunidade, um pesquisador organizou um questionário que enviou, pelo correio, a todas as residências. A resposta ao questionário era facultativa, pois o pesquisador não tinha condições de exigir a resposta. Nesse questionário perguntava-se o número de filhos por mulher moradora na residência. Você acha que os dados assim obtidos seriam tendenciosos?

Os dados devem ser tendenciosos porque é razoável esperar que: a) mulheres com muitos filhos responderiam, pensando na possibilidade de algum tipo de ajuda, como instalação de uma creche no bairro; b) mulheres que recentemente tiveram o primeiro filho também responderiam; c) muitas das mulheres que não têm filhos não responderiam; d) mulheres com filhos adultos e emancipados não responderiam.

1.8.4 – Um pesquisador pretende levantar dados sobre o número de moradores por domicílio, usando a técnica de amostragem sistemática. Para isso, o pesquisador visitará cada domicílio selecionado. Se nenhuma pessoa estiver presente na ocasião da visita, o pesquisador excluirá o domicílio da amostra. Esta última determinação torna a amostra tendenciosa. Por quê?

Nos domicílios onde moram muitas pessoas, será mais fácil o pesquisador encontrar pelo menos uma pessoa por ocasião de sua visita. Então, é razoável admitir que os domicílios com poucos moradores tenham maior probabilidade de serem excluídos da amostra.

1.8.5 – Muitas pessoas acreditam que as famílias se tornaram menores. Suponha que, para estudar essa questão, um pesquisador selecionou uma amostra de 2.000 casais e perguntou quantos filhos eles tinham, quantos filhos tinham seus pais e quantos filhos tinham seus avós. O procedimento produz dados tendenciosos. Por quê?

Os casais de gerações anteriores que não tiveram filhos não têm possibilidade de ser selecionados para a amostra. Por outro lado, os casais de gerações anteriores que tiveram muitos filhos terão grande probabilidade de ser amostrados.

1.8.6 – Para estudar atitudes religiosas, um sociólogo sorteia 10 membros de uma grande igreja para compor uma amostra casual simples. Nota, então, que a amostra ficou composta por nove mulheres e um homem. O sociólogo se espanta: “A amostra não é aleatória! Quase só tem mulher.” O que você diria?

Se a amostra é ou não aleatória depende de como foi selecionada e não de sua composição. As probabilidades envolvidas no processo de constituir uma amostra aleatória podem determinar amostras atípicas.

1.8.7 – Para avaliar a expectativa de pais de adolescentes em relação às possibilidades de estudo de seus filhos, foram distribuídos 5.000 questionários pelos estados do sul do Brasil. Retornaram 1.032. Cerca de 60% dos respondentes diziam que a maior preocupação deles era com o preço que

se paga para um jovem cursar a universidade. Você considera esse resultado uma boa estimativa para o número de pais preocupados com essa questão?

Não é uma boa estimativa porque os respondentes foram relativamente poucos (cerca de 20%). Ainda, tendem a responder pais que querem seus filhos na universidade e estão preocupados com os custos.

1.8.8 – Um dentista quer levantar o tipo de documentação que seus colegas arquivam, quando fazem um tratamento ortodôntico. A documentação depende do caso, mas também envolve questões legais e de bom senso do ortodontista. Para essa pesquisa, o dentista elabora um questionário que envia, por correio, a todos os profissionais inscritos no conselho de odontologia. O dentista provavelmente não receberá respostas de todos. Você saberia dizer algumas das razões de isso acontecer?

Razões possíveis: 1) Nem todos os endereços que constam dos arquivos de um conselho estão atualizados. 2) Nem todas as pessoas que recebem questionários por correio o respondem, seja porque não têm tempo, têm preguiça ou inércia, imaginam razões espúrias para terem sido contatadas etc. 3) Não dão respostas por correio pessoas que têm alguma dificuldade de chegar ao correio, seja porque moram longe, porque não gostam de andar ou não têm condução própria, porque não têm hábito de enviar correspondência, porque a secretaria não leva correspondência ao correio etc. 4) Dos que não têm nenhum dos motivos citados, ainda deixaria de responder o profissional que não tem boa documentação de casos ou não a tem em ordem. 5) Provavelmente também não respondem profissionais que estejam enfrentando problema de ordem financeira, legal, de admissão em cursos etc.

1.8.9 – Para estudar o uso de serviços de saúde por mulheres em idade reprodutiva moradoras de uma grande capital, um pesquisador buscou na Fundação Instituto Brasileiro de Geografia e Estatística (IBGE) as subdivisões da cidade utilizadas em censos, conhecidas como setores censitários. Como você procederia para tomar uma amostra de mulheres moradoras nesses setores e em idade reprodutiva?

Cada setor pode ser considerado como um conglomerado. Podem ser sorteados quatro setores. Depois, em cada setor, escolhe-se um ponto ao acaso e, a partir daí, tira-se uma amostra sistemática. A unidade amostral é um domicílio com mulheres em idade reprodutiva, de 10 a 49 anos. Devem ser excluídas do estudo mulheres que não queiram participar.

1.9 – EXERCÍCIOS PROPOSTOS

1.9.1 – Dada uma população de quatro pessoas, Antônio, Luís, Pedro e Carlos, escreva as amostras casuais simples de tamanho 2 que podem ser obtidas.

1.9.2 – Descreva três formas diferentes de obter uma amostra sistemática de quatro elementos de uma população de oito elementos, A, B, C, D, E, F, G e H.

1.9.3 – Dada uma população de 40 alunos, descreva uma forma de obter uma amostra casual simples de seis alunos.

1.9.4 – Organize uma lista com 10 nomes de pessoas em ordem alfabética. Depois descreva uma forma de obter uma amostra sistemática de cinco nomes.

1.9.5 – Em uma pesquisa de mercado para serviços odontológicos tomou-se a lista telefônica, onde os nomes dos assinantes estão organizados em ordem alfabética do último sobrenome, e se amostrou o décimo de cada 10 assinantes. Critique esse procedimento.

1.9.6 – Um fiscal precisa verificar se as farmácias da cidade estão cumprindo um novo regulamento. A cidade tem 40 farmácias, mas como a fiscalização demanda muito tempo, o fiscal resolveu optar por visitar uma amostra de 10 farmácias. O cumprimento do regulamento que é, evidentemente, desconhecido do fiscal está apresentado na tabela a seguir. Com base na tabela a seguir:

- a) escolha uma amostra para o fiscal;
- b) estime, com base na amostra, a proporção de farmácias que estão cumprindo o regulamento;
- c) com base nos dados da população, estime o parâmetro;
- d) você obteve uma boa estimativa?

Dados sobre cumprimento do regulamento:

<i>Cumprimento do regulamento</i>			
1. Sim	11. Não	21. Sim	31. Sim
2. Sim	12. Sim	22. Sim	32. Sim
3. Não	13. Não	23. Não	33. Não
4. Sim	14. Não	24. Sim	34. Sim
5. Sim	15. Sim	25. Não	35. Sim
6. Não	16. Não	26. Não	36. Não
7. Sim	17. Sim	27. Não	37. Não
8. Não	18. Não	28. Sim	38. Não
9. Não	19. Não	29. Não	39. Sim
10. Sim	20. Sim	30. Não	40. Sim

1.9.7 – A maneira de fazer a pergunta pode influenciar a resposta da pessoa que responde. Basicamente, existem dois tipos de questões: a “questão fechada” e a “questão aberta”. Na “questão fechada” o pesquisador fornece uma série de respostas possíveis e a pessoa que responde deve apenas assinalar a alternativa, ou as alternativas, que lhe convém. A “questão aberta” deve ser respondida livremente. Imagine que um dentista quer levantar dados sobre hábitos de higiene oral das pessoas de uma comunidade. Escreva então uma “questão fechada” e uma “questão aberta”.

1.9.8 – Uma classe tem quatro alunos. Eles foram submetidos a uma prova e suas notas foram: João, 10; José, 6; Paulo, 4; Pedro, 0. Calcule a média da classe (parâmetro). Depois, construa todas as amostras de tamanho 2 e calcule a média de cada uma (estatísticas). Verifique que a média das estatísticas é igual ao parâmetro.

1.9.9 – Um fabricante de produtos alimentícios pede a você para escolher uma cidade do seu Estado para fazer o teste de um novo produto. Como você escolheria a cidade: por sorteio ou usaria o seu julgamento do que considera uma “cidade típica” do Estado?

- 1.9.10 – Pretende-se obter uma amostra dos alunos de uma universidade para estimar o percentual que tem trabalho remunerado. a) Qual é a população em estudo? b) Qual é o parâmetro que se quer estimar? c) Você acha que se obteria uma boa amostra dos alunos no restaurante universitário? d) No ponto de ônibus mais próximo?
- 1.9.11 – Um editor de livros técnicos quer saber se os leitores preferem capas de cores claras com desenhos, ou capas simples de cores mais escuras. Se o editor pedir a você para estudar a questão, como você definiria a população do estudo?
- 1.9.12 – Um dentista quer estudar a porcentagem de policiais militares com distúrbios na articulação temporo-mandibular. Calcule ao tamanho da amostra, considerando que o dentista quer um nível de confiança de 95% ($z = 2$), uma margem de erro de 8% ($d = 8\%$) e que, na população, a porcentagem de pessoas com esse tipo de distúrbio é 35%.

(página deixada intencionalmente em branco)

Apresentação de Dados em Tabelas

2

(página deixada intencionalmente em branco)

Você já aprendeu que os estatísticos coletam informações. Essas informações podem ser sobre peso de pessoas, eficiência de medicamentos, incidência de doenças, causas de morte, quantidade de hemoglobina no sangue, estresse, ansiedade etc. Neste Capítulo vamos aprender como essas informações são organizadas para facilitar a leitura. Mas antes vamos aprender o que são dados e o que são variáveis.

2.1 – DADOS E VARIÁVEIS

Variável é uma condição ou característica das unidades da população; a variável pode assumir valores diferentes em diferentes unidades. Por exemplo, a idade das pessoas residentes no Brasil é uma variável. *Dados* são os valores da variável em estudo, obtidos por meio de uma amostra.

Exemplo 2.1: Dados e variáveis.

O dono de uma academia de ginástica quer saber a opinião de seus clientes sobre a qualidade dos serviços que presta. O que é variável e o que são dados nesse problema?

Solução

A variável de interesse é a opinião dos clientes. Os *dados* serão obtidos somente quando o dono da academia começar a pedir aos clientes que dêem uma nota a cada serviço. Então, se for pedido que o cliente dê uma nota de zero a 5 a cada serviço que utiliza — os dados coletados poderão ser, por exemplo, 4, 3, 2, 4, 1 etc., por serviço.

As variáveis são classificadas em dois tipos:

- quantitativas ou numéricas;
- qualitativas ou categorizadas.

Uma variável é *qualitativa* ou *categorizada* quando os dados são distribuídos em categorias mutuamente exclusivas. São exemplos de variáveis qualitativas: time de futebol do qual a pessoa é torcedora (se a pessoa torce por um time, não torce pelo outro); sexo (é masculino ou é feminino); cidade de nascimento (se a pessoa nasceu em Niterói, automaticamente fica excluída a possibilidade de ter nascido em outra cidade).

Uma variável é *quantitativa* ou *numérica* quando é expressa por números. São exemplos de variáveis quantitativas: idade, estatura, número de crianças numa escola, número de lápis numa caixa.

As variáveis qualitativas ou categorizadas são classificadas em dois tipos:

- Nominal;
- Ordinal.

A variável é *nominal* quando os dados são distribuídos em categorias mutuamente exclusivas, mas são indicadas em qualquer ordem. São variáveis nominais: cor de cabelos (loiro, castanho, preto, ruivo), tipo de sangue (O, A, B, AB), gênero (masculino, feminino), religião (espírita, católico, evangélico, outras) etc.

A variável é *ordinal* quando os dados são distribuídos em categorias mutuamente exclusivas que têm ordenação natural. São variáveis ordinais: escolaridade (primeiro grau, segundo grau, terceiro grau), classe social (A, B, C, D, E), gravidade de uma doença (leve, moderada, severa) etc.

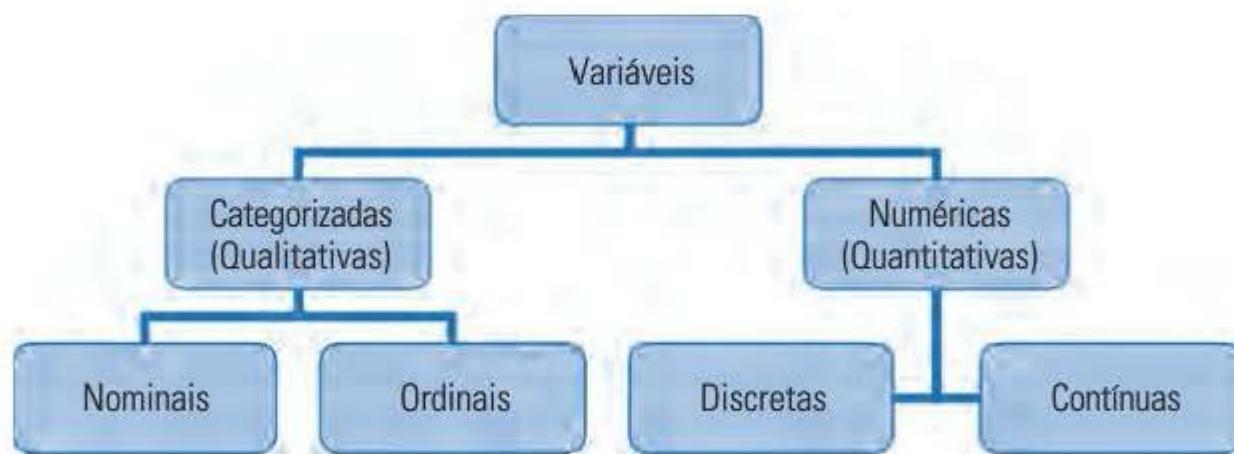
As variáveis quantitativas ou numéricas são classificadas em dois tipos:

- Discreta;
- Contínua.

A variável *discreta* só pode assumir alguns valores em um dado intervalo. São variáveis discretas: número de filhos (nenhum, 1, 2, 3, 4 etc.), quantidade de moedas num bolso (zero, 1, 2, 3 etc.), número de pessoas numa sala.

A variável *contínua* assume qualquer valor num dado intervalo. São variáveis contínuas: peso, tempo de espera, quantidade de chuva etc.

Os dados são do mesmo tipo que o das variáveis. Por exemplo, uma variável discreta produz dados discretos. Veja o organograma:



2.2 – APURAÇÃO DE DADOS

Dados são registrados em fichas, em cadernos, em computador. Para obter apenas os dados de interesse para sua pesquisa, você deve fazer uma *apuração*. Se a variável for qualitativa, a apuração se resume a simples contagem. Veja como isto pode ser feito.

Para estudar a razão de sexos¹ dos recém-nascidos em uma maternidade e seus pesos ao nascer, um pesquisador obteve uma amostra sistemática de 1.000 prontuários de recém-nascidos e escreveu numa folha de papel:

Masculino

Feminino

Depois examinou todos os prontuários e fez, então, um traço na linha que indicava cada sexo, toda vez que o prontuário registrava que o recém-nascido era desse sexo. Cada quadrado, cortado pela diagonal, representa cinco recém-nascidos. O total é dado pelo número de traços em cada linha.

Masculino	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> ... <input checked="" type="checkbox"/> <input type="checkbox"/> = 509
Feminino	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/> ... <input type="checkbox"/> = 491

Quando a variável é quantitativa, é preciso anotar, na apuração, cada valor observado. Para apurar dados de peso ao nascer², o pesquisador deve anotar o número do prontuário e o peso ao nascer, numa folha de papel. O número do prontuário, escrito ao lado do peso ao nascer, facilita a posterior verificação da apuração.

<i>Nº do prontuário</i>	<i>Peso ao nascer</i>
10 525	3,250
10 526	2,010
...	...
10 624	2,208

¹Razão de sexos: número de homens por 100 mulheres.

²A apuração de peso ao nascer pode ser feita por sexo se o interesse é comparar pesos ao nascer de meninos e meninas.

2.3 – COMPONENTES DAS TABELAS

Os dados devem ser apresentados em tabelas construídas de acordo com as normas técnicas ditadas pela Fundação Instituto Brasileiro de Geografia e Estatística (IBGE) (1993)³. As tabelas devem ser colocadas perto do ponto do texto em que são mencionadas pela primeira vez. Devem ser inseridas na ordem em que aparecem no texto.

Veja a Tabela 2.1, que obedece às normas técnicas. De acordo com essas normas, uma tabela deve ter título, corpo, cabeçalho e coluna indicadora. O *título* explica o que a tabela contém. O *corpo* é formado pelos dados, em linhas e colunas. O *cabeçalho* especifica o conteúdo das colunas. A *coluna indicadora* especifica o conteúdo das linhas.

Exemplo 2.2: Componentes de uma tabela.

TABELA 2.1
População residente no Brasil, segundo o sexo, de acordo com o censo demográfico de 2000.

<i>Sexo</i>	<i>População residente</i>
Masculino	83.576.015
Feminino	86.223.155
Total	169.799.170

Fonte: IBGE (2003)³

Na Tabela 2.1, observe o título:

População residente no Brasil, segundo o sexo, de acordo com o censo demográfico de 2000.

O cabeçalho é constituído por:

<i>Sexo</i>	<i>População residente</i>
-------------	----------------------------

³As normas do IBGE são excelentes. Veja em: <http://www1.ibge.gov.br/home/estatistica/populacao/censo2000/tabelabrasil111.shtm>. Disponível em 20 de abril de 2008. Veja também: VIEIRA, S. *Elementos de Estatística*. São Paulo, Atlas, 5 ed. 2003.

A coluna indicadora é constituída pelas especificações:

<i>Sexo</i>
Masculino
Feminino
Total

O corpo da tabela é formado pelos números:

<i>População residente</i>
83.576.015
86.223.155
169.799.170

Toda tabela deve ser delimitada por *traços horizontais*, mas *não* deve ser delimitada por *traços verticais*. Os traços verticais podem ser feitos somente para *separar* as colunas. O cabeçalho deve ser separado do corpo da tabela por um traço horizontal.

As tabelas podem conter *fonte e notas*. *Fonte* é a entidade, ou pesquisador, ou pesquisadores que publicaram ou forneceram os dados. Veja a Tabela 2.1: a fonte é a Fundação Instituto Brasileiro de Geografia e Estatística (IBGE), que publicou os dados.

As *notas* esclarecem aspectos relevantes do levantamento dos dados ou da apuração. Veja a nota apresentada na Tabela 2.2, a qual informa que, na apuração, foram suprimidos os casos com idade, ou local de residência ignorados.

Exemplo 2.3: Uma tabela com nota de rodapé.

TABELA 2.2
Número de internações hospitalares, de mulheres, pelo Sistema Único de Saúde (SUS). Brasil, 2005.

Grupo de doenças	Número
Gravidez, parto e puerpério	2.640.438
Doenças do aparelho respiratório	736.012
Doenças do aparelho circulatório	612.415
Doenças do aparelho geniturinário	507.295
Doenças infecciosas e parasitárias	480.165
Doenças do aparelho digestivo	452.894
Transtornos mentais e comportamentais	105.354
Neoplasias	355.570
Causas externas	233.787
Demais causas	801.123
Total	6.925.053

Fonte: Ministério da Saúde/SE/Datasus — Sistema de Informações Hospitalares do SUS — SIH/SUS.
Nota: Suprimidos os casos com idade ou local de residência ignorados.

2.4 – APRESENTAÇÃO DE DADOS QUALITATIVOS

Quando observamos *dados qualitativos*, classificamos cada unidade da amostra em uma dada categoria. Nossa conhecimento sobre os dados aumenta se contarmos quantas unidades caem em cada categoria. A idéia seguinte é resumir as informações na forma de uma tabela que mostre as contagens (freqüências) em cada categoria. Temos então uma *tabela de distribuição de freqüências*.

Exemplo 2.4: Uma tabela de distribuição de freqüências para dados ordinais.

Foram entrevistados 2.500 brasileiros, com 16 anos ou mais, para saber a opinião deles sobre determinado técnico de futebol. Veja o que eles responderam: 1.300 achavam que o técnico era bom, 450 achavam regular e 125 achavam ruim; 625 não tinham opinião ou não quiseram opinar. Como se organizam estes dados em uma tabela de distribuição de freqüências?

Solução

Na Tabela 2.3 estão as respostas dadas pelos entrevistados (primeira coluna) e as freqüências dessas respostas (segunda coluna). A soma das freqüências é 2.500 (número de entrevistados).

TABELA 2.3
Opinião dos brasileiros sobre determinado técnico de futebol.

<i>Respostas</i>	<i>Freqüência</i>
Bom	1.300
Regular	450
Ruim	125
Não sabe	625
Total	2.500

Nas tabelas de distribuição de freqüências, é usual fornecer a *proporção* (freqüência relativa) de unidades que caem em cada categoria. Para obter a freqüência relativa de uma dada categoria, calcule:

$$\text{Freqüência relativa} = \frac{\text{Freqüência}}{\text{Tamanho da amostra}}$$

Exemplo 2.5: Uma tabela de distribuição de freqüências e freqüências relativas.

Calcule as freqüências relativas dos dados apresentados na Tabela 2.3.

Solução

Na Tabela 2.4 estão as respostas dadas pelos entrevistados (primeira coluna), as freqüências dessas respostas (segunda coluna) e as freqüências relativas (terceira coluna). Note que as freqüências relativas somam 1,00.

TABELA 2.4
Opinião dos brasileiros sobre determinado técnico de futebol.

<i>Respostas</i>	<i>Freqüência</i>	<i>Freqüência relativa</i>
Bom	1.300	$\frac{1300}{2500} = 0,52$
Regular	450	$\frac{450}{2500} = 0,18$
Ruim	125	$\frac{125}{2500} = 0,05$
Não sabe	625	$\frac{625}{2500} = 0,25$
Total	2.500	1,00

As freqüências relativas são, em geral, dadas em porcentagens. Para transformar uma freqüência relativa em porcentagem, basta multiplicar por 100. No exemplo dado na Tabela 2.4, a *freqüência relativa* de respostas “bom” é 0,52. Multiplicando esse resultado por 100, temos a *porcentagem*, que é 52%. Este resultado — 52% de “bom” — é bem entendido pelas pessoas.

As freqüências relativas dadas em porcentagens fornecem a informação mais relevante. Mas sempre convém exibir o total (tamanho da amostra), que é indicador da credibilidade da informação dada⁴.

2.5 – TABELAS DE CONTINGÊNCIA

Muitas vezes os elementos da amostra ou da população são classificados de acordo com duas variáveis qualitativas. Os dados devem então ser apresentados em *tabelas de contingência*, isto é, em tabelas de dupla entrada, cada entrada relativa a uma das variáveis.

Exemplo 2.6: Uma tabela de contingência.

Foram feitos diagnósticos de depressão em 500 estudantes com idades entre 10 e 17 anos, metade de cada sexo. Foram identificados 98 casos de depressão, sendo 62 no sexo feminino. Apresente os dados em uma tabela.

Solução

Note que os dados estão classificados segundo duas variáveis: sexo e presença de depressão.

TABELA 2.5
Sexo e presença de depressão.

Sexo	Depressão		Total
	Sim	Não	
Masculino	36	214	250
Feminino	62	188	250
Total	98	402	500

As tabelas de contingência podem apresentar freqüências relativas em porcentagens, além das freqüências. O tamanho da amostra é sempre im-

⁴Não tem sentido fornecer resultados em porcentagens quando a amostra é muito pequena. Por exemplo, não tem sentido fornecer porcentagens se a amostra fosse constituída de cinco ou seis pessoas.

portante, porque não se pode confiar em resultados obtidos com base em amostras muito pequenas — e calcular porcentagens sobre alguns poucos casos.

Exemplo 2.7: Uma tabela de contingência com freqüências relativas.

Para verificar se o risco de óbito neonatal é maior quando a gestante é diabética, foram obtidos os dados apresentados na Tabela 2.6. Discuta.

TABELA 2.6
Óbito neonatal e diabetes mellitus.

Gestante	<i>Óbito neonatal</i>		Total	<i>Percentual de óbitos</i>
	Sim	Não		
Diabética	3	21	24	12,5%
Não-diabética	21	830	851	2,5%
Total	24	851	875	

O risco de óbito neonatal, dado pelo percentual de óbitos, é maior quando a gestante é diabética.

2.6 – APRESENTAÇÃO DE DADOS NUMÉRICOS

Os dados numéricos são apresentados na ordem em que são coletados. Geralmente são obtidos dados relativos a diversas variáveis em cada paciente. Os pacientes são identificados nas pesquisas por números.

Exemplo 2.8: Uma tabela com dados numéricos.

Para estudar o desempenho cardíaco de pacientes submetidos à diálise renal, foram obtidos valores de diversas variáveis de interesse da Cardiologia. Na Tabela 2.7 são apresentadas apenas algumas informações, para mostrar como se apresentam dados numéricos.

TABELA 2.7

Idade em anos completos, tempo de diálise em meses, altura em metros, peso em quilogramas, pressão sistólica e diastólica em milímetros de mercúrio de mulheres submetidas à diálise renal.

Número da paciente	Idade	Tempo de diálise	Altura	Peso	Pressão sistólica	Pressão diastólica
1	45	14	1,60	62,0	140	85
2	62	54	1,65	52,5	100	70
3	38	52	1,55	67,8	140	100
4	26	34	1,59	48,2	165	105
5	35	18	1,58	46,0	170	105
6	44	71	1,48	40,4	150	100
7	53	39	1,69	67,7	155	95
8	44	79	1,59	55,5	160	105
9	58	23	1,62	63,0	175	110
10	55	64	1,51	50,3	155	105
11	24	16	1,79	77,0	160	95
12	70	46	1,51	44,0	150	95
13	56	48	1,58	64,0	175	110

Dados numéricos também podem ser apresentados em *tabelas de distribuição de freqüências*. Se os dados são *discretos*, para organizar a tabela de distribuição de freqüências:

- escreva os dados em ordem crescente.
- conte quantas vezes cada valor se repete.
- organize a tabela como já foi feito para dados qualitativos, colocando no lugar das categorias, os valores numéricos em ordem natural. Veja o Exemplo 2.9.

Exemplo 2.9: Uma tabela de distribuição de freqüências para dados discretos.

As faltas ao trabalho de 30 empregados de uma clínica em determinado semestre estão na Tabela 2.8. A partir dela, faça uma tabela de distribuição de freqüências.

TABELA 2.8

Número de faltas dadas por 30 empregados de uma clínica no semestre.

1	3	1	1	0	1	0	1	1	0
2	2	0	0	0	1	2	1	2	0
0	1	6	4	3	3	1	2	4	0

Solução

TABELA 2.9

Distribuição do número de faltas de 30 empregados de uma clínica no semestre.

Número de faltas	Freqüência	Percentual
0	9	30,0
1	10	33,3
2	5	16,7
3	3	10,0
4	2	6,7
5	0	0,0
6	1	3,3
Total	30	100,0

Tabelas com grande número de dados não oferecem ao leitor visão rápida e global do fenômeno. Observe os dados apresentados na Tabela 2.10: é difícil dizer como os valores se distribuem. Por esta razão, dados contínuos — desde que em grande número — são apresentados em *tabelas de distribuição de freqüências*.

Exemplo 2.10: Uma tabela com dados contínuos.
TABELA 2.10
Peso ao nascer de nascidos vivos, em quilogramas.

2,522	3,200	1,900	4,100	4,600	3,400
2,720	3,720	3,600	2,400	1,720	3,400
3,125	2,800	3,200	2,700	2,750	1,570
2,250	2,900	3,300	2,450	4,200	3,800
3,220	2,950	2,900	3,400	2,100	2,700
3,000	2,480	2,500	2,400	4,450	2,900
3,725	3,800	3,600	3,120	2,900	3,700
2,890	2,500	2,500	3,400	2,920	2,120
3,110	3,550	2,300	3,200	2,720	3,150
3,520	3,000	2,950	2,700	2,900	2,400
3,100	4,100	3,000	3,150	2,000	3,450
3,200	3,200	3,750	2,800	2,720	3,120
2,780	3,450	3,150	2,700	2,480	2,120
3,155	3,100	3,200	3,300	3,900	2,450
2,150	3,150	2,500	3,200	2,500	2,700
3,300	2,800	2,900	3,200	2,480	-
3,250	2,900	3,200	2,800	2,450	-

Para construir uma tabela de distribuição de freqüências com dados contínuos:

- Ache o valor máximo e o valor mínimo do conjunto de dados.
- Calcule a *amplitude*, que é a diferença entre o valor máximo e o valor mínimo.
- Divida a amplitude dos dados pelo número de faixas que pretende organizar (no caso do Exemplo 2.10, as faixas são de peso). Essas faixas recebem, tecnicamente, o nome de *classes*.
- O resultado da divisão é o *intervalo de classe*. É sempre melhor arredondar esse número para um valor mais alto, o que facilita o trabalho.
- Organize as classes, de maneira que a primeira contenha o menor valor observado.

Observe os dados apresentados na Tabela 2.10. O menor valor é 1,570 kg e o maior valor 4,600 kg. A amplitude dos dados é:

$$4,600 - 1,570 = 3,030$$

Vamos definir *sete classes*. Então calcule:

$$3,030 \div 7 = 0,433$$

Arredonde esse valor para 0,500 e construa a primeira classe, que será de 1,5 kg a 2,0 kg (esta classe contém o menor valor); depois, construa a segunda classe, que será de 2,0 kg a 2,5 kg, e assim por diante, como mostra o esquema dado a seguir:

$$\begin{aligned} 1,5 &\leftarrow 2,0 \\ 2,0 &\leftarrow 2,5 \\ 2,5 &\leftarrow 3,0 \\ 3,0 &\leftarrow 3,5 \\ 3,5 &\leftarrow 4,0 \\ 4,0 &\leftarrow 4,5 \\ 4,5 &\leftarrow 5,0 \end{aligned}$$

Na classe de 1,5 kg até menos de 2,0 kg são colocados desde nascidos com 1,5 kg até os que nasceram com 1,999 kg; na classe de 2,0 kg até menos de 2,5 kg são colocados desde nascidos com 2,0 kg até os que nasceram com 2,499 kg e assim por diante. Logo, cada classe cobre um intervalo de 0,5 kg. É mais fácil trabalhar com intervalos de classe iguais.

Denominam-se *extremos de classe* os limites dos intervalos de classe. Deve ficar claro, na tabela de distribuição de freqüências, se os valores iguais aos extremos estão ou não incluídos na classe. Veja a notação usada no exemplo. A primeira classe é:

$$1,5 \leftarrow 2,0.$$

Isto significa que o intervalo é *fechado à esquerda*, isto é, pertencem à classe os valores iguais ao extremo inferior da classe (por exemplo, 1,5 na primeira classe). Também significa que o intervalo é *aberto à direita*, isto é, não pertencem à classe os valores iguais ao extremo superior (por exemplo, o valor 2,0 não pertence à primeira classe).

Exemplo 2.11: Uma tabela de distribuição de freqüências para dados contínuos.

Para dar idéia geral sobre peso ao nascer de nascidos vivos, o pesquisador quer apresentar não os pesos observados — mas o número de nascidos vivos por faixas de peso. A Tabela 2.11 apresenta a distribuição de freqüências.

TABELA 2.11
Distribuição de freqüências para peso ao nascer de nascidos vivos, em quilogramas.

<i>Classe</i>	<i>Freqüência</i>
1,5 – 2,0	3
2,0 – 2,5	16
2,5 – 3,0	31
3,0 – 3,5	34
3,5 – 4,0	11
4,0 – 4,5	4
4,5 – 5,0	1

É importante lembrar aqui que existem outras maneiras de indicar se os extremos de classe estão, ou não, incluídos em determinada classe. Aliás, a Fundação Instituto Brasileiro de Geografia e Estatística (IBGE) usa notação diferente. Para dados de idade, por exemplo, escreve: “De 0 até 4 anos”, “De 5 até 9 anos”, “De 10 até 14 anos” e assim por diante. A classe “De 0 até 4 anos” inclui desde indivíduos que acabaram de nascer até indivíduos que estão na véspera de completar 5 anos.

O número de classes deve ser escolhido pelo pesquisador, em função do que ele quer mostrar. Em geral, convém estabelecer de 5 a 20 classes. Se o número de classes for demasiado pequeno (por exemplo, 3), perde-se muita informação. Se o número de classes for grande (por exemplo, 30), têm-se pormenores desnecessários. Não existe um número “ideal” de classes para um conjunto de dados, embora existam até fórmulas para estabelecer quantas classes devem ser construídas.

Os resultados obtidos por meio de fórmulas podem servir como referência — mas não devem ser entendidos como obrigatórios. Para usar uma dessas fórmulas, faça n indicar o *número de dados*. O *número de classes* será um inteiro próximo de k , obtido pela fórmula:

$$k = \sqrt{n}$$

ou então, por esta segunda fórmula:

$$k = 1 + 3,222 \times \log n$$

Exemplo 2.12: Cálculo do número de classes.

Para entender como se obtém o número de classes por meio de fórmula, reveja a Tabela 2.11. Como $n = 100$, aplicando a primeira fórmula dada, tem-se que:

$$k = \sqrt{n} = \sqrt{100} = 10$$

Aplicando a segunda fórmula, obtém-se:

$$k = 1 + 3,222 \times \log n = 1 + 3,222 \times \log 100 = 7,444$$

Para obter o número de classes apresentadas no Exemplo 2.11, foi aplicada a segunda fórmula e, por isto, foram construídas *sete classes*.

Numa distribuição de freqüências, o extremo inferior da primeira classe, o extremo superior da última classe ou ambos *podem* não estar definidos. Ainda, os intervalos de classe podem ser diferentes.

Exemplo 2.13: Uma tabela de distribuição de freqüências para dados contínuos com classes de tamanhos diferentes e extremo superior não definido.

Para dar uma idéia geral sobre pressão sanguínea sistólica de mulheres com 30 anos de idade, o pesquisador apresentou não os valores observados — mas o número de mulheres por faixas de pressão. Veja a Tabela 2.12, que também é um exemplo em que o extremo superior da última classe não está definido.

TABELA 2.12

Distribuição de freqüências para a pressão sanguínea sistólica, em milímetros de mercúrio, de mulheres com 30 anos de idade.

<i>Classe</i>	<i>Freqüência</i>
90 – 100	6
100 – 105	11
105 – 110	12
110 – 115	17
115 – 120	18
120 – 125	11
125 – 130	9
130 – 135	6
135 – 140	4
140 – 150	4
150 – 160	1
160 e mais	1

As tabelas de distribuição de freqüências mostram a distribuição da variável, *mas perdem em exatidão*. Por exemplo, a Tabela 2.12 mostra que seis mulheres apresentaram pressão sanguínea sistólica entre 90 e 100, mas não dá informação exata sobre a pressão de cada uma delas.

2.7 – EXERCÍCIOS RESOLVIDOS

2.7.1 – Converta as seguintes proporções em porcentagens: 0,09; 0,955; 0,33; 0,017.

Basta multiplicar por 100, para obter: 9%; 95,5%; 33%; 1,7%.

2.7.2 – Converta as seguintes porcentagens em proporções: 35,5%; 53,1%; 50%; 46,57%.

Basta dividir por 100, para obter: 0,355; 0,531; 0,50; 0,4657.

2.7.3 – Para estudar a distribuição dos erros cometidos por alunos nas tomadas radiográficas, foi feito um levantamento de dados na seção de Radiologia de uma faculdade de odontologia. Calcule as freqüências relativas e os totais.

TABELA 2.13
Erros em tomadas radiográficas.

<i>Erros</i>	<i>Freqüência</i>
Posição do paciente	598
Fatores de exposição	288
Processamento	192
Produção de artefatos	101
Posição do chassi	83
Outros fatores	53

TABELA 2.14
Erros em tomadas radiográficas.

<i>Erros</i>	<i>Freqüência</i>	<i>Freqüência relativa</i>
Posição do paciente	598	45,5%
Fatores de exposição	288	21,9%
Processamento	192	14,6%
Produção de artefatos	101	7,7%
Posição do chassi	83	6,3%
Outros fatores	53	4,0%
Total	1.315	100,0%

2.7.4 – De acordo com o Sistema Nacional de Informações Tóxico-Farmacológicas (Sinitox) em 2005 foram registrados 23.647 casos de intoxicação humana no Brasil por animais peçonhentos. Desse total, 8.208 foram atribuídos a escorpiões, 4.944 a serpentes, 4.661 a aranhas e 5.834 a outros animais peçonhentos. Apresente esses dados em uma tabela.

TABELA 2.15
Casos de intoxicação humana por animal peçonhento, ocorridos no Brasil em 2005, segundo o animal.

	Total	Porcentagem
Escorpião	8.208	34,71
Serpente	4.944	20,91
Aranha	4.661	19,71
Outros animais	5.834	24,67
Total	23.647	100,00

Fonte: Sinitox (2005)⁵

2.7.5 – Construa uma tabela de distribuição de freqüências para apresentar os dados da Tabela 2.16.

TABELA 2.16
Pressão arterial, em milímetros de mercúrio, de cães adultos anestesiados.

130	105	120	111	99	116	82
107	125	100	107	120	143	115
135	130	135	127	90	104	136
100	145	125	104	101	102	101
134	158	110	102	90	107	124
121	135	102	119	115	125	117
107	140	121	107	113	93	103

Para determinar o número de classes pode ser usada a fórmula:

$$k = 1 + 3,222 \times \log n$$

onde n é igual a 49. Então,

$$k = 1 + 3,222 \times \log 49 = 6,4$$

De acordo com a fórmula, podem ser constituídas seis ou sete classes. Como o menor valor observado é 82 e o maior valor é 158, é razoável construir classes com intervalos iguais a 10, a partir de 80. O número de classes será, então, oito, um pouco maior do que o estabelecido pela fórmula.

⁵<http://www.saude.rj.gov.br/animaispeconhentos/estatisticas.html>. Disponível em 30 de maio de 2008.

TABELA 2.17
Distribuição da pressão arterial, em milímetros de mercúrio, de cães adultos anestesiados.

Classe	Número
80 – 90	1
90 – 100	4
100 – 110	16
110 – 120	8
120 – 130	9
130 – 140	7
140 – 150	3
150 – 160	1

2.7.6 – Imagine⁶ que você quer comparar as distribuições de freqüências da mesma variável, para homens e mulheres, separadamente. No entanto, o número de mulheres é consideravelmente maior do que o número de homens. Você compararia as freqüências ou as freqüências relativas? Por quê? Dê um exemplo.

Você deve comparar as freqüências relativas. As freqüências não são comparáveis, uma vez que as amostras são de tamanhos diferentes. Imagine que são 200 mulheres e 50 homens e que, para uma dada classe, a freqüência seja de quatro, em ambas as distribuições. Isto significa 2% das mulheres ($4/200 = 0,02$) e 8% dos homens ($4/50 = 0,08$), uma diferença muito grande.

2.8 – EXERCÍCIOS PROPOSTOS

2.8.1 – Especifique o tipo das seguintes variáveis: a) peso de pessoas; b) marcas comerciais de um mesmo analgésico (mesmo princípio ativo); c) temperatura de pessoas; d) quantidade anual de chuva na cidade de São Paulo; e) religião; f) número de dentes permanentes irrompidos em uma criança; g) número de bebês nascidos por dia em uma maternidade; h) comprimento de cães.

2.8.2 – Faça uma tabela para mostrar que, das 852 pessoas entrevistadas sobre determinado assunto, 59 não tinham opinião ou não conheciam o assunto, 425 eram favoráveis e as demais eram contrárias.

⁶MINIUM, E. W., CLARKE, R. C., COLADARCI, T. *Elements of Statistical Reasoning*. New York, Wiley, 2ed. 1999. p.33.

2.8.3 – Complete a Tabela 2.18.

TABELA 2.18
Distribuição das notas de 200 alunos.

<i>Nota do aluno</i>	<i>Freqüência</i>	<i>Freqüência relativa</i>
De 9 a 10		0,08
De 8 a 8,9	36	
De 6,5 a 7,9	90	
De 5 a 6,4	30	
Abaixo de 5	28	
Total	200	1,0

2.8.4 – Uma doença pode ser classificada em três estágios (leve, moderado e severo). Foram examinados 20 pacientes e obtidos os dados: moderado, leve, leve, severo, leve, moderado, moderado, moderado, leve, leve, severo, leve, moderado, moderado, leve, severo, moderado, moderado, moderado, leve. Com base nestes dados: a) determine a freqüência de cada categoria; b) calcule a freqüência relativa de cada categoria.

2.8.5 – Qual é o erro na distribuição de freqüências dada em seguida?

<i>Classe</i>
20 – 30
30 – 40
40 – 50
60 – 70
70 e mais

2.8.6 – São dados os tipos de sangue de 40 doadores que se apresentaram no mês em um banco de sangue: B; A; O; A; A; A; B; O; B; A; A; AB; O; O; A; O; O; A; A; B; A; A; O; O; A; O; A; O; O; A; O; AB; O; O; A; AB; B; B. Coloque os dados em uma tabela de distribuição de freqüências.

2.8.7 – Dos 80 alunos que fizeram um curso de Estatística, 70% receberam grau B e 5% grau C. Quantos (freqüência) alunos receberam grau A, supondo que não tenha sido conferido nenhum outro grau?

2.8.8 – Foram avaliadas, por cirurgiões-dentistas com especialização em Ortodontia, crianças no estágio de dentadura decidua, entre 3 e 6 anos de idade. Não tinham hábitos de sucção 615. Das demais, 190 tinham o hábito de sucção do polegar, 588 usavam chupeta, 618 usavam mamadeira. Apresente os dados em tabela. Calcule o total e as freqüências relativas.

2.8.9 – Os pesos dos bombeiros que trabalham em determinada cidade variam entre 70 kg e 118 kg. Indique os limites de 10 classes nas quais os pesos dos bombeiros possam ser agrupados.

2.8.10 – O número de enfermeiros em serviço varia muito em um hospital. Foi feita uma distribuição de freqüências com as seguintes classes: 20 |– 25; 25 |– 30; 30 |– 35; 35 |– 40; 40 |– 45; 45 |– 50. Qual é o intervalo de classes e qual é o intervalo de toda a distribuição?

2.8.11 – Construa uma tabela de distribuição de freqüências para apresentar os dados da Tabela 2.19, usando intervalos de classes iguais. Depois faça outra tabela, com os seguintes intervalos: 1 dia, 2 ou 3 dias, de 4 a 7 dias, de 8 a 14 dias, mais de 14 dias.

TABELA 2.19
Tempo de internação, em dias, de pacientes acidentados no trabalho em um dado hospital.

7	8	1	7	13	6
12	12	3	17	4	2
4	15	2	14	3	5
10	8	9	8	5	3
2	7	14	12	10	8
1	6	4	7	7	11

2.8.12 – São dados o valor máximo e o valor mínimo de dois conjuntos, A e B, de dados; no primeiro conjunto, $n = 50$ e no segundo, $n = 100$. No conjunto A, o valor mínimo é 24 e o valor máximo é 70; no conjunto B, o valor mínimo é 187 e o valor máximo é 821. Dê os intervalos de classe para cada conjunto.

2.8.13 – Com base nos dados apresentados na Tabela 2.20, calcule o percentual de pacientes que abandonaram o tratamento contra a tuberculose pulmonar (taxa de abandono), segundo a zona de moradia.

TABELA 2.20
Número de pacientes segundo o abandono do tratamento contra tuberculose pulmonar e a zona de moradia.

<i>Zona de moradia</i>	<i>Abandono do tratamento</i>	
	<i>Sim</i>	<i>Não</i>
Urbana	15	80
Rural	70	35

2.8.14 – Perguntou-se, a 100 dentistas, se eles rotineiramente enfatizavam, no consultório, métodos de prevenção de cáries e doenças gengivais. A resposta de 78 dentistas foi “sim”. Os demais disseram “não”. Apresente estes dados em uma tabela de distribuição de freqüências e discuta os resultados. Os dados mostram que os dentistas adotam a prática da prevenção?

2.8.15 – Calcule as freqüências relativas para os dados apresentados na Tabela 2.21 e comente.

TABELA 2.21
Número de óbitos por grupos de causas. Brasil, 2004.

<i>Grupos de causas</i>	<i>Número</i>	
	<i>Masculino</i>	<i>Feminino</i>
Doenças infecciosas e parasitárias	27.437	18.615
Neoplasias	76.065	64.724
Doenças do aparelho circulatório	150.383	135.119
Doenças do aparelho respiratório	55.785	46.369
Afecções originadas no período perinatal	17.530	13.165
Causas externas	107.032	20.368
Demais causas definidas	88.563	75.399

Fonte: Ministério da Saúde/ SVS- Sistema de Informações sobre Mortalidade - SIM⁷

Notas:

1. As análises devem considerar as limitações de cobertura e qualidade da informação da causa de óbito.
2. Estão suprimidos os óbitos sem definição de causa.

⁷<http://tabnet.datasus.gov.br/cgi/tabcgi.exe?edb2006/c04.def>. Disponível em 4 de maio de 2008.

2.8.16 – Calcule as freqüências relativas para os dados apresentados na Tabela 2.22 e aponte a faixa etária de maior risco.

TABELA 2.22

Pacientes portadores de carcinoma epidermóide de base de língua, segundo a faixa etária, em anos.

Faixa etária	Número
30 – 40	10
40 – 50	66
50 – 60	119
60 – 70	66
70 – 80	24
80 e mais	5

2.8.17 – Com base nos dados apresentados na Tabela 2.23, calcule o percentual de órgãos aproveitados (taxa de aproveitamento para cada órgão).

TABELA 2.23

Número de órgãos obtidos de doadores cadáveres.

Órgão	Número de doadores	Número de órgãos aproveitados
Rim	105	210
Coração	105	45
Fígado	105	20
Pulmões	105	17

(página deixada intencionalmente em branco)

Apresentação de Dados em Gráficos

3

(página deixada intencionalmente em branco)

Gráficos ajudam a visualizar a distribuição das variáveis. Neste Capítulo vamos aprender como apresentar dados em gráficos, seguindo as normas nacionais ditadas pela Fundação Instituto Brasileiro de Geografia e Estatística (IBGE)¹. Todo gráfico deve apresentar título e escala. O título deve ser colocado abaixo do gráfico. As escalas devem crescer da esquerda para a direita e de baixo para cima. As legendas explicativas devem ser colocadas, de preferência, à direita do gráfico.

3.1 – APRESENTAÇÃO DE DADOS QUALITATIVOS

3.1.1 – Gráfico de barras

O *gráfico de barras* é usado para apresentar variáveis qualitativas, sejam elas nominais ou ordinais. Para construir um *gráfico de barras*:

- Desenhe o sistema de eixos cartesianos.
- Escreva as categorias da variável estudada no eixo das abscissas (eixo horizontal).
- Escreva as freqüências ou as freqüências relativas (porcentagens) no eixo das ordenadas (eixo vertical), obedecendo a uma escala.
- Desenhe barras verticais de mesma largura para representar as categorias da variável em estudo. A altura de cada barra deve ser dada pela freqüência ou pela freqüência relativa (geralmente em porcentagem) da categoria.
- Coloque legendas nos dois eixos e título na figura.

Exemplo 3.1: Um gráfico de barras.

Foram entrevistadas 100 pessoas que haviam se submetido a uma cirurgia estética reparadora. Perguntadas se consideravam que a cirurgia havia melhorado a aparência delas, responderam como segue: 66 disseram que sim, 20 disseram que em parte, 8 disseram que não e 6 não quiseram responder. Organize os dados em uma tabela de distribuição de freqüências e desenhe o gráfico de barras.

¹As normas do IBGE são excelentes. Veja essas normas em: <http://www1.ibge.gov.br/home/estatistica/populacao/censo2000/tabelabrasil111.shtm>. Disponível em 24 de abril de 2008. Veja também: VIEIRA, S. **Elementos de Estatística**, São Paulo, Atlas, 5 ed. 2003.

Solução**TABELA 3.1**
Você acha que a cirurgia melhorou sua aparência?

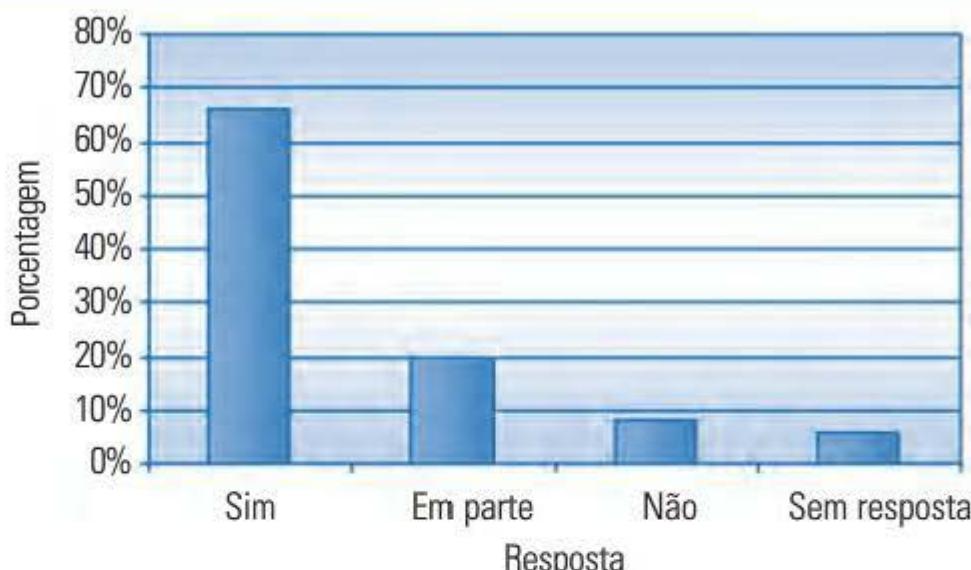
Respostas	Freqüência	Porcentagem
Sim	66	66
Em parte	20	20
Não	8	8
Sem resposta	6	6
Total	100	100

**FIGURA 3.1** Você acha que a cirurgia melhorou sua aparência?

Para facilitar a leitura dos percentuais de cada categoria, podem ser feitas linhas auxiliares (grades).

Exemplo 3.2: Gráfico de barras com grades.

Com os dados da Tabela 3.1, faça um gráfico de barras com linhas auxiliares.

Solução**FIGURA 3.2** Você acha que a cirurgia melhorou sua aparência?

Os percentuais podem ser apresentados acima das barras.

Exemplo 3.3: Gráfico de barras com percentuais nas barras.

Com os dados da Tabela 3.1, faça um gráfico de barras, mas escreva os percentuais acima das barras.

Solução**FIGURA 3.3** Você acha que a cirurgia melhorou sua aparência?

Os gráficos de barras podem ser feitos com perspectiva, isto é, em três dimensões. Por isso, são conhecidos como gráficos em 3D. Eles são agradáveis de ver, mas difíceis de compreender quando apresentam muitas categorias.

Exemplo 3.4: Gráfico de barras com 3D.

Com os dados da Tabela 3.1, faça um gráfico de barras em três dimensões.

Solução

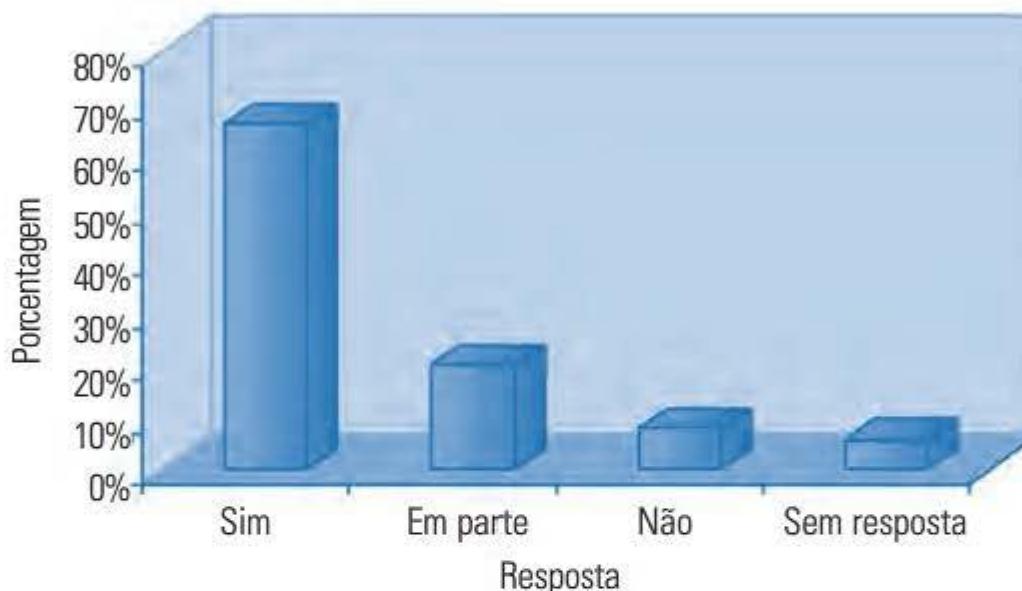


FIGURA 3.4 Você acha que a cirurgia melhorou sua aparência?

Nos gráficos de barras, as barras podem ser apresentadas na posição horizontal, como mostra o Exemplo 3.5.

Exemplo 3.5: Gráfico de barras (horizontais).

Os dados sobre a etiologia de fraturas e corpos estranhos encontrados na face de 46 pacientes, por meio de radiografias panorâmicas feitas em um centro de radiologia, estão na Tabela 3.2. Desenhe um gráfico de barras, mas com as barras em posição horizontal.

Solução

TABELA 3.2
Distribuição dos pacientes quanto à etiologia da fratura ou presença de corpo estranho.

Etiologia	Freqüência
Acidente de trânsito	16
Agressão	13
Arma de fogo	7
Queda	4
Acidente em esportes	2
Assalto	2
Cirurgia ortognática	2
Total	46



FIGURA 3.5 Distribuição dos pacientes quanto à etiologia da fratura ou presença de corpo estranho.

3.1.2 – Gráfico de setores

O gráfico de setores² é especialmente indicado para apresentar variáveis nominais, desde que o número de categorias seja pequeno. Para construir um *gráfico de setores*:

- trace uma circunferência (uma circunferência tem 360°). Essa circunferência representará o total, ou seja, 100%.
- divida a circunferência em tantos setores quantas sejam as categorias da variável em estudo, mas o ângulo de cada setor precisa ser calculado: é igual à *proporção* de respostas na categoria, multiplicada por 360° .
- marque, na circunferência, os ângulos calculados; separe com o traçado dos raios.
- escreva a legenda e coloque título na figura.

Exemplo 3.6: Gráfico de setores.

Por meio de radiografias panorâmicas feitas em um centro de radiologia, foram constatados fraturas e corpos estranhos na face de 46 pacientes, 29 homens e 17 mulheres. Faça um gráfico de setores para mostrar a distribuição por sexo desses pacientes.

Solução

TABELA 3.3
Distribuição por sexo de pacientes com fraturas e corpos estranhos na face.

<i>Sexo</i>	<i>Freqüência</i>	<i>Proporção</i>
Masculino	29	0,63
Feminino	17	0,37
Total	46	1,00

Para fazer o gráfico de setores, é preciso calcular o ângulo de cada setor. Para o sexo masculino, calcule o ângulo:

$$0,63 \times 360 = 226,8$$

e para o feminino, calcule:

$$0,37 \times 360 = 133,2$$

²O gráfico de setores é mais conhecido como gráfico de pizza. Este não é, entretanto, o nome técnico.

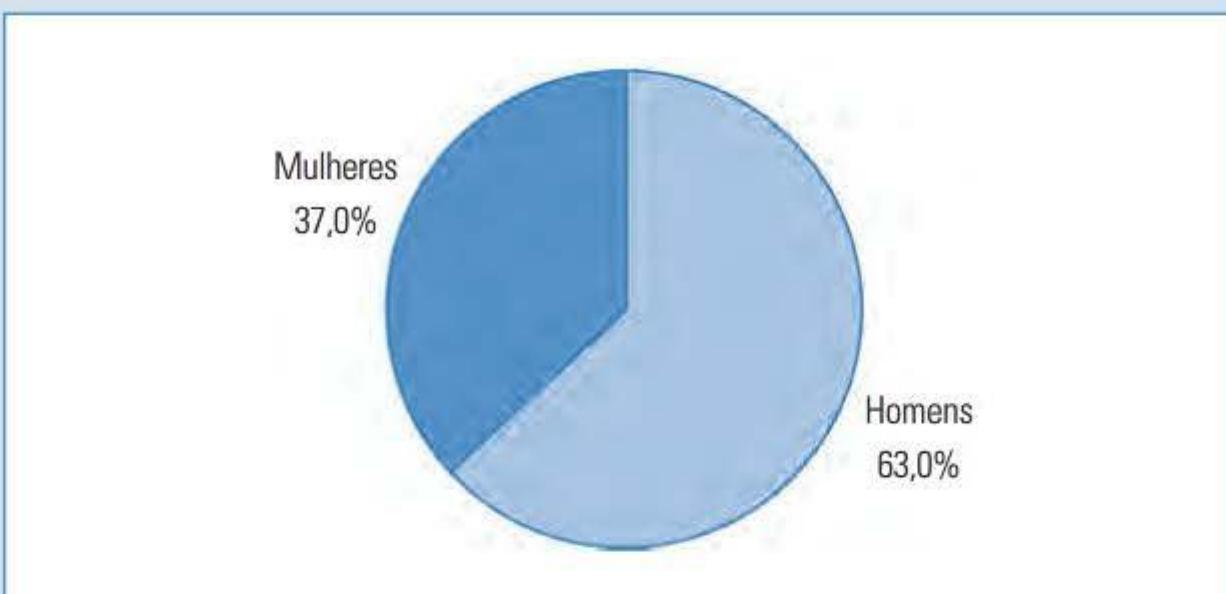


FIGURA 3.6 Distribuição de pacientes com fraturas e corpos estranhos na face segundo o sexo.

Os gráficos de setores podem ser feitos em três dimensões. Esse tipo de apresentação aparece em muitas revistas, mas deve ser evitado porque dificulta a avaliação da proporção de cada categoria.

Exemplo 3.7: Gráfico de setores em 3D.

Com os dados da Tabela 2.3, faça um gráfico de setores em três dimensões.

Solução

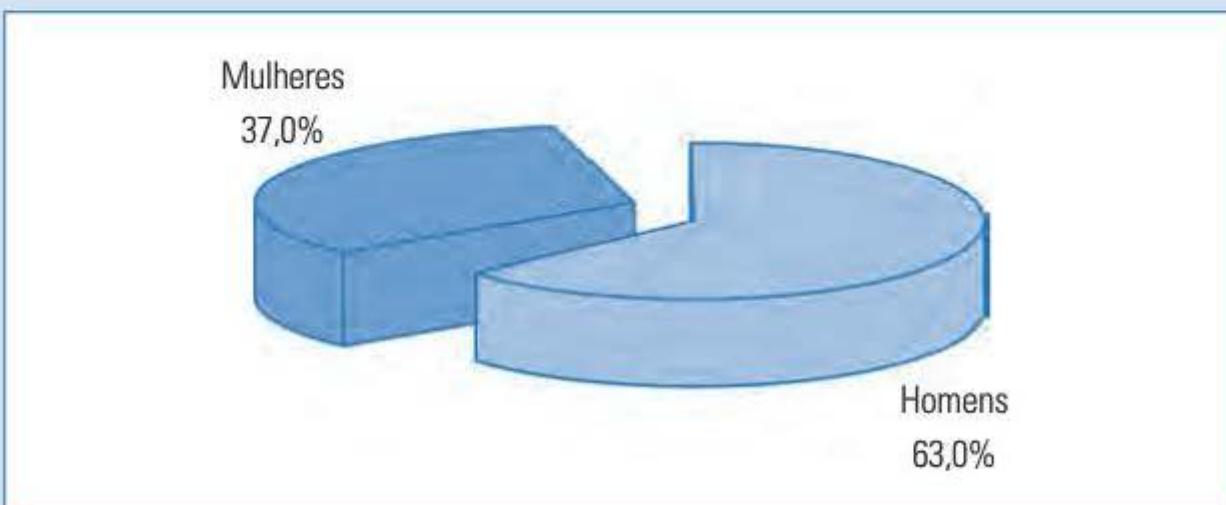


FIGURA 3.7 Distribuição de pacientes com fraturas e corpos estranhos na face segundo o sexo.

3.2 – APRESENTAÇÃO DE DADOS NUMÉRICOS

3.2.1 – Diagrama de linhas

Dados numéricos são, muitas vezes, apresentados em *tabelas de distribuição de freqüências*. Se os dados são *discretos*, as tabelas de distribuição de freqüências apresentam os valores numéricos na ordem natural, em lugar das categorias que aparecem nas distribuições de freqüências de dados qualitativos. Reveja o Exemplo 2.9 do Capítulo 2.

Para construir um *diagrama de linhas*:

- Escreva os valores assumidos pela variável no eixo das abscissas (eixo horizontal).
- Escreva as freqüências ou freqüências relativas (porcentagens) no eixo das ordenadas (eixo vertical).
- Desenhe barras verticais com pequena largura (para evidenciar que os dados são discretos) a partir dos pontos marcados no eixo das abscissas. Os comprimentos das barras são dados pelas freqüências ou pelas freqüências relativas (geralmente em porcentagem).
- Coloque legendas nos dois eixos e título na figura.

Exemplo 3.8: Diagrama de linhas.

As faltas ao trabalho de 30 empregados de uma clínica em determinado semestre estão na Tabela 2.8 do Capítulo 2. A partir dela, foi feita uma tabela de distribuição de freqüências. Faça o diagrama de linhas.

Solução

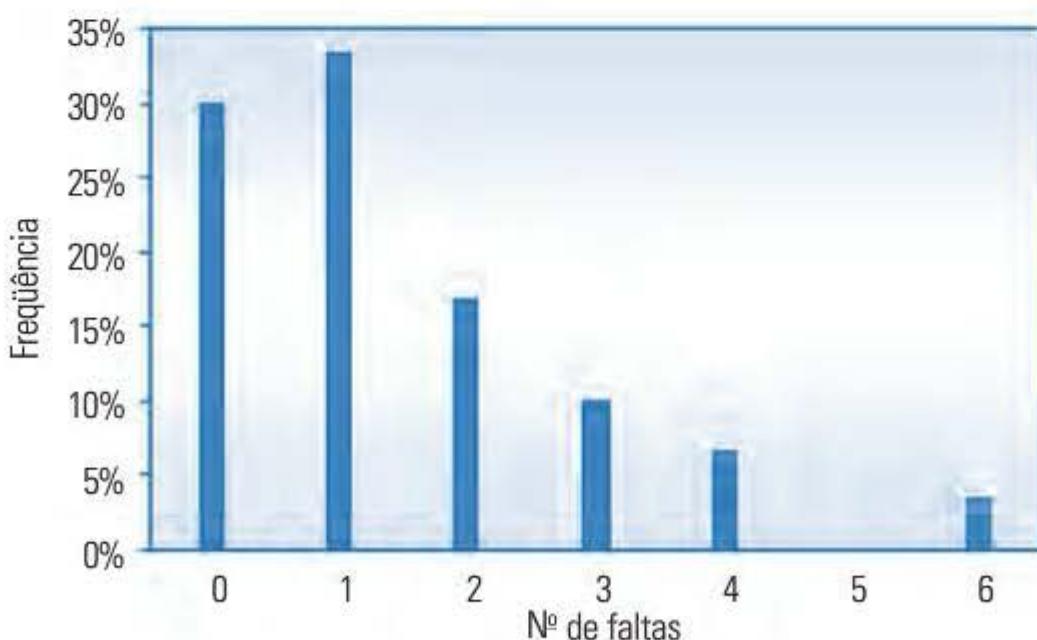


FIGURA 3.8 Diagrama de linhas para a distribuição do número de faltas ao trabalho de 30 empregados de uma clínica no semestre.

3.2.2 – Gráfico de pontos

Os dados contínuos — ao contrário dos discretos — são, na maioria das vezes, uns diferentes dos outros. Veja o Exemplo 3.9: os valores são todos diferentes entre si. Quando em pequeno número, os dados contínuos podem ser apresentados por meio de um gráfico de pontos.

Para fazer um gráfico de pontos (ou diagrama de pontos):

- Desenhe uma linha (na verdade, o eixo das abscissas) com escala, de maneira que nela caibam todos os dados.
- Desenhada a linha, ponha sobre ela pontos que representem os dados, obedecendo à escala.
- Coloque legenda no eixo e título na figura.

Exemplo 3.9: Tempo de sobrevivência após transplante renal.

O número de dias que sete pacientes submetidos a um transplante renal sobreviveram, após a cirurgia em determinado hospital, foi: 17, 5, 48, 120, 651, 64, 150. Apresente esses dados em um gráfico de pontos.

Solução

Para fazer um gráfico de pontos (ou diagrama de pontos), comece desenhando uma linha (eixo das abscissas) que vá do zero até 700, porque o maior número é 651. Desenhada a linha, ponha os pontos que vão representar os dados sobre ela, sempre obedecendo à escala como mostra a Figura 3.9.

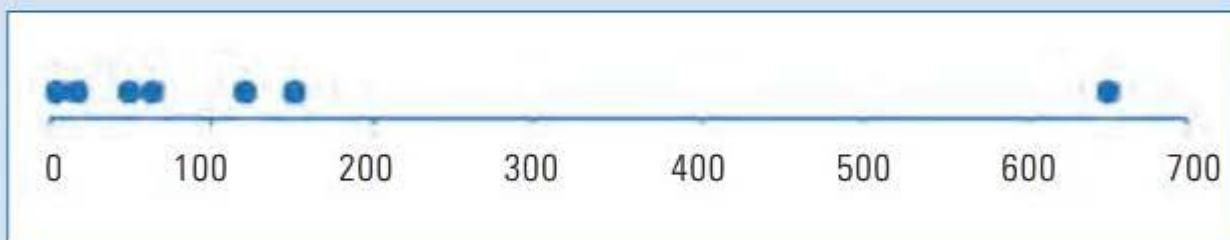


FIGURA 3.9 Diagrama de pontos para os dados de sobrevivência a transplante renal.

3.2.3 – Histograma

Quando os dados são contínuos e a amostra é grande não se pode fazer um gráfico de pontos. É mais conveniente condensar os dados, isto é, organizar uma tabela de distribuição de freqüências³ e desenhar um *histograma*. Para construir um histograma:

- Trace, primeiro, o sistema de eixos cartesianos.

³Faça, de preferência, tabelas de freqüência com intervalos iguais. Se os intervalos de classe forem diferentes, não se pode fazer o histograma como ensinado aqui. Consulte textos mais avançados.

- Apresente as classes no eixo das abscissas. Se os intervalos de classe forem *iguais*, trace barras retangulares com bases iguais, que correspondam aos intervalos de classe.
- Desenhe as barras com alturas iguais às freqüências (ou às freqüências relativas) das respectivas classes. As barras devem ser justapostas, para evidenciar a natureza contínua da variável.
- Coloque legendas nos dois eixos e título na figura.

Exemplo 3.10: Histograma.

Faça um histograma para apresentar os dados mostrados em distribuição de freqüências na Tabela 2.11 do Capítulo 2.

Solução

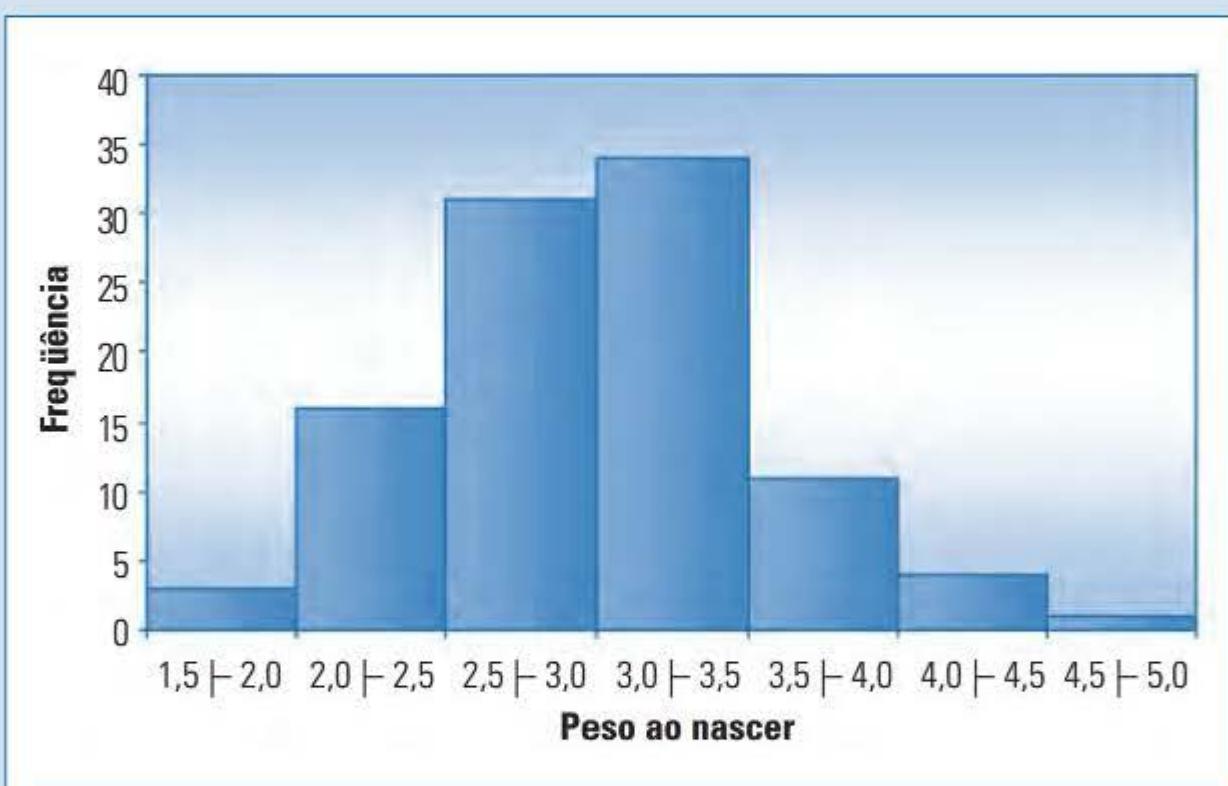


FIGURA 3.10 Histograma para peso ao nascer de nascidos vivos, em quilogramas.

3.2.4 – Polígono de freqüências

Os dados apresentados em tabela de distribuição de freqüências também podem ser mostrados em gráficos denominados *polígonos de freqüências*. Para fazer esse tipo de gráfico:

- Trace o sistema de eixos cartesianos.
- Marque, no eixo das abscissas, pontos que correspondam aos valores centrais⁴ das classes.

⁴Valor central ou ponto médio de uma classe é a média dos dois extremos de classe.

- Marque, no eixo das ordenadas, as freqüências de classe.
- Una os pontos por segmentos de reta.
- Feche o polígono unindo os extremos da figura com o eixo horizontal (nos pontos de abscissas iguais aos valores centrais de uma classe imediatamente inferior à primeira e de uma classe imediatamente superior à última).
- Coloque legendas nos dois eixos e título na figura.

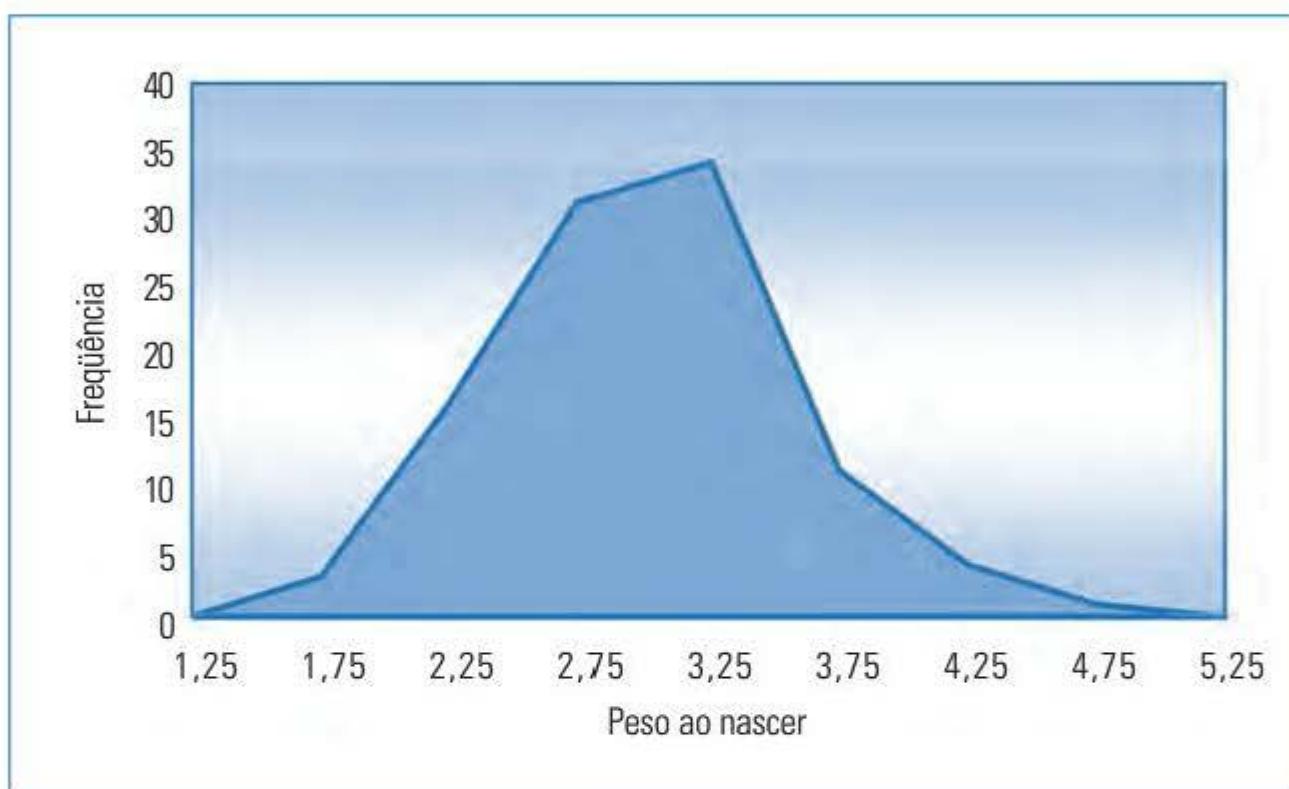


FIGURA 3.11 Polígono de freqüências para peso ao nascer de nascidos vivos, em quilogramas.

3.3 – OBSERVAÇÕES

1. As barras, no gráfico de barras, tanto podem ser desenhadas na posição horizontal como na vertical. A apresentação gráfica é a mesma. Só o programa Excel (muito usado para fazer gráfico) nomeia o gráfico da Figura 3.1 como gráfico de colunas. Se as categorias tiverem nomes extensos como é o caso do Exemplo 3.5, prefira desenhar as barras na posição horizontal, porque isso facilita a leitura.
2. Em geral, as pessoas são mais capazes de comparar comprimentos de barras do que ângulos de gráficos de pizza. Por isso, desenhe pizzas somente quando o número de categorias for pequeno.
3. Se você pretende desenhar um histograma, organize a tabela de distribuição de freqüências com classes iguais.

3.4 – EXERCÍCIOS RESOLVIDOS

3.4.1 – Faça um gráfico de barras e um gráfico de setores para apresentar os dados da Tabela 2.15 do Capítulo 2.

O gráfico de barras está na Figura 3.12 e o gráfico de setores está na Figura 3.13.

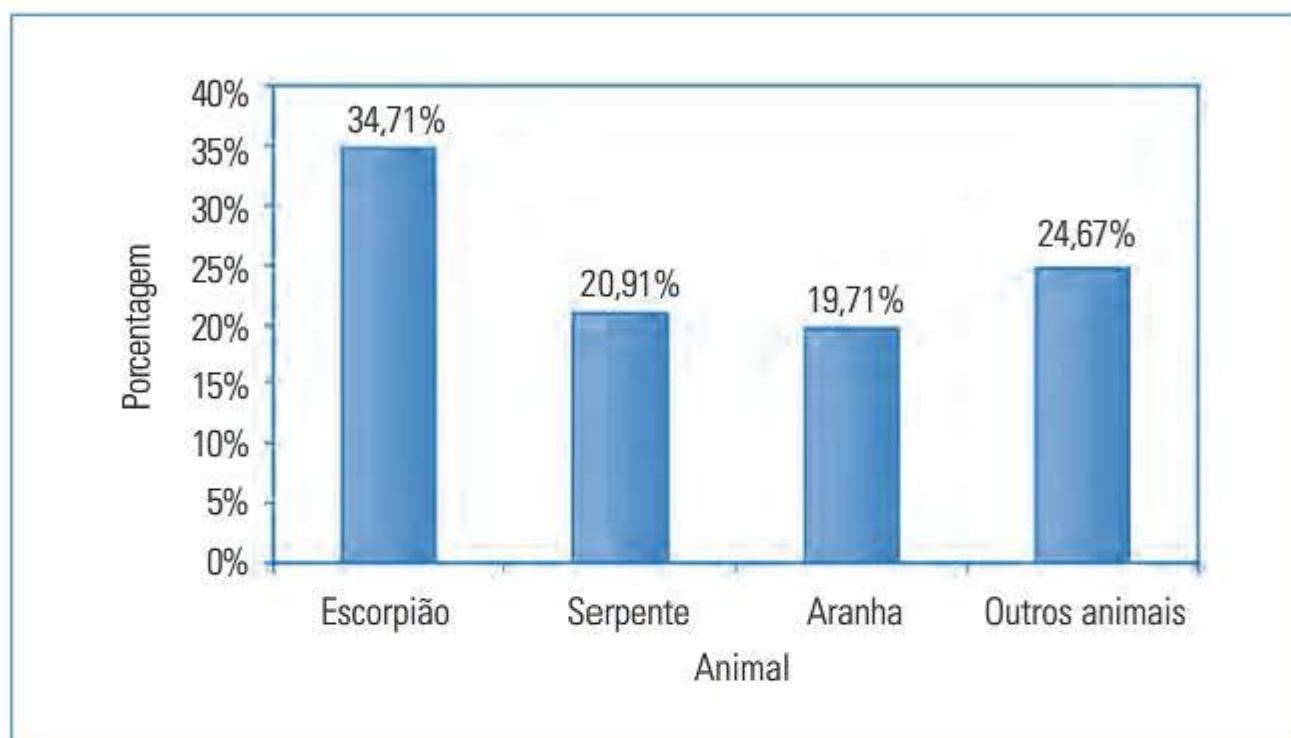


FIGURA 3.12 Casos de intoxicação humana por animal peçonhento, ocorridos no Brasil em 2005, segundo o animal.

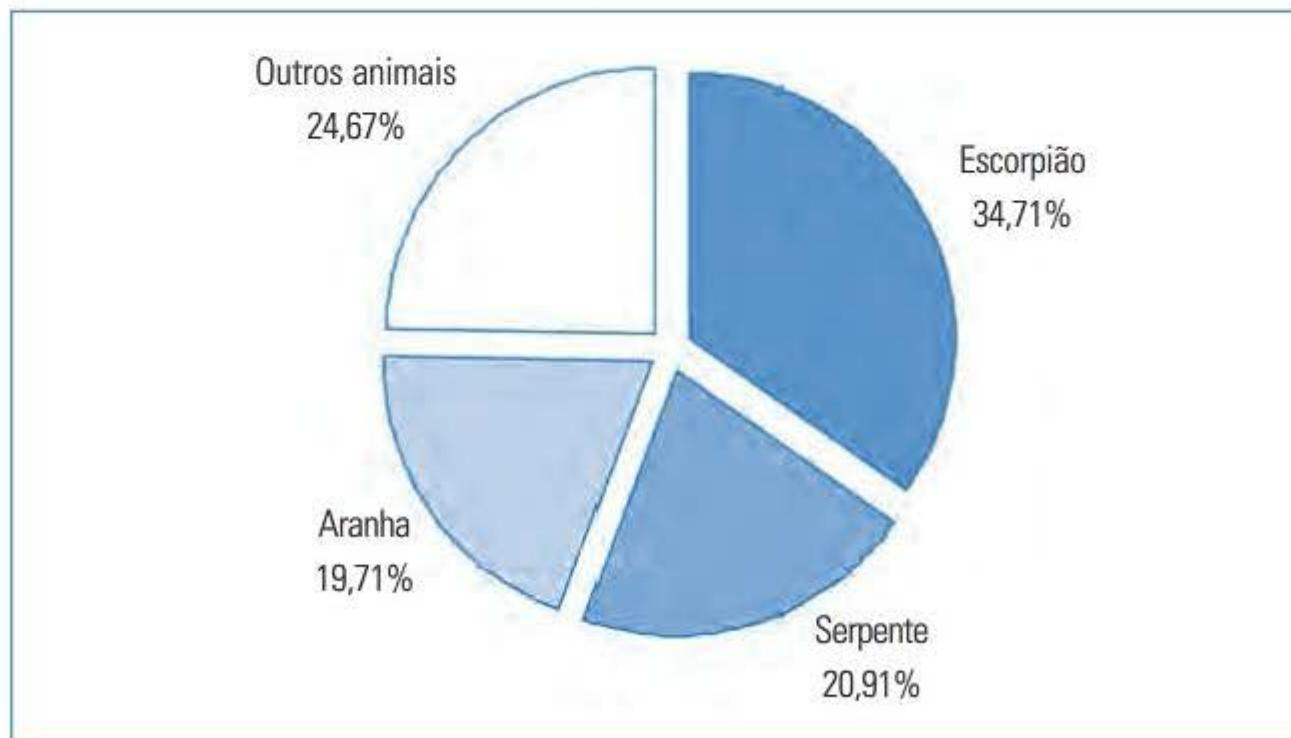


FIGURA 3.13 Casos de intoxicação humana por animal peçonhento, ocorridos no Brasil em 2005, segundo o animal.

3.4.2 – Faça um histograma e um polígono de freqüências para apresentar dados da Tabela 2.17 do Capítulo 2.

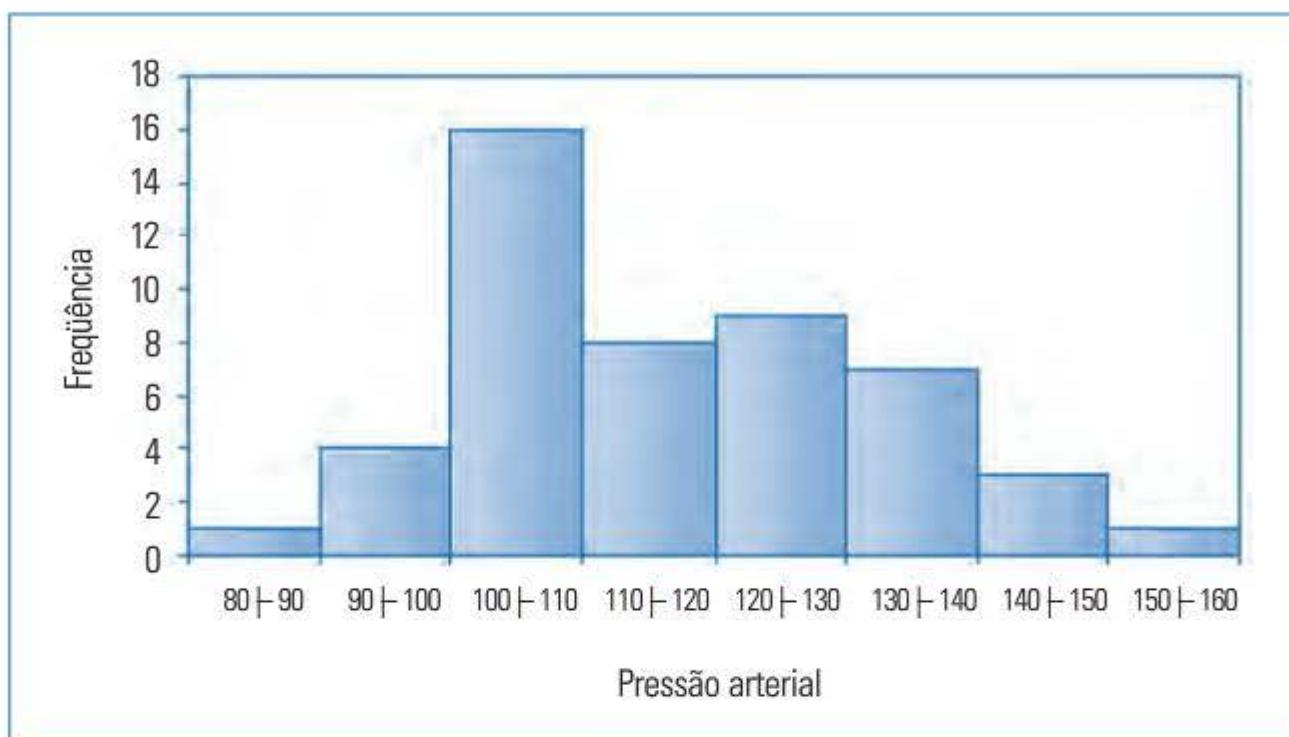


FIGURA 3.14 Distribuição da pressão arterial, em milímetros de mercúrio, de cães adultos anestesiados.

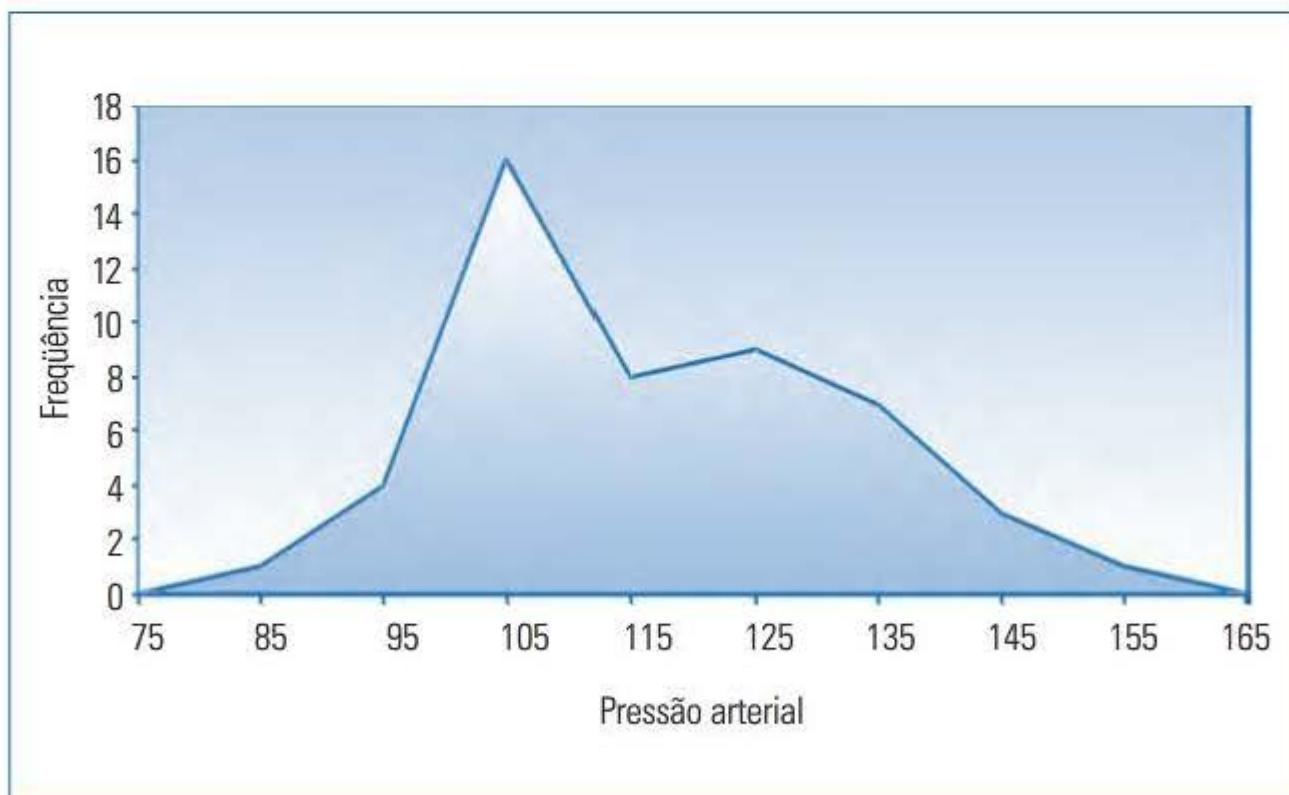


FIGURA 3.15 Distribuição da pressão arterial, em milímetros de mercúrio, de cães adultos anestesiados.

3.4.3 – Por que uma pessoa que conhece determinado assunto preferiria olhar uma tabela de distribuição de freqüências em vez de um gráfico? Qual seria um argumento razoável contra essa postura?

Como podem ser construídos gráficos muito diferentes com base nos mesmos dados, a interpretação, com base apenas neles, pode não ser confiável. Por outro lado, a apresentação gráfica, que faz ressaltar determinadas características dos dados, ajuda o pesquisador. Às vezes, é melhor observar tanto dados como gráfico⁵.

3.4.4 – Quando um gráfico deve ser grande? Quando deve ser pequeno?

O gráfico deve ser grande quando os valores que apresenta precisam ser lidos. Um gráfico pequeno mostra apenas as características gerais do conjunto de dados.

3.5 – EXERCÍCIOS PROPOSTOS

3.5.1 – Desenhe um gráfico de setores para apresentar a distribuição de freqüências que você construiu conforme pedido no Exercício 2.8.4.

3.5.2 – Desenhe um gráfico de barras para apresentar a distribuição de freqüências que você construiu conforme pedido no Exercício 2.8.6.

3.5.3 – Desenhe um gráfico de setores para apresentar a distribuição de freqüências que você construiu conforme pedido no Exercício 2.8.8.

3.5.4 – Desenhe um histograma para apresentar a distribuição de freqüências que você construiu usando intervalos de classes iguais, conforme pedido no Exercício 2.8.11.

3.5.5 – Desenhe dois gráficos de setores (um para cada zona de moradia) para apresentar a distribuição de freqüências que você construiu conforme pedido no Exercício 2.8.13.

3.5.6 – Desenhe um gráfico de barras (as barras na posição horizontal) para apresentar a distribuição de freqüências que você construiu conforme pedido no Exercício 2.8.15.

⁵Veja mais explicações no Capítulo 6.

3.5.7 – Desenhe um histograma para apresentar a distribuição de freqüências que você construiu conforme pedido no Exercício 2.8.16.

3.5.8 – Desenhe um gráfico de barras (as barras na posição horizontal) para apresentar a taxa de aproveitamento para cada órgão, usando os dados apresentados na Tabela 2.23 do Capítulo 2.

3.5.9 – Com base nos dados apresentados na Tabela 3.4, faça uma tabela de distribuição de freqüência. Desenhe um histograma.

TABELA 3.4

Pressão sanguínea diastólica de 30 enfermeiros que trabalham em um hospital.

81	89	91	81	79	82
70	80	92	64	73	86
87	74	72	75	90	96
83	79	82	82	78	85
77	83	85	87	88	80

3.5.10 – Com base nos dados apresentados na Tabela 3.4, faça uma tabela de distribuição de freqüências. Desenhe um polígono de freqüências.

(página deixada intencionalmente em branco)

Medidas de Tendência Central

4

(página deixada intencionalmente em branco)

Muitas pessoas preferem — para entender as características gerais de um conjunto de dados — olhar uma figura¹. Daí a importância dos métodos gráficos descritos no Capítulo 3. No entanto, *medidas numéricas* são mais úteis do que gráficos para mostrar o padrão geral dos dados. Além de serem mais exatas, elas podem ser escritas e faladas. Neste Capítulo, veremos as *medidas de tendência central*. Antes, porém, de descrever essas medidas, precisamos apresentar alguns símbolos matemáticos.

4.1 – SÍMBOLOS MATEMÁTICOS

Para representar uma amostra com n unidades, escrevemos:

$$x_1, x_2, x_3, \dots, x_i, \dots, x_n$$

O subscrito i indica a posição da medida; x_i é a i -ésima observação, num conjunto de n observações. Portanto x_1 representa a primeira observação, x_2 representa a segunda e assim por diante.

Exemplo 4.1: Peso de bebês.

São dados os pesos, em quilogramas, de cinco recém-nascidos em um hospital, na ordem em que eles nasceram:

$$3,500; 2,850; 3,370; 2,250; 3,970.$$

Escreva esse conjunto de dados na notação geral e identifique n .

Solução

Em termos dos símbolos, podemos escrever:

$$x_1 = 3,500; x_2 = 2,850; x_3 = 3,370; x_4 = 2,250; x_5 = 3,970.$$

O último subscrito, — no caso, 5 — dá o tamanho da amostra.

Com relação ao Exemplo 4.1 na seqüência x_1, x_2, x_3, x_4, x_5 , não existe ordem com relação à grandeza dos dados. O bebê menor não é, necessariamente, o primeiro da amostra, nem o bebê maior precisa ser o último. Qualquer que for a amostra, os valores $x_1, x_2, x_3, \dots, x_n$ estarão na ordem em que foram coletados. Os pontos significam “e assim por diante”.

A soma dos valores $x_1, x_2, x_3, \dots, x_n$ é escrita como segue:

$$x_1 + x_2 + x_3 + \dots + x_n$$

ou de forma muito mais compacta:

$$\sum_{i=1}^n x_i$$

¹Já disse alguém: “Um desenho vale por mil palavras”.

que se lê *somatório de x índice i, i de 1 a n.* O símbolo Σ que indica o somatório, é a letra grega sigma maiúscula. O subscrito $i = 1$, sob Σ , indica que o índice i deve ser substituído por números inteiros em ordem crescente sucessivamente, começando por 1 e terminando em n .

Exemplo 4.2: A notação de somatório.

Lembre o exemplo 4.1. Os pesos dos bebês eram:

$$x_1 = 3,500; x_2 = 2,850; x_3 = 3,370; x_4 = 2,250; x_5 = 3,970.$$

Calcule a soma desses pesos, mas faça a indicação da soma usando a notação de somatório.

Solução

Em termos dos símbolos, podemos escrever:

$$\begin{aligned}\sum_{i=1}^5 x_i &= x_1 + x_2 + \dots + x_5 \\ &= 3,500 + 2,850 + 3,370 + 2,250 + 3,970 \\ &= 15,940\end{aligned}$$

Quando é fácil saber o número de parcelas que devem ser somadas pelo próprio texto, podemos escrever apenas $\sum x$ em lugar $\sum_{i=1}^n x_i$.

4.2 – MÉDIA DA AMOSTRA

A medida de tendência central mais conhecida e mais utilizada é a *média aritmética*, ou simplesmente média. Como se calcula uma média?

A média aritmética de um conjunto de dados é obtida somando todos os dados e dividindo o resultado pelo número deles.

$$\text{Média} = \frac{\text{Soma de todos os dados}}{\text{Tamanho da amostra}}$$

A média, que se indica média por \bar{x} (lê-se: x-traço ou x-barra), tem uma fórmula:

$$\bar{x} = \frac{\sum x}{n}$$

que se lê: x-traço é igual ao somatório de x , dividido por n .

Exemplo 4.3: A média da circunferência abdominal de 10 pessoas.

Um professor de Educação Física mediou a circunferência abdominal de 10 homens que se apresentaram em uma academia de ginástica. Obteve os valores, em centímetros: 88; 83; 79; 76; 78; 70; 80; 82; 86; 105. Calcule a média.

Solução

Some todos os dados e divida o resultado pelo tamanho da amostra, que é 10. Então:

$$\bar{x} = \frac{88 + 83 + 79 + 76 + 78 + 70 + 80 + 82 + 86 + 105}{10} = \frac{827}{10} = 82,7$$

ou seja, os homens mediram, em média, 82,7 cm de circunferência abdominal.

A média indica o centro de gravidade do conjunto de dados. Para entender essa afirmativa, observe a Figura 4.1, que apresenta os dados do Exemplo 4.3. Imagine que o eixo das abscissas sejam os braços de uma balança e que cada ponto tenha uma unidade de massa. Para haver equilíbrio, é preciso que o fulcro da balança esteja sob a média, isto é, no ponto em que está a flecha. Então a média é a abscissa do centro de gravidade.



FIGURA 4.1 Distribuição de dados de circunferência abdominal, em centímetros, sobre um eixo e a respectiva média.

Quando a amostra é grande e os dados são discretos, podem ocorrer valores repetidos. Nesses casos, como vimos no Capítulo 2, é razoável organizar os dados em uma *tabela de distribuição de freqüências*. Veja a Tabela 4.1

TABELA 4.1
Uma tabela de distribuição de freqüências.

Dados	Freqüência
x_1	f_1
x_2	f_2
.	.
x_n	f_n
Total	Σf

A média aritmética de dados agrupados em uma tabela de distribuição de freqüências, isto é, de x_1, x_2, \dots, x_n que se repetem f_1, f_2, \dots, f_n vezes na amostra, é

$$\bar{x} = \frac{\sum xf}{\sum f}$$

Exemplo 4.4: A média do número de filhos.

Para calcular a média do número de filhos em idade escolar que têm os funcionários de uma empresa, a psicóloga que trabalha em Recursos Humanos obteve uma amostra de 20 funcionários. Os dados estão apresentados em seguida. Como se calcula a média?

TABELA 4.2
Número de filhos em idade escolar de 20 funcionários.

1	0	1	0
2	1	2	1
2	2	1	5
0	1	1	1
3	0	0	0

Solução

Primeiro, é preciso construir a tabela de distribuição de freqüências. Veja a Tabela 4.3.

TABELA 4.3
Distribuição de freqüências para o número de filhos em idade escolar de 20 funcionários.

<i>Número de filhos em idade escolar</i>	<i>Freqüência</i>
0	6
1	8
2	4
3	1
4	0
5	1

Os cálculos intermediários para obter a média estão na Tabela 4.4. É preciso multiplicar cada valor possível (x) pela respectiva freqüência (f), somar e dividir a soma pelo tamanho da amostra $n = (\sum f)$.

TABELA 4.4
Cálculos auxiliares.

<i>Número de filhos em idade escolar (x)</i>	<i>Freqüência (f)</i>	<i>Produto (xf)</i>
0	6	0
1	8	8
2	4	8
3	1	3
4	0	0
5	1	5
Total	$\sum f = 20$	$\sum xf = 24$

A média é obtida dividindo 24 por 20, que resulta em 1,2 filho em idade escolar por funcionário. Aplicando a fórmula:

$$\bar{x} = \frac{0 \times 6 + 1 \times 8 + 2 \times 4 + 3 \times 1 + 4 \times 0 + 5 \times 1}{6 + 8 + 4 + 1 + 0 + 1} = \frac{24}{20} = 1,2$$

Em certos casos — principalmente quando a variável é contínua e a amostra é grande — são apresentadas apenas as tabelas de distribuição de freqüências — os dados brutos não são fornecidos. Para calcular a média de dados agrupados em classes, é preciso calcular o *valor central* de cada classe. O valor central é a média dos dois extremos de classe. Veja o exemplo 4.5.

Exemplo 4.5: A média de peso ao nascer de nascidos vivos.

No Exemplo 2.11 do Capítulo 2, os dados foram agrupados em faixas de peso. Os nascidos vivos com pesos entre 1,5 (inclusive) e 2,0 kg (exclusive) constituíram a primeira classe, os nascidos vivos com pesos entre 2,0 (inclusive) e 2,5 kg (exclusive) constituíram a segunda classe e assim por diante. Nesse caso, como se calcula a média?

TABELA 4.5
Nascidos vivos segundo o peso ao nascer, em quilogramas.

<i>Classe</i>	<i>Freqüência</i>
1,5 – 2,0	3
2,0 – 2,5	16
2,5 – 3,0	31
3,0 – 3,5	34
3,5 – 4,0	11
4,0 – 4,5	4
4,5 – 5,0	1

Solução

Primeiro, é preciso obter o valor central de cada classe. Para isso, some os valores mínimo e máximo da classe e divida por dois. A classe 1,5 |– 2,0 tem valor mínimo 1,5 e valor máximo 2,0. O valor central da classe é:

$$\frac{1,5 + 2,0}{2} = \frac{3,5}{2} = 1,75$$

A classe 2,0 |– 2,5 tem valor mínimo 2,0 e valor máximo 2,5. O valor central da classe é:

$$\frac{2,0 + 2,5}{2} = \frac{4,5}{2} = 2,25$$

Proceda da mesma forma para obter os demais valores centrais de classe. Para calcular a média, construa uma tabela com os cálculos auxiliares. Escreva as classes, os valores centrais (x^*), as freqüências (f) de classe e os produtos x^*f , como mostra a Tabela 4.6.

TABELA 4.6
Cálculos auxiliares.

Classe	Valor central (x^*)	Freqüência (f)	Produto (x^*f)
1,5 – 2,0	1,75	3	5,25
2,0 – 2,5	2,25	16	36
2,5 – 3,0	2,75	31	85,25
3,0 – 3,5	3,25	34	110,5
3,5 – 4,0	3,75	11	41,25
4,0 – 4,5	4,25	4	17
4,5 – 5,0	4,75	1	4,75
Soma		$\sum f = 100$	$\sum x^*f = 300,00$

A média é obtida dividindo 300 por 100, que dá 3,00 ou, aplicando a fórmula:

$$\bar{x} = \frac{1,75 \times 3 + 2,25 \times 16 + \dots + 4,75 \times 1}{3 + 16 + \dots + 1} = \frac{300}{100} = 3,00$$

ou seja, a média do peso ao nascer, nessa amostra, é 3,00 kg.

A média é, de longe, a medida de tendência central mais usada e, por isso, mais conhecida — quem nunca ouviu falar na *média de aprovação* em determinada disciplina, ou no *tempo médio de uma viagem* (por exemplo, de São Paulo ao Rio de Janeiro) ou na *idade média dos jogadores de futebol*? Em certas circunstâncias, porém, é melhor usar outras medidas de tendência central, como a *mediana* ou a *moda*. Mas o que é mediana e o que é moda?

4.3 – MEDIANA DA AMOSTRA

Mediana é o valor que ocupa a posição central do conjunto dos dados ordenados.

A mediana divide a amostra em duas partes: uma com números menores ou iguais à mediana, outra com números maiores ou iguais à mediana. Quando o número de dados é *ímpar*, existe um único valor na posição central. Esse valor é a mediana. Por exemplo, o conjunto de dados,

$$\{3; 5; 9\}$$

tem mediana 5, porque 5 é o valor que está no centro do conjunto, quando os números são escritos em ordem crescente. Quando o número de dados é *par*, existem dois valores na posição central. A mediana é a média desses dois valores. Por exemplo, o conjunto,

$$\{3; 5; 7; 9\}$$

tem a mediana 6, porque 6 é a média de 5 e 7, que estão na posição central dos números ordenados.

Exemplo 4.6: Calculando a mediana do peso de bebês.

Calcule a mediana do peso, em quilogramas, de cinco bebês nascidos em um hospital, dados no Exemplo 4.1.

Solução

Coloque os dados em ordem crescente como segue:

$$2,250; 2,850; 3,370; 3,500; 3,970.$$

A mediana é o valor que está na posição central, ou seja, 3,370 kg.

Em algumas circunstâncias a mediana mais bem descreve a tendência central dos dados. É o caso dos conjuntos com *dados discrepantes*, isto é, dados de conjuntos que têm um ou alguns valores bem maiores ou bem menores que os demais. Veja o Exemplo 4.7: o valor 42, que é discrepante, “puxa” a média para cima, embora não afete a grandeza da mediana.

Exemplo 4.7: Escolhendo entre média e mediana.

Calcule a média e a mediana dos dados: 42, 3, 9, 5, 7, 9, 1, 9.

Solução

Para obter a média, calcule:

$$\bar{x} = \frac{42+3+9+5+7+9+1+9}{8} = \frac{85}{8} = 10,625$$

Para obter a mediana, é preciso ordenar os dados:

1, 3, 5, 7, 9, 9, 9, 42

Como o número de dados é par, a mediana é a média aritmética dos valores 7 e 9, que ocupam a posição central dos dados ordenados. Então a mediana é 8.

A média é maior do que a mediana porque 42, que é um valor discrepante, "puxa" a média para cima.

Existem casos, porém, em que o uso da média aritmética é mais razoável do que a mediana, mesmo que haja um valor discrepante. Como exemplo, considere que você jogou três vezes na loteria e ganhou:

- na primeira vez, $x_1 = \text{R\$ } 0,00$;
- na segunda vez, $x_2 = \text{R\$ } 0,00$;
- na terceira vez, $x_3 = \text{R\$ } 1.000.000,00$.

Qual medida melhor descreve o seu ganho? A mediana é zero (diga isso aos seus parentes), mas a média é $1/3$ do valor de x_3 (e esse valor diz mais sobre seu ganho nas três tentativas).

4.4 – MODA DA AMOSTRA

Moda é o valor que ocorre com maior freqüência.

Exemplo 4.8: Determinando a moda.

Determine a moda dos dados: 0, 0, 2, 5, 3, 7, 4, 7, 8, 7, 9, 6.

Solução

A moda é 7, porque é o valor que ocorre o maior número de vezes.

Um conjunto de dados pode não ter moda porque nenhum valor se repete maior número de vezes, ou ter duas ou mais modas. Assim, o conjunto de dados

0, 2, 4, 6, 8, 10

não tem moda e o conjunto

1, 2, 2, 3, 4, 4, 5, 6, 7

tem duas modas: 2 e 4.

Quando uma tabela de distribuição de freqüências apresenta grande quantidade de dados, é importante destacar a classe de maior freqüência, a chamada *classe modal*. Essa classe mostra a área em que os dados estão concentrados.

Exemplo 4.9: A moda de idade no Brasil no ano 2000.

É dada a distribuição da população brasileira segundo a faixa de idade, no Censo 2000. Determine a classe modal.

TABELA 4.7
População brasileira presente, segundo a faixa de idade. Brasil, Censo 2000.

Faixa de idade	Número de pessoas
De 0 a 9 anos	32.918.055
De 10 a 19 anos	35.287.882
De 20 a 29 anos	29.991.180
De 30 a 39 anos	25.290.473
De 40 a 49 anos	19.268.235
De 50 a 59 anos	12.507.316
De 60 a 69 anos	8.182.035
De 70 a 79 anos	4.521.889
De 80 a 89 anos	1.570.905
De 90 a 99 anos	236.624
99 anos e mais	24.576
Total	169.799.170

Fonte: IBGE (2003)²

²Em <http://www1.ibge.gov.br/home/estatistica/populacao/censo2000/tabelabrasil111.shtm>. Disponível em 14 de março de 2008.

Solução

A classe modal é “de 10 a 19 anos”, porque é a classe com maior freqüência. Então a moda, no ano 2000, era ter de 10 até 19 anos.

A moda também pode ser usada para descrever dados qualitativos. Nesse caso, a moda é a *categoria* que ocorre com maior freqüência.

Exemplo 4.10: A moda para tipo de sangue.

Veja os dados apresentados na Tabela 4.8. Qual é a moda?

TABELA 4.8
Distribuição de indivíduos segundo o grupo sanguíneo.

<i>Grupo sanguíneo</i>	<i>Freqüência</i>
O	550
A	456
B	132
AB	29
Total	1.167

Solução

Nessa amostra, o grupo sanguíneo O ocorreu com maior freqüência. Então a moda nessa amostra é sangue tipo O.

A moda é bastante informativa quando o conjunto de dados é grande. Se o conjunto de dados for relativamente pequeno (menos de 30 observações), você pode até obter a moda, mas, na maioria das vezes, ela não terá qualquer sentido prático. A média e a mediana fornecem, nesses casos, melhor descrição da tendência central dos dados.

4.5 – EXERCÍCIOS RESOLVIDOS

4.5.1 – Com base nos dados da Tabela 4.9, calcule o peso médio dos ratos em cada idade.

TABELA 4.9
Peso, em gramas, de ratos machos da raça Wistar segundo a idade, em dias.

<i>Número do rato</i>	<i>Idade</i>				
	30	34	38	42	46
1	76	95	99	122	134
2	81	90	101	125	136
3	50	60	62	72	85
4	47	50	57	72	84
5	63	79	82	94	110
6	65	75	79	88	98
7	63	74	79	88	100
8	64	74	92	96	98

Para obter a média aritmética aos 30 dias, basta calcular:

$$\bar{x} = \frac{76 + 81 + 50 + 47 + 63 + 65 + 63 + 64}{8} = \frac{509}{8} = 63,6$$

Da mesma forma, para 34 dias obtém-se:

$$\bar{x} = \frac{95 + 90 + 60 + 50 + 79 + 75 + 74 + 74}{8} = \frac{597}{8} = 74,6$$

As médias para as demais idades são obtidas de maneira idêntica. Essas médias, apresentadas na Tabela 4.10, mostram que o peso médio dos ratos aumenta com a idade.

TABELA 4.10
Médias, em gramas, dos pesos de grupos de oito ratos machos Wistar, segundo a idade, em dias.

<i>Idade</i>	<i>Média</i>
30	63,6
34	74,6
38	81,4
42	94,6
46	105,6

4.5.2 – Determine a mediana dos dados apresentados na Tabela 2.8 do Capítulo 2.

Para obter a mediana, os dados da Tabela 2.8 (faltas ao trabalho de 30 empregados de uma clínica em determinado semestre) foram arranjados em ordem crescente na Tabela 4.11.

TABELA 4.11

Faltas ao trabalho de 30 empregados de uma clínica em determinado semestre, em ordem crescente.

0	0	0	0	0	0	0	0	0	1
1	1	1	1	1	1	1	1	1	2
2	2	2	2	3	3	3	4	4	6

Como o número de dados (30) é par, a mediana é a média aritmética dos dois valores (em negrito) que ocupam a posição central, ou seja, a mediana é 1. Portanto, metade dos empregados faltou um dia ou não faltou no semestre.

4.5.3 – Foi feito um experimento para testar o efeito de um antiinflamatório (droga que tem, também, efeito analgésico) em pacientes com osteoartrite. Os pacientes foram sorteados para receber placebo ($2 \times$ ao dia) ou droga (60 mg $2 \times$ ao dia). Os dados, apresentados na Tabela 4.12, são uma medida da dor à noite (0 = nenhuma dor; 100 = dor extrema), relatada pelo paciente. Calcule as diferenças entre os valores obtidos no final e no início da pesquisa para placebo e para droga. Calcule as médias dessas diferenças. Discuta.**TABELA 4.12**

Dados de dor referidos pelo paciente numa escala de zero a 100, segundo o tratamento.

Placebo		Antiinflamatório	
Início	Final	Início	Final
80	70	80	60
70	50	75	50
75	50	45	25
75	85	50	20
65	65	60	30

TABELA 4.13

Dados de dor referidos pelo paciente numa escala de zero a 100 e diferenças entre início e final do tratamento.

Placebo			Antiinflamatório		
Início	Final	Diferença	Início	Final	Diferença
80	70	-10	80	60	-20
70	50	-20	75	50	-25
75	50	-25	45	25	-20
75	85	10	50	20	-30
65	65	0	60	30	-30
365	320	-45	310	185	-125

Nota: A última linha é o total ou soma.

As médias das diferenças são -9,0 para placebo e -25,0 para o antiinflamatório. A diminuição da dor foi maior quando se usou antiinflamatório.

4.6 – EXERCÍCIOS PROPOSTOS

4.6.1 – Determine média, mediana e moda dos seguintes conjuntos de dados:

- a) 8; 3; 0; 6; 8.
- b) 8; 16; 2; 8; 6.
- c) 4; 16; 10; 6; 20; 10.
- d) 0; -2; 3; -1; 5.
- e) 2;-1; 0; 1; 2; 1; 9.

4.6.2 – Imagine que você está dirigindo um carro numa estrada e observa que o número de carros que você ultrapassa é igual ao número de carros que ultrapassam você. Nesse caso, a velocidade de seu carro corresponde — considerando as velocidades de todos esses carros — a qual medida de tendência central?

4.6.3 – Dado um conjunto de dados, qual das medidas de tendência central (média, mediana e moda) corresponde sempre a um valor numérico do conjunto?

4.6.4 – Quatro pessoas reunidas numa sala têm, em média, 20 anos. Se uma pessoa com 40 anos entrar na sala, qual passa a ser a idade média do grupo?

4.6.5 – Na Tabela 4.14 estão taxas de glicose em miligramas por 100 ml de sangue em ratos machos da raça Wistar com 30 dias de idade, que serão usados em um experimento para o teste de determinada droga. Ache média e mediana.

TABELA 4.14

Taxa de glicose em miligramas por 100 ml de sangue, de oito ratos machos da raça Wistar com 30 dias de idade.

<i>Nº do rato</i>	<i>Taxa de glicose</i>
1	101
2	98
3	97
4	104
5	95
6	105

4.6.6 – Na Tabela 4.15 estão apresentados estaturas, em metros, pesos, em quilogramas, e pressão arterial, em milímetros de mercúrio de pacientes hospitalizados porque tiveram um acidente vascular cerebral (AVC), mais conhecido como derrame. Calcule a média e a mediana para cada variável.

TABELA 4.15

Estaturas, em metros, pesos, em quilogramas, e pressão arterial, em milímetros de mercúrio de 11 pacientes hospitalizados.

<i>Nº do paciente</i>	<i>Estatura</i>	<i>Peso</i>	<i>Pressão arterial</i>
1	1,75	90	180
2	1,58	60	200
3	1,80	80	140
4	1,65	76	220
5	1,80	70	170
6	1,73	65	150
7	1,68	72	140
8	1,65	70	140
9	1,65	75	180
10	1,75	70	160
11	1,65	70	140

4.6.7 – Com os dados apresentados na Tabela 4.16, calcule o número médio de dentes cariados, para cada sexo.

TABELA 4.16
Escolares de 12 anos, segundo o número de dentes cariados e o sexo.

<i>Número de dentes cariados</i>	<i>Sexo</i>	
	<i>Masculino</i>	<i>Feminino</i>
0	16	13
1	2	5
2	3	3
3	2	2
4	2	2

4.6.8 – Para estudar o tempo de latência de um sonífero usando ratos de laboratório, um pesquisador administrou o sonífero a 10 ratos e determinou o tempo que eles demoravam em dormir. Dos 10 ratos, dois demoraram meio minuto, quatro demoraram 1 minuto, três demoraram 1 minuto e meio e um rato não dormiu. Calcule o tempo médio de latência.

4.6.9 – Determine a média, mediana e a moda para cada sexo dos dados apresentados na Tabela 4.17.

TABELA 4.17
Consumo diário de sal, em gramas por dia, segundo o sexo.

	<i>Sexo</i>	
	<i>Masculino</i>	<i>Feminino</i>
	6	4
	9	10
	6	6
	8	8
	7	6
	6	8

4.6.10 – Determine a média, a mediana e a moda, para cada sexo, dos dados apresentados na Tabela 4.18.

TABELA 4.18
Volume diário de urina, em litros, por sexo

<i>Sexo</i>	
<i>Masculino</i>	<i>Feminino</i>
0,5	0,9
1,4	0,6
0,9	0,5
0,8	1,3
1,3	0,8
0,5	0,7

4.6.11 – Determine a mediana e a moda para os dados apresentados na Tabela 4.19 e interprete.

TABELA 4.19
Tempo de retorno, em dias, às atividades de pacientes submetidas a histerectomia.

<i>Nº da paciente</i>	<i>Tempo de retorno</i>
1	20
2	30
3	15
4	20
5	40
6	50
7	25
8	30
9	15
10	35

4.6.12 – Determine a média dos dados apresentados na Tabela 4.20.

TABELA 4.20

Teor de vitamina C (miligramas de ácido ascórbico em 100 ml) em 10 caixas de 100 ml de suco de maçã encontrado no mercado.

<i>Nº da caixa</i>	<i>Teor de vitamina C</i>
1	2,5
2	4,9
3	4,1
4	0,8
5	2,4
6	5,7
7	3,3
8	7,4
9	1,6
10	3,5

4.6.13 – A média, a mediana e a moda podem ser iguais? Dê um exemplo.

4.6.14 – Qual das medidas de tendência central não pode ser calculada para os dados da Tabela 4.21? Por quê?

TABELA 4.21

Número de reclamações recebidas pela diretoria de empregados de uma clínica em determinado semestre, distribuídas segundo o sexo.

<i>Número de reclamações</i>	<i>Sexo</i>	
	<i>Masculino</i>	<i>Feminino</i>
0	16	13
1	8	3
2	3	3
3	2	1
4 ou mais	2	3

Medidas de Dispersão para uma Amostra

5

(página deixada intencionalmente em branco)

As medidas de tendência central resumem a informação contida em um conjunto de dados, mas não contam toda a história. Por exemplo, é fato de observação diária que, na mesma cidade, a temperatura varia ao longo do dia. Ainda, no mesmo dia, registram-se temperaturas muito diferentes em diferentes lugares do mundo. O peso das pessoas varia ao longo da vida e a quantidade de dinheiro que carregam nos bolsos varia em função das circunstâncias. Por causa da *variabilidade*, a média, a mediana e a moda que estudamos no Capítulo 4 não bastam para descrever um conjunto de dados: elas informam a *tendência central*, mas nada dizem sobre a variabilidade.

Para entender este ponto, imagine dois domicílios: no primeiro moram sete pessoas, todas com 22 anos de idade. A média de idade dos moradores desse domicílio coletivo (uma “república”) é, evidentemente, 22 anos. No segundo domicílio também moram sete pessoas: um casal, ela com 17 e ele com 23 anos, dois filhos, um com 2, outro com 3 anos, a mãe da moça, com 38 anos de idade, e um seu outro filho, de 8 anos, e a avó da moça, com 65 anos. Nesse segundo domicílio, a média de idade também é 22 anos. No entanto, “idade média de 22 anos” descreve bem a situação no primeiro domicílio, mas não no segundo.

As medidas de tendência central são tanto mais descriptivas de um conjunto de dados quanto menor for a *variabilidade*. Então, quando você apresenta medidas de tendência central para descrever um conjunto de dados, deve fornecer também uma *medida de variabilidade ou dispersão*. Veremos, neste Capítulo, algumas medidas usadas para medir variabilidade.

5.1 – MÍNIMO, MÁXIMO E AMPLITUDE

O mínimo de um conjunto de dados é o número de menor valor.

O máximo de um conjunto de dados é o número de maior valor.

Para medir variabilidade, você pode fornecer os valores mínimo e máximo do conjunto de dados e calcular a *amplitude* usando a fórmula:

$$\text{amplitude} = \text{máximo} - \text{mínimo}$$

A amplitude de um conjunto de dados, definida como a diferença entre o máximo e o mínimo, é uma medida de dispersão ou variabilidade.

Exemplo 5.1: Mínimo, máximo e amplitude das idades das crianças.

As idades das crianças que estão no pátio de uma escola são: 3, 6, 5, 7 e 9 anos. Faça uma tabela para apresentar o tamanho da amostra, a média, o mínimo, o máximo e a amplitude.

Solução

Para obter a média, você precisa calcular:

$$\bar{x} = \frac{3 + 6 + 5 + 7 + 9}{5} = 6$$

Para obter a amplitude, você ordena os dados como segue: 3, 5, 6, 7, 9. A amplitude é:

$$\text{amplitude} = 9 - 3 = 6$$

TABELA 5.1
Estatísticas das idades das crianças.

Estatísticas	Resultados
Tamanho da amostra	5
Média	6
Mínimo	3
Máximo	9
Amplitude	6

Alguns autores fornecem os valores mínimos e máximos para descrever seus dados e não fornecem a amplitude. Isto está certo, porque esses valores são, muitas vezes, mais úteis. Por exemplo, se alguém informar que os policiais que estão na ativa em certa corporação têm idades entre 18 e 52 anos, estará fornecendo informação mais útil do que se disser que a amplitude das idades é 34 anos. De qualquer modo, a idéia de que os dados de um conjunto têm amplitude de variação é básica em Estatística.

A amplitude é fácil calcular e é fácil de interpretar. Mas essa medida não mede bem a variabilidade por uma razão simples: para calculá-la, usam-se apenas os *dois valores extremos*. Então dois conjuntos de dados podem ter variabilidades diferentes e apresentar a mesma amplitude. Ainda, um valor discrepante — por ser muito grande, ou muito pequeno — faz a amplitude aumentar muito. Como dizem os estatísticos, a amplitude é muito *sensível* aos valores discrepantes.

Exemplo 5.2: Amplitude do barulho do tráfego.

São dados em seguida o barulho do tráfego em duas esquinas, medido em decibéis durante os cinco dias úteis de determinada semana. Calcule as amplitudes.

1^a esquina: 52,0; 54,5; 54,0; 51,0; 54,4; 55,0.

2^a esquina: 54,0; 51,5; 52,0; 51,0; 53,0; 77,1.

Solução

1^a esquina: amplitude = 55,0 – 51,0 = 4,0

2^a esquina: amplitude = 77,1 – 51,0 = 26,1

Note que, na segunda esquina, houve um dia em que o barulho foi bem maior do que nos demais dias da semana. Ocorreu, então, o que os estatísticos chamam de *valor discrepante*. Esse valor (77,1) aumentou, em muito, a amplitude dos dados da segunda esquina.

5.2 – QUARTIL

A mediana, que você viu no Capítulo 4, divide um conjunto de dados em dois subconjuntos com o mesmo número de dados:

- o que antecede a mediana (dados iguais ou menores do que a mediana);
- o que sucede a mediana (dados iguais ou maiores do que a mediana).

Se o número de observações for grande (digamos, maior do que 30), o conceito de mediana pode ser estendido da seguinte forma: a mediana divide o conjunto de dados em *duas metades*; os quartis — como o nome sugere — dividem o conjunto de dados em *quatro quartos*.

Os quartis dividem um conjunto de dados em quatro partes iguais. Os quartis são, portanto, três: o primeiro quartil, o segundo quartil (que é a mediana) e o terceiro quartil.

Para obter os quartis¹:

- Organize os dados em ordem crescente. Ache a mediana (que é, também, o segundo quartil); marque esse valor.

¹Os métodos usados para calcular os quartis têm pequenas diferenças. Se você calcular os quartis para o exemplo 5.3 usando o Excel, encontrará valores diferentes. Os valores calculados aqui são os quartis (em inglês, *quartiles*). O outro método, usado no Excel, calcula as “dobradiças” (em inglês, *hinges*).

- Ache o primeiro quartil, da seguinte forma: tome o conjunto de dados à esquerda da mediana; o primeiro quartil é a mediana do novo conjunto de dados.
- Ache o terceiro quartil, da seguinte forma: tome o conjunto de dados à direita dessa mediana; o terceiro quartil é a mediana do novo conjunto de dados.

Exemplo 5.3: Obtendo os quartis de um conjunto com número ímpar de dados.

Determine os quartis do conjunto de dados: 1, 2, 3, 4, 5, 5, 7, 9, 10.

Solução

Os dados já estão ordenados. Para obter a mediana, observe que o número de dados é ímpar. Então a mediana é o valor central, ou seja, é 5.

1, 2, 3, 4, 5, 6, 7, 9, 10.



Para obter o primeiro quartil, separe os dados menores do que a mediana. A mediana desses dados (2,5) é o primeiro quartil.

1, 2, 3, 4.



Para obter o terceiro quartil, separe os dados maiores do que a mediana. A mediana desses dados (8) é o terceiro quartil.

6, 7, 9, 10.



Lembre-se de que a amplitude é muito *sensível* aos valores discrepantes, isto é, a amplitude pode mudar completamente se for incluída uma observação muito maior, ou muito menor, do que as outras. Então também se define a *distância interquartílica* como medida de dispersão.

Distância interquartílica é a distância entre o primeiro e o terceiro quartil.

$$\text{Distância interquartílica} = \text{Terceiro quartil} - \text{Primeiro quartil}$$

Exemplo 5.4: Distância interquartílica para o barulho do tráfego.

Reveja os dados do exemplo 5.2. Calcule as distâncias interquartílicas.

1^a esquina: 52,0; 54,5; 54,0; 51,0; 54,4; 55,0.

2^a esquina: 54,0; 51,5; 52,0; 51,0; 53,0; 77,1.

Solução

Para achar a distância interquartílica, primeiro ordene os dados. Depois ache os quartis. Então:

Para a 1^a esquina: 51,0; 52,0; 54,0; 54,4; 54,5; 55,0.

Mediana: 54,2

1^o quartil: 52,0

3^o quartil: 54,5.

$$\text{Distância interquartílica} = 54,5 - 52,0 = 2,5.$$

Para a 2^a esquina: 51,0; 51,5; 52,0; 53,0; 54,0; 77,1.

Mediana: 52,5

1^o quartil: 51,5

3^o quartil: 54,0.

$$\text{Distância interquartílica} = 54,0 - 51,5 = 2,5.$$

Note que, embora as amplitudes apresentadas no Exemplo 5.3 sejam muito diferentes, as distâncias interquartílicas são iguais.

5.2.1 – Diagrama de caixa (*Box plot*)

As medidas que acabamos de ver esclarecem a informação contida em um conjunto de dados. O diagrama de caixa mostra isso claramente. Para desenhar o diagrama, são necessárias cinco medidas: mínimo, primeiro quartil, mediana, terceiro quartil, máximo.

Para desenhar um diagrama de caixa:

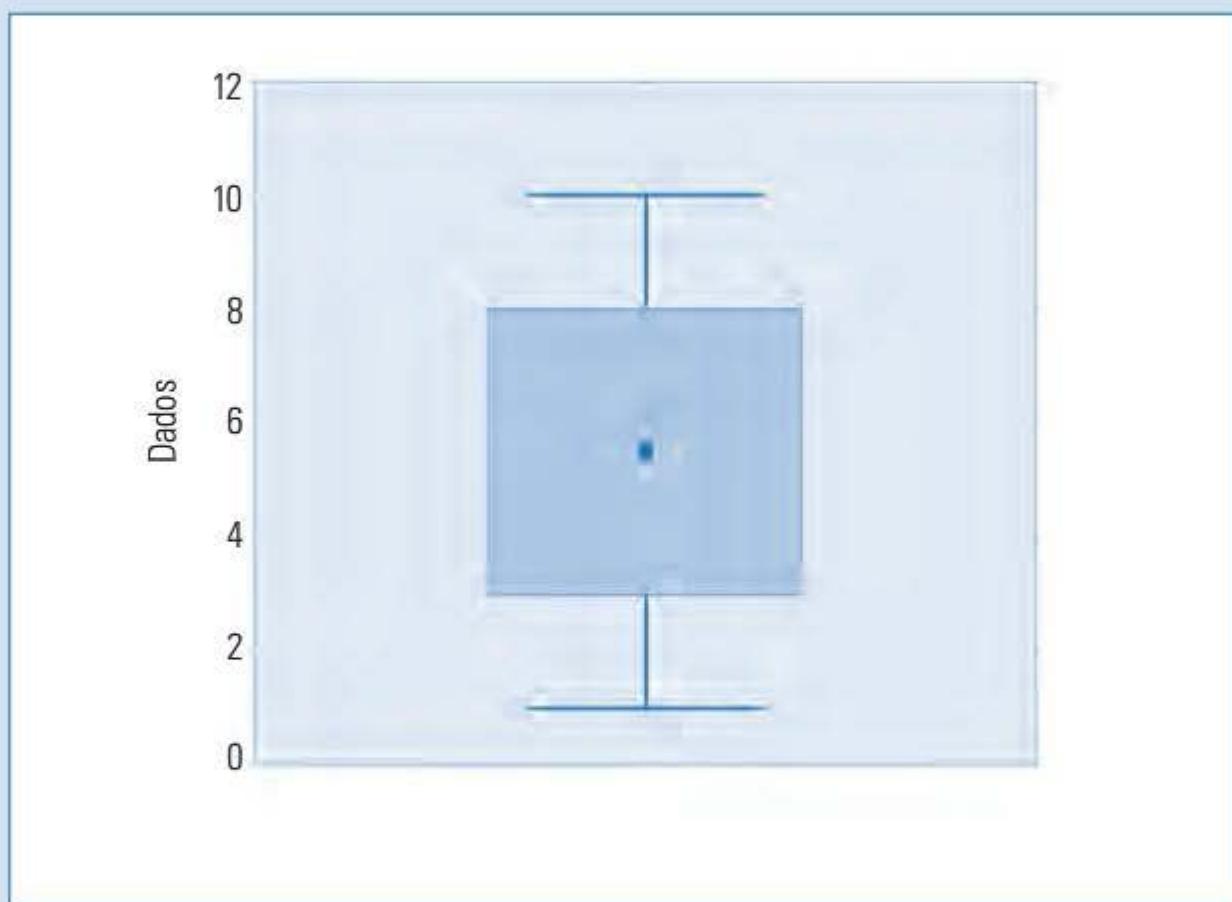
- Desenhe um segmento de reta em posição vertical, para representar a amplitude dos dados.
- Marque, nesse segmento, o primeiro, o segundo e o terceiro quartis.
- Desenhe um retângulo (*box*) de maneira que o lado superior e o lado inferior passem exatamente sobre os pontos que marcam o primeiro e o terceiro quartis.
- Faça um ponto para representar a mediana (obedecendo a escala).

Exemplo 5.5: Um diagrama de caixa.

Desenhe um diagrama de caixa para apresentar o conjunto de dados: 1; 2; 3; 4; 5; 6; 7; 8; 9; 10.

Solução

- Mínimo: 1
- Primeiro quartil: 3
- Mediana: 5,5
- Terceiro quartil: 8
- Máximo: 10.

**FIGURA 5.1** Diagrama de caixa.

O retângulo do diagrama de caixa é dado pela *distância interquartílica*. Esse retângulo contém cerca de 50% dos dados que estão no centro da distribuição.

5.3 – DESVIO PADRÃO DA AMOSTRA

O desvio padrão é uma medida de variabilidade muito recomendada porque mede bem a dispersão dos dados e permite, por conta disso, interpretação de interesse. Mas para calcular o desvio padrão, é preciso, primeiro, calcular a variância. Vamos, então, entender o que é variância.

5.3.1 – Introduzindo a variância

Quando a média é usada como medida de tendência central, ou seja, quando a média indica o centro, podemos calcular o desvio de cada observação em relação à média como segue:

$$\begin{aligned} \text{Desvio} &= \text{observação} - \text{média} \\ d_i &= x_i - \bar{x} \end{aligned}$$

Se os desvios forem pequenos, os dados estão aglomerados em torno da média; logo, a variabilidade é pequena. Por outro lado, desvios grandes significam observações dispersas em torno da média e, portanto, variabilidade grande. Mas veja no Exemplo 5.6 como calcular desvios em relação à média.

Exemplo 5.6: Desvios em relação à média.

Dadas as idades de cinco crianças do Exemplo 5.1, isto é, 3, 6, 5, 7 e 9 anos, calcule os desvios em relação à média.

Solução

Os desvios são obtidos subtraindo a média de cada observação. No caso, a média é 6 anos. Os desvios estão apresentados na Tabela 5.2.

TABELA 5.2
Cálculo dos desvios.

<i>Observação</i> <i>x</i>	<i>Desvio</i> <i>x</i> − \bar{x}
3	$3 - 6 = -3$
6	$6 - 6 = 0$
5	$5 - 6 = -1$
7	$7 - 6 = 1$
9	$9 - 6 = 3$

É preciso resumir todos os desvios em relação à média numa *única* medida de variabilidade. Calcular a média dos desvios pode parecer, à primeira vista, sugestão lógica. No entanto, existem desvios positivos e negativos. A soma dos desvios negativos é sempre igual à soma dos positivos. Aliás, é este o motivo de a média ser uma boa medida de tendência central: o “peso” dos desvios negativos é igual ao “peso” dos desvios positivos. Isto pode ser verificado no Exemplo 5.6:

$$-3 + 0 - 1 + 1 + 3 = 0$$

ou em qualquer outro exemplo.

Para obter uma medida de variabilidade usando os desvios em relação à média, é preciso eliminar os sinais, antes de somar. Uma maneira de eliminar sinais é elevar ao quadrado. A soma assim obtida é chamada *soma dos quadrados dos desvios*. A partir dessa soma, obtém-se a *variância*. Veja a definição de variância da amostra, que se indica por s^2 .

Variância da amostra é a soma dos quadrados dos desvios de cada observação em relação à média, dividida por $(n - 1)$.

$$s^2 = \frac{\sum(x - \bar{x})^2}{n-1}$$

Para calcular a variância:

- calcule os desvios, de cada observação em relação à média;
- eleve cada desvio ao quadrado;
- some os quadrados;
- divida o resultado por $n-1$ (n é o número de observações).

Exemplo 5.7: Calculando a variância.

No Exemplo 5.6 foram calculados os desvios em relação à média para os dados do Exemplo 4.1. Calcule a variância.

Solução

TABELA 5.3
Cálculo da variância.

<i>Observação</i> <i>x</i>	<i>Desvio</i> $x - \bar{x}$	<i>Quadrado do desvio</i> $(x - \bar{x})^2$
3	$3 - 6 = -3$	$(-3)^2 = 9$
6	$6 - 6 = 0$	$0^2 = 0$
5	$5 - 6 = -1$	$(-1)^2 = 1$
7	$7 - 6 = 1$	$1^2 = 1$
9	$9 - 6 = 3$	$3^2 = 9$

A soma dos quadrados dos desvios é:

$$\sum(x - \bar{x})^2 = 20$$

A variância é

$$s^2 = \frac{20}{4} = 5$$

A variância quantifica a variabilidade dos dados em termos de desvios da média ao quadrado mas — embora seja referida como *média dos quadrados dos desvios* — usamos o divisor $n-1$, em lugar de n . Esse divisor, $n-1$, são os *graus de liberdade*² associados à variância.

5.3.2 – Definindo o desvio padrão

É importante notar que o cálculo da variância envolve *quadrados* de desvios. Então a unidade de medida da variância é igual ao *quadrado* da medida das observações. Veja o Exemplo 5.8: as observações são medidas em minutos. Então a variância é dada em minutos ao quadrado, o que não tem sentido prático.

²A soma dos desvios é sempre zero. Então, dados os valores de $n - 1$ desvios, é possível calcular o valor do que estiver faltando. Reveja o exemplo 5.6, que tem $n = 5$ desvios. Dados quatro deles, por exemplo, -3, 0, -1 e 1, é fácil verificar que a soma deles é -3. Para que seja zero, é preciso somar 3 — exatamente o desvio que não foi incluído na soma. Os graus de liberdade representam o número de desvios que estão “livres” para variar (podem ter qualquer valor) — o último está determinado, porque a soma dos desvios é, necessariamente, zero.

Para obter uma medida de variabilidade na mesma unidade de medida dos dados, extrai-se a raiz quadrada da variância. Obtém-se, assim, o desvio padrão.

Desvio padrão é a raiz quadrada da variância, com sinal positivo.

O desvio padrão é uma medida de variabilidade muito usada porque mede bem a dispersão dos dados.

$$s = \sqrt{\text{variância}} = \sqrt{\frac{\sum(x - \bar{x})^2}{n-1}}$$

Exemplo 5.8: Calculando o desvio padrão.

É dada a duração, em minutos, das chamadas telefônicas feitas em três consultórios médicos. Calcule a média, a variância e o desvio padrão.

Solução

TABELA 5.4
Tempo, em minutos, de chamadas telefônicas feitas em uma manhã, em três consultórios médicos.

Consultório A	Consultório B	Consultório C
4	9	9
6	1	1
4	5	1
6	5	2
5	1	8
5	9	9

TABELA 5.5
Estatísticas calculadas.

Estatísticas	Consultório A	Consultório B	Consultório C
Média	5	5	5
Variância	0,8	12,8	16,4
Desvio padrão	0,89	3,58	4,05

A duração, em minutos, das chamadas telefônicas feitas nos três consultórios médicos foi, em média, a mesma, isto é, 5 minutos. No entanto, a duração das chamadas variou muito, de consultório para consultório. Compare, por exemplo, o desvio padrão 0,89 minuto, do consultório A, com o desvio padrão 4,05 minutos, no consultório C.

5.3.3 – Uma fórmula prática para calcular a variância

A fórmula dada na Seção 5.3.1 para calcular a variância da amostra pode ser desenvolvida algebricamente. Obtém-se, então, uma segunda fórmula, que, embora pareça mais complicada à primeira vista, permite que o cálculo da variância seja feito com menor número de operações aritméticas. Prefira esta segunda fórmula, se você faz cálculos à mão.

$$s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}$$

Exemplo 5.9: Calculando a variância pela fórmula prática.

São dados os tempos em minutos que seis meninos permaneceram sobre seus *skates*: 4; 6; 4; 6; 5; 5. Calcule a variância usando a nova fórmula.

Solução

TABELA 5.6
Cálculo da variância.

x	x^2
4	16
6	36
4	16
6	36
5	25
5	25
$\sum x = 30$	
$\sum x^2 = 154$	

Então a variância é:

$$s^2 = \frac{154 - \frac{(30)^2}{6}}{5} = 0,8$$

5.4 – COEFICIENTE DE VARIAÇÃO

O *coeficiente de variação* é a razão entre o desvio padrão e a média. O resultado é multiplicado por 100, para que o coeficiente de variação seja dado em porcentagem. Então:

$$CV = \frac{s}{\bar{x}} \times 100$$

Para entender como se interpreta o coeficiente de variação, imagine dois grupos de pessoas: no primeiro grupo, as pessoas têm idades 3, 1 e 5 anos e a média é, evidentemente, 3 anos; no segundo grupo, as pessoas têm idades 55, 57 e 53 anos, com média de 55 anos.

Observe que, nos dois grupos, a dispersão dos dados é a mesma: ambos têm variância $s^2 = 4$. Mas as diferenças de 2 anos são muito mais importantes no primeiro grupo, que tem média 3, do que no segundo grupo, que tem média 55. Agora, veja os coeficientes de variação. No primeiro grupo, o coeficiente de variação é:

$$CV = \frac{2}{3} \times 100 = 66,67\%$$

e, no segundo grupo, o coeficiente de variação é:

$$CV = \frac{2}{55} \times 100 = 3,64\%$$

Um coeficiente de variação de 66,67% indica que a dispersão dos dados em relação à média é muito grande, ou seja, a *dispersão relativa* é alta. Um coeficiente de variação de 3,64% indica que a dispersão dos dados em relação à média é pequena. Em outras palavras, diferenças de 2 anos são relativamente mais importantes no primeiro grupo, que tem média 3 (o coeficiente de variação é 66,67%), do que no segundo grupo, que tem média 55 (o coeficiente de variação é 3,64%). Então o coeficiente de variação mede a *dispersão dos dados em relação à média*.

É importante notar que o coeficiente de variação pode ser expresso em porcentagem porque é *adimensional*, isto é, não tem unidade de medida. Isto acontece porque média e desvio padrão são medidos na mesma unidade de medida — então elas se cancelam. Por ser adimensional, o coeficiente de variação é útil para comparar a dispersão relativa de variáveis medidas em diferentes unidades. Veja o Exercício 5.5.3.

5.5 – EXERCÍCIOS RESOLVIDOS

5.5.1 – São dados os níveis de colesterol de cinco pessoas: 260; 160; 200; 210; 240. Calcule média e a variância.

TABELA 5.7
Cálculo da média e da variância.

Nível de colesterol	Desvio em relação à média	Quadrado do desvio
260	46	2.116
160	-54	2.916
200	-14	196
210	-4	16
240	26	676

Para obter a média, é preciso calcular a soma dos níveis de colesterol:

$$260 + 160 + 200 + 210 + 240 = 1.070$$

A média é:

$$\bar{x} = \frac{1.070}{5} = 214,0$$

Verifique que a soma dos desvios das observações em relação à média é igual a zero:

$$46 - 54 - 14 - 4 + 26 = 0$$

Para obter a variância, é preciso calcular:

$$46^2 + (-54)^2 + (-14)^2 + (-4)^2 + 26^2 = 2.116 + 2.916 + 196 + 16 + 676 = 5.920$$

A variância é:

$$s^2 = \frac{5.920}{4} = 1.480,00$$

5.5.2 – Dados os seguintes conjuntos de dados, veja qual tem menor variância e quais têm maior variância, sem fazer cálculos.

- a) 7; 7; 7; 7
- b) 6; 7; 7; 8
- c) 6; 8; 10; 12
- d) 106; 108; 110; 112

O conjunto a) tem a menor variância, pois os dados são iguais entre si. Os conjuntos c) e d) têm variâncias iguais (variam de 2 em 2) e maiores do que as dos outros dois.

5.5.3 – Calcule a média, o desvio padrão e o coeficiente de variação dos dados apresentados na Tabela 5.8. Comente os resultados.

TABELA 5.8

Peso, em quilogramas, e comprimento, em centímetros, de 10 cães.

<i>Peso</i>	<i>Comprimento</i>
23	104
22	107
21	103
21	105
17	100
28	104
19	108
14	91
19	102
19	99

- a) Para peso: a média é 20,3 kg e o desvio padrão é 3,74 kg. O coeficiente de variação é 18,42%.
- b) Para comprimento: a média é 102,3 cm e o desvio padrão é 4,85 cm. O coeficiente de variação é 4,74%.

Não se podem comparar desvios padrões de peso e comprimento porque as unidades de medida são diferentes. Mas os coeficientes de variação podem ser comparados porque são adimensionais. É fácil ver que a dispersão relativa dos dados de peso ($CV = 18,42\%$) é maior do que a dispersão relativa dos dados de comprimento ($CV = 4,74\%$). Isto significa que os dados de comprimento variam menos em relação à média do que os dados de peso.

5.5.4 – Determine os quartis³ do conjunto de dados: 1, 2, 2, 5, 5, 7, 8, 10, 11, 11.

Os dados já estão ordenados. Para obter a mediana, note que o número de dados é par. Então a mediana é a média dos dois valores centrais, ou seja, de 5 e 7, que é 6.

$$1, 2, 2, 5, 5, 7, 8, 10, 11, 11.$$

↔

³Os métodos usados para calcular os quartis têm pequenas diferenças. Se você calcular os quartis para o Exemplo 4.5 usando o Excel, encontrará: 1º quartil = 2,75; 3º quartil = 9,5. Não é o método ensinado aqui.

Para obter o primeiro quartil, separe os dados menores do que a mediana 6. O primeiro quartil é a mediana desses dados, ou seja, é 2.

1, 2, 2, 5, 5.

↑

Para obter o terceiro quartil, separe os dados iguais ou maiores do que a mediana. O terceiro quartil é a mediana desses dados, ou seja, é 10.

7, 8, 10, 11, 11.

↑

5.5.5 – Para comparar dois programas de treinamento para executar um serviço especializado, foi feito um experimento. Dez homens foram selecionados ao acaso para serem treinados pelo método A e outros 10 para serem treinados pelo método B. Terminado o treinamento, todos os homens fizeram o serviço e foi registrado o tempo em que cada um desempenhou a tarefa. Os dados estão na Tabela 5.9. Desenhe dois diagramas de caixa e compare.

TABELA 5.9
Tempo, em minutos, despendido em executar o serviço, segundo o método de treinamento.

<i>Método</i>	
<i>A</i>	<i>B</i>
15	23
20	31
11	13
23	19
16	23
21	17
18	28
16	26
27	25
24	28

Método A:

- Mínimo: 11
- Primeiro quartil: 16
- Mediana: 19
- Terceiro quartil: 23
- Máximo: 27

Método B:

- Mínimo: 13
- Primeiro quartil: 19
- Mediana: 24
- Terceiro quartil: 28
- Máximo: 31

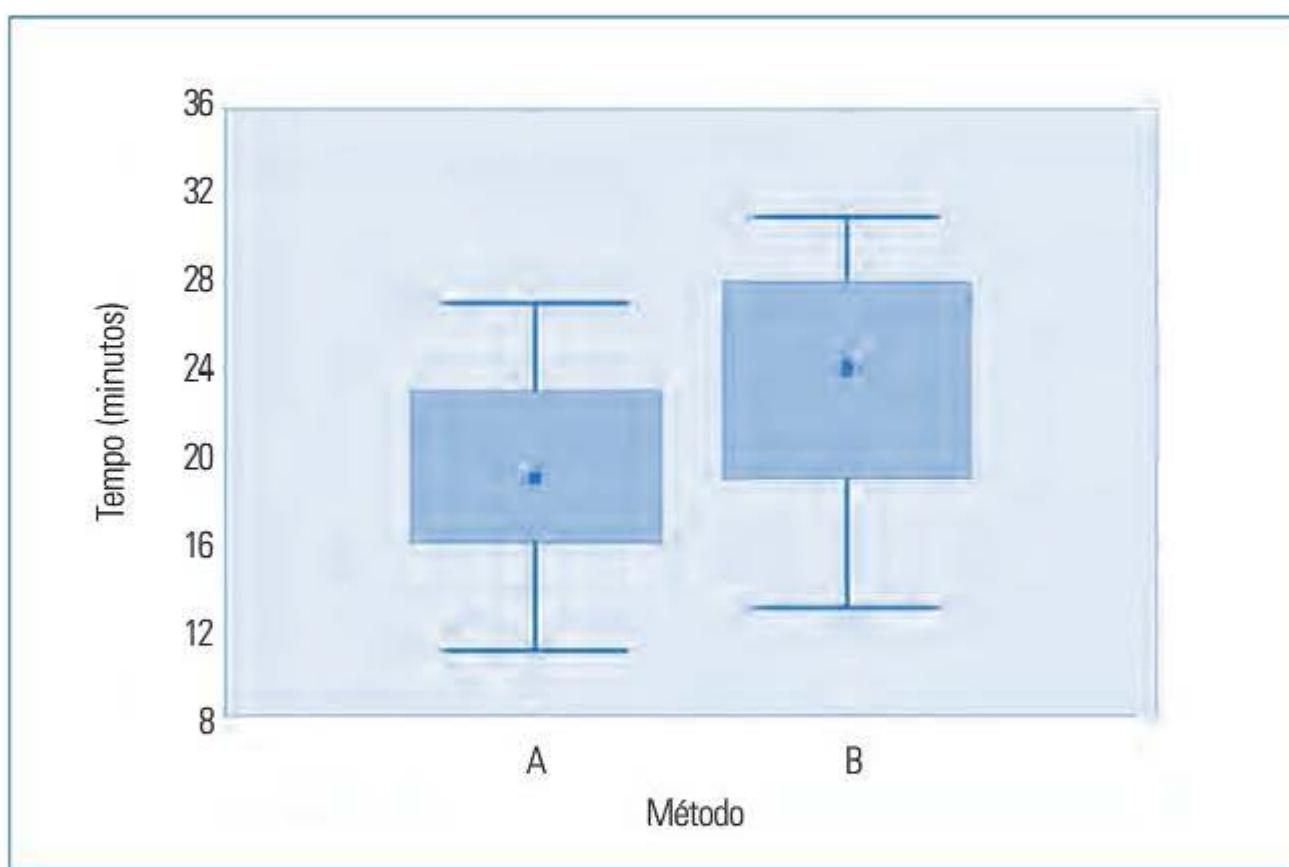


FIGURA 5.2 Comparação de dois diagramas de caixa.

A Figura 5.2 mostra que a mediana do tempo despendido por homens treinados pelo método A foi menor. A variabilidade é praticamente a mesma para os dois métodos. Prefira o método A.

5.5.6 – Calcule a variância e o desvio padrão dos dados apresentados na Tabela 4.9 do Capítulo 4, em cada idade. Comente o resultado.

A variância é dada pela fórmula:

$$s^2 = \frac{\sum x^2 - (\sum x)^2}{n-1}$$

Usando uma calculadora, obtém-se:

a) Para 30 dias de idade:

$$\sum x^2 = 33.305; \sum x = 509; (\sum x)^2 = 259.081$$

b) Para 34 dias de idade:

$$\sum x^2 = 46.043; \sum x = 597; (\sum x)^2 = 356.409$$

c) Para 38 dias de idade:

$$\sum x^2 = 54.765; \sum x = 651; (\sum x)^2 = 423.801$$

d) Para 42 dias de idade:

$$\sum x^2 = 74.417; \sum x = 757; (\sum x)^2 = 573.049$$

e) Para 46 dias de idade:

$$\sum x^2 = 92.041; \sum x = 845; (\sum x)^2 = 714.025$$

Para calcular o desvio padrão basta extrair a raiz quadrada da variância. Os valores dos desvios padrões estão apresentados na Tabela 5.10. É fácil ver que os desvios padrões aumentam com a idade. Portanto, a dispersão dos dados em torno da média aumenta com a idade.

TABELA 5.10
Desvio padrão do peso, em gramas, de grupos de oito ratos machos da raça Wistar, segundo a idade, em dias.

<i>Idade</i>	<i>Desvio padrão</i>
30	11,5
34	14,6
38	16,0
42	19,9
46	20,0

5.6 – EXERCÍCIOS PROPOSTOS

5.6.1 – Dados os valores 5, 3, 2 e 1, ache: a) \sum mínimo; b) o máximo; c) a amplitude.

5.6.2 – Dados os valores 3, 8, 5, 6, 4, 3 e 6, ache: a) $\sum x$; b) $\sum(x - \bar{x})^2$

5.6.3 – Calcule a média e o desvio padrão para o seguinte conjunto de dados: 3; 9; 4; 1; 3.

5.6.4 – A variância de uma amostra é 100 e a soma de quadrados dos desvios é 500. Qual é o tamanho da amostra?

5.6.5 – A média das idades das quatro pessoas que estão reunidas em uma sala é 20 anos e a variância é zero. Se uma pessoa com 40 anos entrar na sala, qual será a idade média do novo grupo e qual será a variância?

5.6.6 – São dadas, na Tabela 5.11, as notas de três alunos em cinco provas. Calcule, para cada aluno, a média e o desvio padrão das notas obtidas. Discuta.

TABELA 5.11
Notas de quatro alunos em cinco provas.

Aluno	1 ^a prova	2 ^a prova	3 ^a prova	4 ^a prova	5 ^a prova
Antônio	5	5	5	5	5
João	6	4	5	4	6
Pedro	10	10	5	0	0

5.6.7 – Responda às questões: a) O valor do desvio padrão pode ser maior do que o valor da média? b) O valor do desvio padrão pode ser igual ao valor da média? c) O valor do desvio padrão pode ser negativo? d) Quando o desvio padrão é igual a zero?

5.6.8 – Calcule a variância, o desvio padrão e o coeficiente de variação para os dados apresentados no Exercício 4.6.5 do Capítulo 4.

5.6.9 – Os tempos de latência em minutos de um analgésico em seis pacientes foram: 4; 6; 4; 6; 5; 5. Calcule a média e a variância.

5.6.10 – Responda às questões: a) qual é a desvantagem de usar a amplitude para comparar a variabilidade de dois conjuntos de dados? b) a variância pode ser negativa? c) a variância pode ser menor do que o desvio padrão?

5.6.11 – Um professor de Odontologia quer saber se alunos que começam a atender pacientes em disciplinas clínicas têm aumento na pressão sistólica. Mediou então a pressão sistólica de cinco alunos de primeiro ano (que não cursam disciplinas clínicas) e de cinco alunos do segundo ano, logo antes do primeiro atendimento de pacientes. Os dados estão na Tabela 5.12. Calcule as médias e os desvios padrões. Discuta.

TABELA 5.12

Pressão sanguínea sistólica, em milímetros de mercúrio, de alunos, segundo o ano que cursavam.

1º ano	2º ano
113	126
121	131
115	146
123	126
118	126

5.6.12 – Para verificar se duas dietas indicadas para pessoas que precisam perder peso são igualmente eficientes, um médico separou, ao acaso, um conjunto de 12 pacientes em dois grupos. Cada paciente seguiu a dieta designada para seu grupo. Decorrido certo tempo, o médico obteve a perda de peso, em quilogramas, de cada paciente de cada grupo. Os dados estão na Tabela 5.13. Calcule as médias e as variâncias. Discuta.

TABELA 5.13

Perda de peso em quilogramas, segundo a dieta.

Dieta	
A	B
8	7
5	8
6	2
7	5
4	12
6	8

5.6.13 – Calcule as médias e os desvios padrões das notas obtidas por alunos dos cursos diurnos e noturnos de uma universidade brasileira, no Exame Nacional de Cursos (Provão), em determinado ano. Compare.

TABELA 5.14

Notas obtidas por alunos de determinada universidade no Exame Nacional de Cursos (Provão), em determinado ano.

<i>Curso</i>	<i>Curso diurno</i>	<i>Curso noturno</i>
Administração	51,2	47,1
Direito	55,1	59
Matemática	43,3	35,7
Letras	46	46,6
Física	43	43
Química	46,6	46,5
Ciências biológicas	49,5	42,6
Pedagogia	63,3	58,2
História	29,3	29,8

Noções sobre Correlação

6

(página deixada intencionalmente em branco)

Você já deve ter ouvido falar que a pressão arterial aumenta quando a idade avança. Você também já deve ter ouvido falar que o desempenho de um atleta melhora com o treinamento. E você provavelmente já ouviu dizer que o número de cárries diminui com uma higiene oral bem-feita. Estes exemplos mostram que existem *relações entre variáveis* ou, em linguagem nada técnica, que existem variáveis que “andam juntas”.

6.1 – DIAGRAMA DE DISPERSÃO

Vamos pensar em duas variáveis numéricas e — só para facilitar — vamos chamar uma delas de X e a outra de Y . Então cada unidade da amostra fornece dois valores numéricos, um referente à variável X , outro referente à variável Y . Você já sabe calcular a média, o mínimo, o máximo e o desvio padrão de cada uma das duas variáveis. Mas, neste Capítulo, vamos buscar responder às questões:

- Existe relação entre as variáveis X e Y ?
- Que tipo de relação existe entre elas?
- Qual é o grau da relação?

Para estudar a relação entre duas variáveis numéricas, você pode fazer um gráfico da seguinte maneira:

- Trace um sistema de eixos cartesianos e represente uma variável em cada eixo.
- Estabeleça as escalas de maneira a dar ao diagrama o aspecto de um quadrado.
- Escreva os nomes das variáveis nos respectivos eixos e faça, depois, as graduações.
- Desenhe um ponto para representar cada par de valores das variáveis.

O gráfico assim obtido é chamado *diagrama de dispersão*. O diagrama de dispersão permite visualizar a relação entre duas variáveis. Se X e Y crescem *no mesmo sentido*, existe uma *correlação positiva* entre as variáveis. Se X e Y variam em *sentidos contrários*, existe *correlação negativa* entre as variáveis.

Exemplo 6.1: Correlação positiva e correlação negativa.

A Tabela 6.1 apresenta dois conjuntos de pares de valores das variáveis X e Y . A correlação é positiva no Conjunto A porque X e Y crescem juntas; a correlação é negativa no Conjunto B porque X cresce enquanto Y decresce. Observe os diagramas de dispersão da Figura 6.1: é mais fácil ver a relação que existe entre as variáveis nos diagramas.

TABELA 6.1
Dois conjuntos de pares de valores de duas variáveis.

<i>Conjunto A</i>		<i>Conjunto B</i>	
<i>X</i>	<i>Y</i>	<i>X</i>	<i>Y</i>
1	2	1	8
2	0	2	12
3	6	3	8
4	3	4	10
5	9	5	4
6	4	6	9
7	10	7	3
8	8	8	6
9	12	9	0
10	8	10	2

Solução

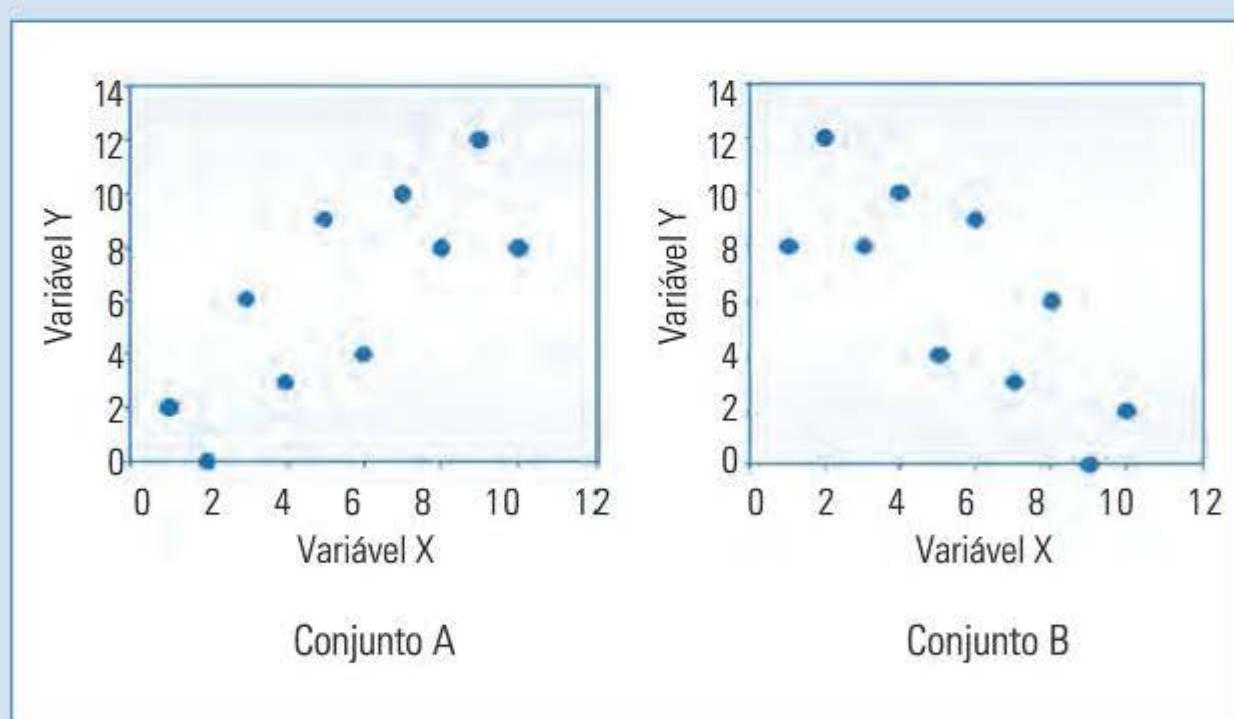


FIGURA 6.1 Correlação positiva (à esquerda) e correlação negativa (à direita).

A correlação será tanto *maior* quanto *menor* for a *dispersão* dos pontos. O Exemplo 6.2 apresenta três gráficos com correlação positiva: quando os pontos estão muito espalhados como no conjunto A, a correlação é *fraca*. Quando os pontos estão concentrados *em torno de uma reta* imaginária como no conjunto B, a correlação é forte.

Exemplo 6.2: Correlação fraca, correlação forte, correlação perfeita.

A Tabela 6.2 apresenta três conjuntos de pares de valores das variáveis X e Y : a correlação é fraca no Conjunto A, é forte no Conjunto B e é perfeita (porque os pontos estão sobre a reta) no Conjunto C. É fácil apreender a intensidade da correlação entre as variáveis de cada um dos conjuntos observando os diagramas de dispersão da Figura 6.2.

TABELA 6.2
Três conjuntos de pares de valores de duas variáveis.

<i>Conjunto A</i>		<i>Conjunto B</i>		<i>Conjunto C</i>	
X	Y	X	Y	X	Y
1	6	1	2	1	3
2	3	2	6	2	4
3	5	3	5	3	5
4	7	4	8	4	6
5	2	5	6	5	7
6	11	6	9	6	8
7	9	7	10	7	9
8	3	8	8	8	10
9	6	9	12	9	11
10	8	10	10	10	12

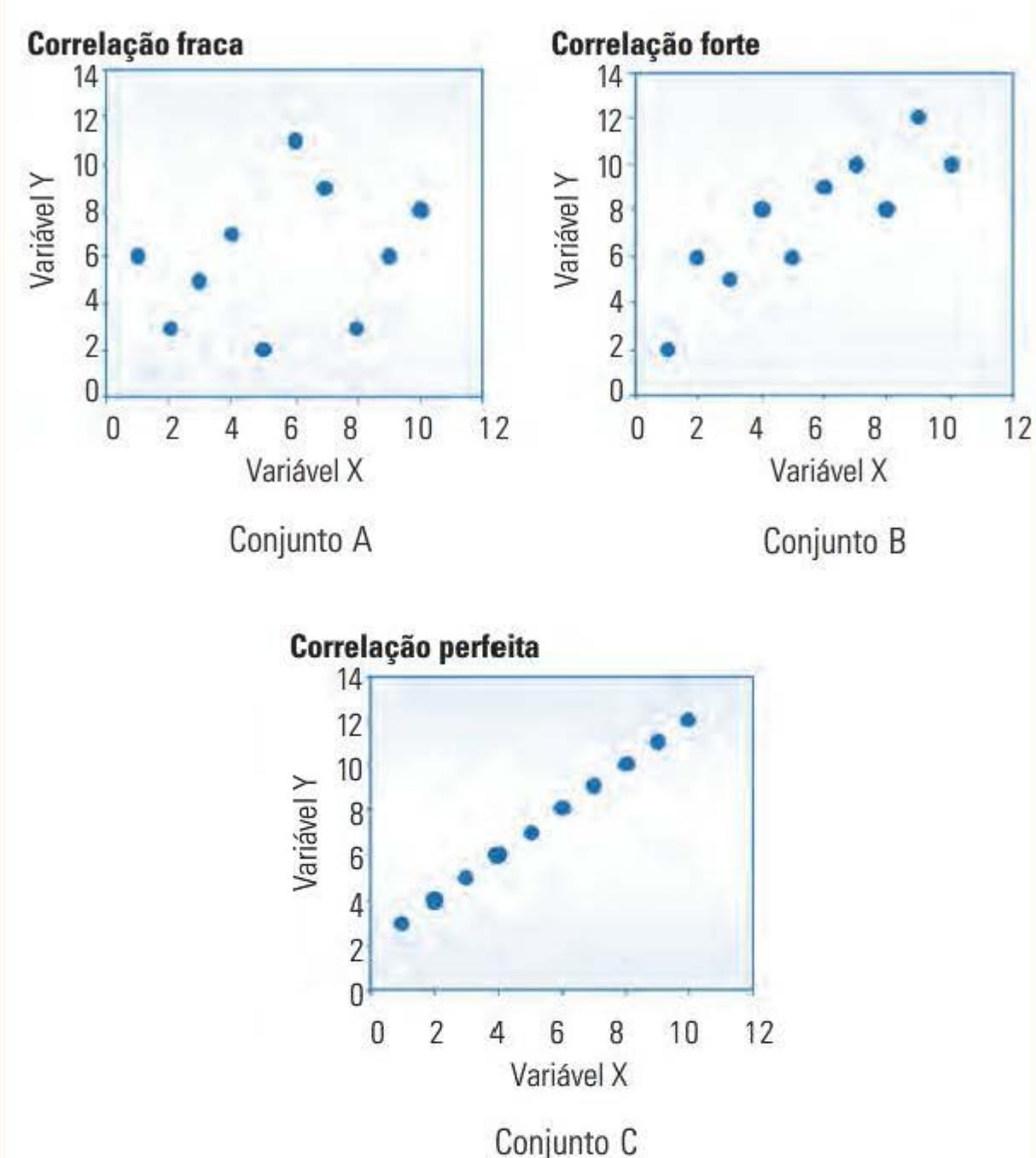


FIGURA 6.2 Correlações fraca, forte e perfeita.

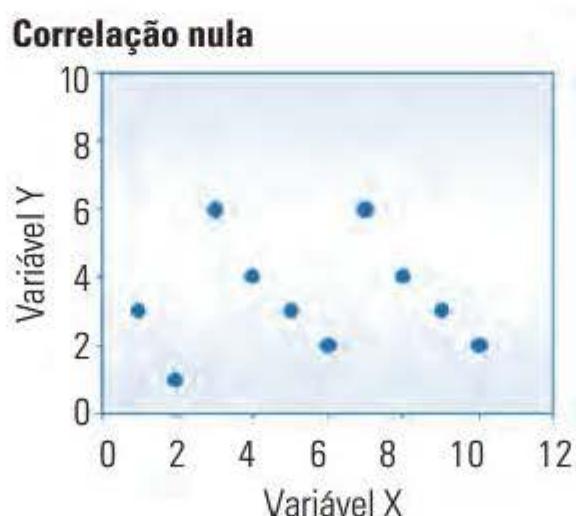
Pode acontecer, no entanto, de a variação de Y não estar relacionada com a variação de X . Nesses casos, o diagrama de dispersão mostra que X cresce e Y varia ao acaso. Dizemos, então, que a correlação entre as variáveis é *nula* ou, o que é o mesmo, que não existe correlação entre as variáveis.

Exemplo 6.3: Correlação nula.

A Tabela 6.3 apresenta um conjunto de pares de valores das variáveis X e Y . O diagrama de dispersão apresentado na Figura 6.3 mostra que não existe qualquer tipo de relação entre as variáveis.

TABELA 6.3
Pares de valores de duas variáveis.

X	Y
1	3
2	1
3	6
4	4
5	3
6	2
7	6
8	4
9	3
10	2

Solução**FIGURA 6.3** Correlação nula.

Quando você olha o diagrama de dispersão, “vê” o tipo de relação entre as variáveis. Se os pontos estão dispersos em torno de uma reta, como acontece nos dois conjuntos de dados mostrados no Exemplo 6.1, a relação entre as variáveis é *linear*. Algumas variáveis têm *relação não-linear*. Veja o Exemplo 6.4: a relação entre as variáveis é não-linear. Neste livro, porém, serão estudadas apenas as relações lineares entre duas variáveis.

Exemplo 6.4: Relação não-linear entre duas variáveis.

Observe o diagrama de dispersão da Figura 6.4, que apresenta os dados X e Y da Tabela 6.4. Note que a relação entre as variáveis é não-linear.

TABELA 6.4
Uma relação não-linear entre duas variáveis.

X	Y
1,5	1,0
2,0	2,0
3,0	3,0
4,0	3,5
5,0	3,0
6,0	2,0
6,5	1,0

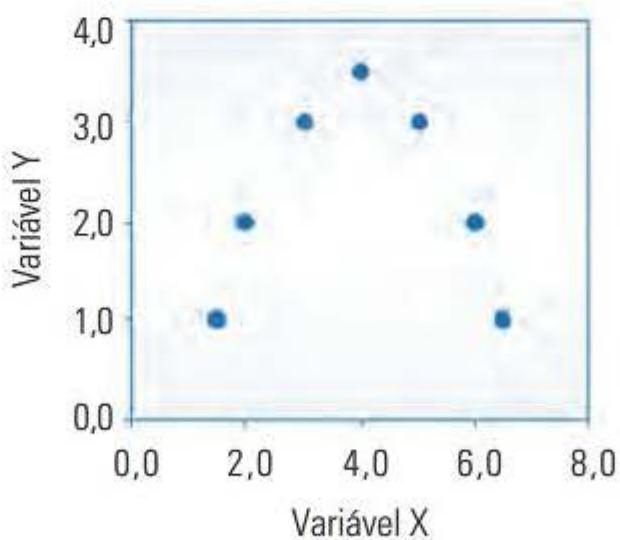


FIGURA 6.4 Uma relação não-linear entre duas variáveis.

6.2 – COEFICIENTE DE CORRELAÇÃO

Existe uma medida para o *grau de correlação linear* entre duas variáveis numéricas¹. Essa medida é o *coeficiente de correlação de Pearson*, que se representa por r e é definido pela fórmula:

$$r = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{\left[\sum X^2 - \frac{(\sum X)^2}{n} \right] \left[\sum Y^2 - \frac{(\sum Y)^2}{n} \right]}}$$

Para entender como se aplica esta fórmula, veja o Exemplo 6.5 e o Exemplo 6.6. Os dados já foram apresentados na Tabela 6.1 e na Figura 6.1.

Exemplo 6.5: Cálculo do coeficiente de correlação.

Reveja os dados apresentados na Tabela 6.1. Calcule o coeficiente de correlação para os dados do Conjunto A.

Para obter o coeficiente de correlação entre X e Y foram feitos os cálculos intermediários que estão na Tabela 6.5. Na última linha dessa tabela estão os somatórios.

TABELA 6.5
Cálculos intermediários para a obtenção do coeficiente de correlação (Conjunto A da Tabela 6.1).

<i>Conjunto A</i>				
X	Y	XY	X^2	Y^2
1	2	2	1	4
2	0	0	4	0
3	6	18	9	36
4	3	12	16	9
5	9	45	25	81
6	4	24	36	16
7	10	70	49	100
8	8	64	64	64
9	12	108	81	144
10	8	80	100	64
$\Sigma X = 55$		$\Sigma Y = 62$	$\Sigma XY = 423$	$\Sigma X^2 = 385$
				$\Sigma Y^2 = 518$

¹ Para estudar a correlação entre variáveis ordinais, calcula-se o coeficiente de correlação de Spearman. Veja em: VIEIRA, S. Bioestatística: Tópicos Avançados. Rio de Janeiro: Campus-Elsevier, 2004.

Substituindo, na fórmula, os somatórios pelos valores calculados na Tabela 6.5 e lembrando que n é o tamanho da amostra (no exemplo $n = 10$), obtemos:

$$r = \frac{423 - \frac{55 \times 62}{10}}{\sqrt{\left[385 - \frac{55^2}{10}\right]\left[518 - \frac{62^2}{10}\right]}}$$

$$r = \frac{82}{\sqrt{82,5 \times 133,6}}$$

$$r = 0,781$$

Exemplo 6.6: Cálculo do coeficiente de correlação.

Reveja os dados apresentados na Tabela 6.1. Calcule o coeficiente de correlação para os dados do Conjunto B.

Para obter o coeficiente de correlação entre X e Y foram feitos os cálculos intermediários apresentados na Tabela 6.6. Na última linha dessa tabela estão os somatórios.

TABELA 6.6
Cálculos intermediários para a obtenção do coeficiente de correlação (Conjunto B da Tabela 6.1).

<i>Conjunto B</i>				
X	Y	XY	X^2	Y^2
1	8	8	1	64
2	12	24	4	144
3	8	24	9	64
4	10	40	16	100
5	4	20	25	16
6	9	54	36	81
7	3	21	49	9
8	6	48	64	36
9	0	0	81	0
10	2	20	100	4
$\Sigma X = 55$		$\Sigma Y = 62$	$\Sigma XY = 259$	$\Sigma X^2 = 385$
				$\Sigma Y^2 = 518$

Substituindo, na fórmula, os somatórios pelos valores calculados na Tabela 6.6 e lembrando que n é o tamanho da amostra (no exemplo $n = 10$), obtemos:

$$r = \frac{259 - \frac{55 \times 62}{10}}{\sqrt{\left[385 - \frac{55^2}{10}\right] \left[518 - \frac{62^2}{10}\right]}}$$

$$r = \frac{-82}{\sqrt{82,5 \times 133,6}}$$

$$r = -0,781$$

O coeficiente de correlação varia entre -1 e $+1$, inclusive, isto é, $-1 \leq r \leq +1$. Veja então como se interpreta o valor do coeficiente de correlação:

- $r = 1$: *correlação perfeita positiva*
- $r = -1$: *correlação perfeita negativa*.
- $r = 0$: *correlação nula*
- $0 < r < 1$: *correlação positiva*
- $-1 < r < 0$: *correlação negativa*

Nas ciências físicas são encontrados valores grandes para os coeficientes de correlação, mas nas ciências da saúde os coeficientes de correlação são bem menores, devido à grande variabilidade dos fenômenos biológicos. Nas ciências do comportamento, são raros coeficientes de correlação iguais ou maiores do que 0,70. Em nenhuma ciência, porém, você encontra coeficientes de correlação iguais a $+1$ ou iguais a -1 .

Mas que valor deve ter o coeficiente de correlação para que a relação entre as variáveis seja julgada, por exemplo, forte? Para ter significado estatístico, o valor do coeficiente de correlação (r) deve ser julgado considerando o tamanho da amostra (n), por meio de um teste estatístico². Uma regra prática para julgar o valor de r , embora rudimentar³, é a seguinte:

- $0 < r < 0,25$ ou $-0,25 < r < 0$: correlação pequena ou nula.
- $0,25 < r < 0,50$ ou $-0,50 < r < -0,25$: correlação fraca.
- $0,50 < r < 0,75$ ou $-0,75 < r < -0,50$: correlação moderada.
- $0,75 < r < 1,00$ ou $-1 < r < -0,75$: correlação forte ou perfeita (perfeita se $r = -1$ ou $r = 1$).

² Veja o teste t no Capítulo 13.

³ A regra é imprecisa, mas serve como primeira aproximação. Ainda, valores de r entre $-0,30$ e $+0,30$, embora possam ter significância estatística, não são perceptíveis nos diagramas. In: COLTON, T. **Statistics in Medicine**. New York: Little, Brown and Company, 1974. p 209-11.

Exemplo 6.5: Altura e peso de pessoas.

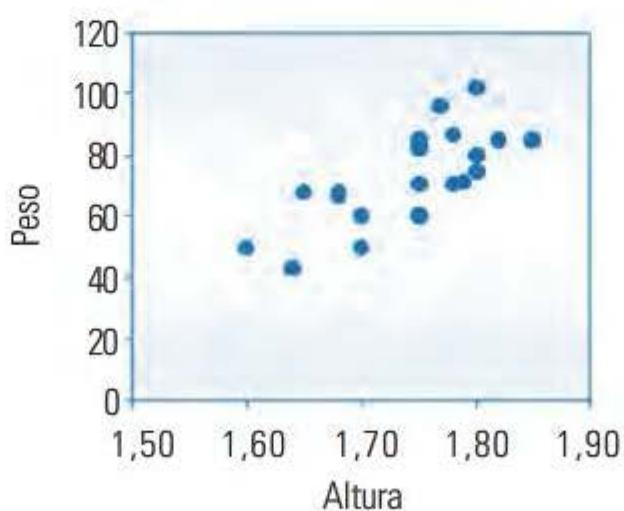
Um fisioterapeuta mediu altura (X), em metros, e peso (Y), em quilogramas, de 22 homens. Como se estuda a correlação entre essas variáveis?

TABELA 6.7

Altura, em metros, e peso, em quilogramas, de 22 homens.

Número	Altura	Peso	Número	Altura	Peso
1	1,70	60	12	1,80	75
2	1,68	68	13	1,79	71
3	1,75	85	14	1,75	70
4	1,68	67	15	1,78	87
5	1,65	68	16	1,77	96
6	1,80	102	17	1,80	80
7	1,75	60	18	1,85	85
8	1,70	60	19	1,78	70
9	1,60	50	20	1,80	80
10	1,82	85	21	1,75	82
11	1,64	43	22	1,70	50

Com um diagrama de dispersão, você "vê" a relação entre as variáveis. Parece razoável considerar que a relação é linear e positiva.

**FIGURA 6.5** Altura, em metros, e peso, em quilogramas, de 22 homens.

O valor do coeficiente de correlação, que mede o grau de correlação entre as variáveis (e você pode calcular), é $r = 0,747$, que pode ser considerada uma correlação positiva forte. Portanto, o peso de um homem está altamente correlacionado com a sua altura.

6.3 – PRESSUPOSIÇÕES

Para calcular o coeficiente de correlação, é preciso que algumas pressuposições estejam satisfeitas.

1. As unidades medidas foram selecionadas *ao acaso* — ou, pelo menos — são representativas de uma grande população.
2. Cada unidade deve fornecer tanto valores de X como de Y .
3. As variáveis X e Y devem ser *medidas independentemente*. Se os valores de Y foram obtidos por uma fórmula que inclui X , o coeficiente de correlação nunca será zero. Por exemplo, se você calcular o coeficiente de correlação entre as notas de aprovação em um curso com as notas obtidas na primeira prova, e a nota de aprovação incluir a nota obtida na primeira prova, o coeficiente de correlação não será zero.

6.4 – CUIDADOS NA INTERPRETAÇÃO DO COEFICIENTE DE CORRELAÇÃO

O diagrama de dispersão dá idéia da relação entre duas variáveis. O coeficiente de correlação de Pearson mede apenas a *relação linear* entre duas variáveis numéricas. Mas para que o valor de r , estudado aqui, tenha significado, é preciso que, no diagrama de dispersão, os pontos estejam espalhados *em torno de uma linha reta*. Portanto, antes de calcular o valor de r , convém desenhar um diagrama de dispersão: se a relação *não* for linear, o valor de r não mede a relação entre as variáveis.

Outro ponto importante é saber que *correlação não implica causa*. Uma correlação positiva entre duas variáveis mostra que essas variáveis crescem no mesmo sentido, mas não indica que aumentos sucessivos em uma das variáveis *causam* aumentos sucessivos na outra variável. Da mesma forma, uma correlação negativa entre duas variáveis mostra apenas que elas variam em sentidos contrários, mas não indica que acréscimos em uma das variáveis *causam* decréscimos na outra variável. Mas cuidado com o chavão: correlação não significa causa! Afinal, pode existir uma relação de causa e efeito entre as variáveis.

De qualquer forma, um exemplo antigo, mas muito interessante, foi dado por um estatístico que mostrou que havia correlação positiva entre o número de recém nascidos e o número de cegonhas em uma pequena cidade da Dinamarca⁴, no decorrer dos anos 30. A correlação entre essas duas variáveis é *espúria*: não indica *relação de causa e efeito*. Existe uma *terceira variável*, o crescimento da cidade, que implicava tanto no número de recém-nascidos (quanto maior a cidade, mais crianças nascem) quanto no número de casas com chaminés, perto das quais as cegonhas faziam seus ninhos.

6.5 – EXERCÍCIOS RESOLVIDOS

6.5.1 – Calcule os coeficientes de correlação para cada um dos três conjuntos de dados apresentados no Exemplo 6.2.

Solução:

Para o conjunto A: $\Sigma X = 55$; $\Sigma Y = 60$; $\Sigma XY = 352$; $\Sigma X^2 = 385$; $\Sigma Y^2 = 434$. Portanto, $r = 0,282$.

Para o conjunto B: $\Sigma X = 55$; $\Sigma Y = 76$; $\Sigma XY = 487$; $\Sigma X^2 = 385$; $\Sigma Y^2 = 654$. Portanto, $r = 0,869$

Para o conjunto C: $\Sigma X = 55$; $\Sigma Y = 75$; $\Sigma XY = 495$; $\Sigma X^2 = 385$; $\Sigma Y^2 = 645$. Portanto, $r = 1,000$.

6.5.2 – Em um trabalho sobre acumulação de placa dental em pacientes jovens, foi obtido tanto um índice clínico para medir a quantidade de placa como o peso seco das placas, em miligramas. Os dados estão na Tabela 6.8. Construa um diagrama de dispersão. Você acha que existe correlação entre as medidas? Se existe, a correlação é linear?

⁴ O exemplo é de Gustav Fischer, que apresentou em gráfico a população da cidade de Oldenburg durante sete anos (de 1930 a 1936) e o número de cegonhas observadas em cada ano. In BOX, G. E. P., HUNTER, W. G., HUNTER, J. S. **Statistics for experimenters**. New York, Wiley, 1978.

Solução:

TABELA 6.8
Peso seco, em miligramas, das placas dentais de 10 pacientes e índice clínico.

<i>Peso seco</i>	<i>Índice clínico</i>
2,3	25
2,8	45
3,5	50
3,7	68
5,8	80
6,9	100
8,2	120
10,5	128
11,9	132
14,2	135

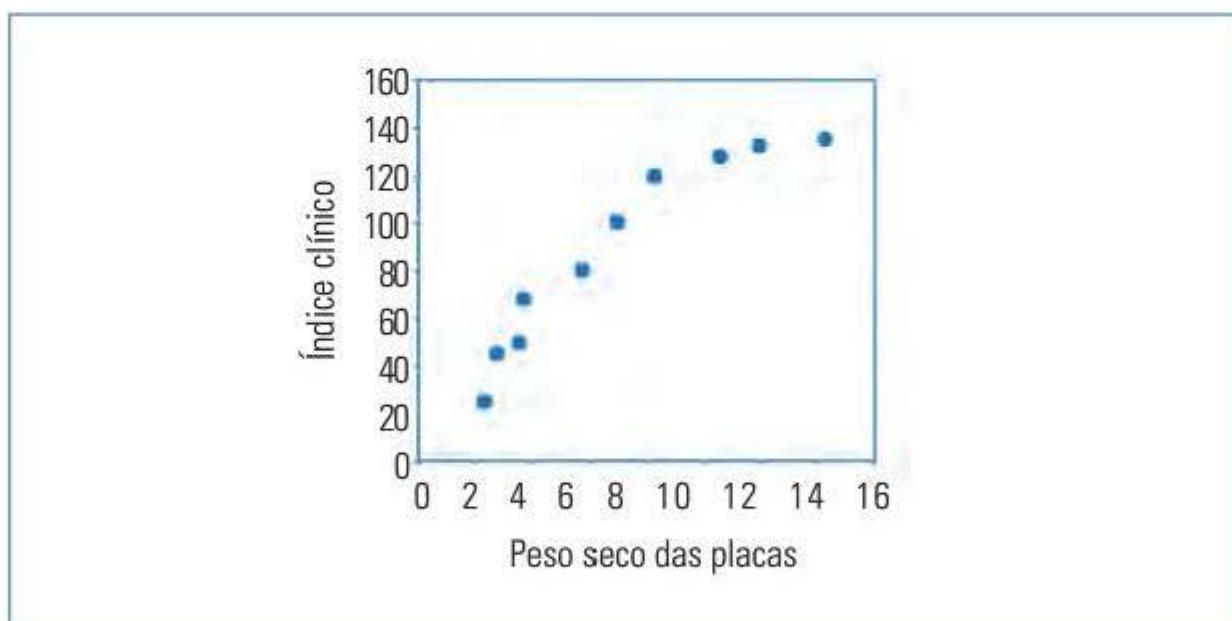


FIGURA 6.6 Índice clínico e peso seco, em miligramas, das placas dentais em 10 pacientes.

Existe correlação positiva entre as variáveis, pois ambas crescem no mesmo sentido. No entanto, essa correlação é não-linear⁵.

⁵ Existe uma explicação para o fato: o índice clínico mede apenas a extensão da área coberta pelas placas e não o volume, que determina o peso.

6.5.3 – Faça um diagrama de dispersão e calcule o coeficiente de correlação para os dados apresentados na Tabela 6.9. Discuta o resultado.

TABELA 6.9

Peso, em quilogramas, e comprimento, em centímetros, de sete recém-nascidos.

<i>Peso</i>	<i>Comprimento</i>
3,5	51
3,7	49
3,1	48
4,2	53
2,8	48
3,5	50
3,2	49

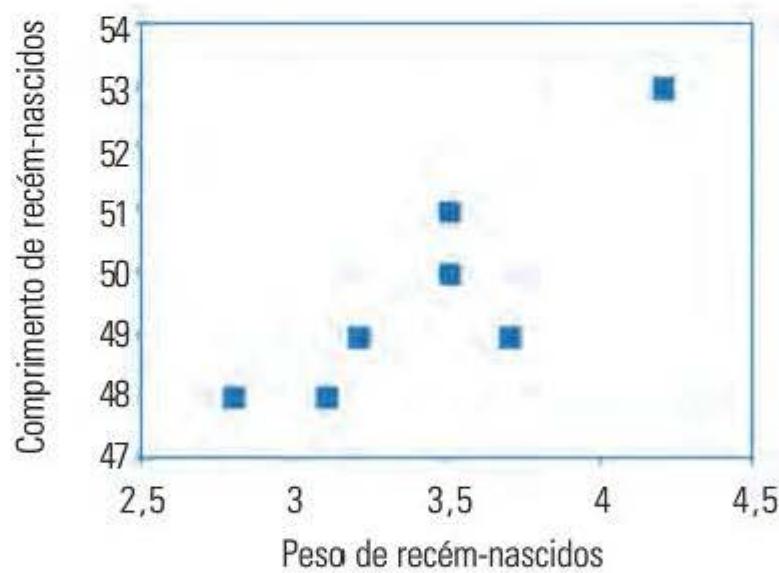


FIGURA 6.7 Peso, em quilogramas, e comprimento, em centímetros, de sete recém-nascidos.

TABELA 6.10
Cálculos intermediários para obtenção do coeficiente de correlação.

Peso (X)	Comprimento (Y)	X ²	Y ²	XY
3,5	51	12,25	2601	178,5
3,7	49	13,69	2401	181,3
3,1	48	9,61	2304	148,8
4,2	53	17,64	2809	222,6
2,8	48	7,84	2304	134,4
3,5	50	12,25	2500	175
3,2	49	10,24	2401	156,8
$\Sigma X = 24$	$\Sigma Y = 348$	$\Sigma X^2 = 83,52$	$\Sigma Y^2 = 17320$	$\Sigma XY = 1197,4$

Usando a fórmula, obtém-se $r = 0,869$, ou seja, existe correlação positiva alta entre peso e comprimento de recém-nascidos.

6.5.4 – A Tabela 6.11 fornece o peso, a estatura e o IMC (índice de massa corporal) de 10 pessoas. É razoável calcular os coeficientes de correlação das três variáveis, combinadas duas a duas? Por exemplo: altura versus peso, altura versus IMC, peso versus IMC?

TABELA 6.11
Peso, em quilogramas, estatura, em centímetros, e IMC de 10 pessoas.

Altura	Peso	IMC
1,56	53,5	21,98
1,58	58,4	23,39
1,61	59,2	22,84
1,62	53,2	20,27
1,65	64	23,51
1,72	57,5	19,44
1,73	67	22,39
1,74	66	21,80
1,79	77	24,03
1,8	66	20,37

Solução:

O IMC é dado pela fórmula:

$$IMC = \frac{\text{Peso}}{\text{Altura} \times \text{Altura}}$$

e indica a condição da pessoa, como segue:

<i>IMC</i>	<i>Condição</i>
Abaixo de 18,5	Abaixo do peso
De 18,5 a 24,9	Peso normal
De 25 a 29,9	Sobrepeso
De 30 a 34,9	Obesidade grau I
De 35 a 39,9	Obesidade grau II
40 e mais	Obesidade grau III

É perfeitamente cabível calcular a correlação entre peso e altura, mas nunca de qualquer dessas variáveis contra IMC, uma vez que esta variável é calculada a partir das outras duas. Calcular a correlação entre peso e IMC, ou entre altura e IMC, por exemplo, entraria em conflito com a pressuposição de independência.

6.6 – EXERCÍCIOS PROPOSTOS

6.6.1 – Explique o que cada um dos seguintes coeficientes de correlação informa sobre a relação entre X e Y: a) $r = 1$; b) $r = -1$; c) $r = 0$; d) $r = 0,90$; e) $r = -0,90$.

6.6.2 – Sem ver os dados, que tipo de correlação você espera entre: a) idade de pessoas adultas e velocidade de corrida; b) número de vendedores na loja e volume de vendas feitas por dia; c) a estatura de um homem e o número de dentes presentes na boca.

6.6.3 – Um estudo mostrou que a taxa de morte por doenças do coração era maior entre motoristas de ônibus do que entre cobradores. A princípio se pensou que o tipo de trabalho fosse a maior causa da doença, mas depois se notou que o tamanho dos uniformes que se fornecia aos motoristas era sempre bem maior que o dos cobradores. O que isto sugere a você?

6.6.4 – Os valores de X e Y devem ser medidos na mesma unidade para que se possa calcular o coeficiente de correlação?

6.6.5 – Indique a afirmativa que mais bem descreve o diagrama (a), o diagrama (b) e o diagrama (c), apresentados na Figura 6.8.

1. Forte correlação positiva
2. Forte correlação negativa.
3. Correlação nula ou próxima de nula
4. Correlação positiva fraca
5. Correlação negativa fraca
6. Correlação perfeita positiva
7. Correlação perfeita negativa.

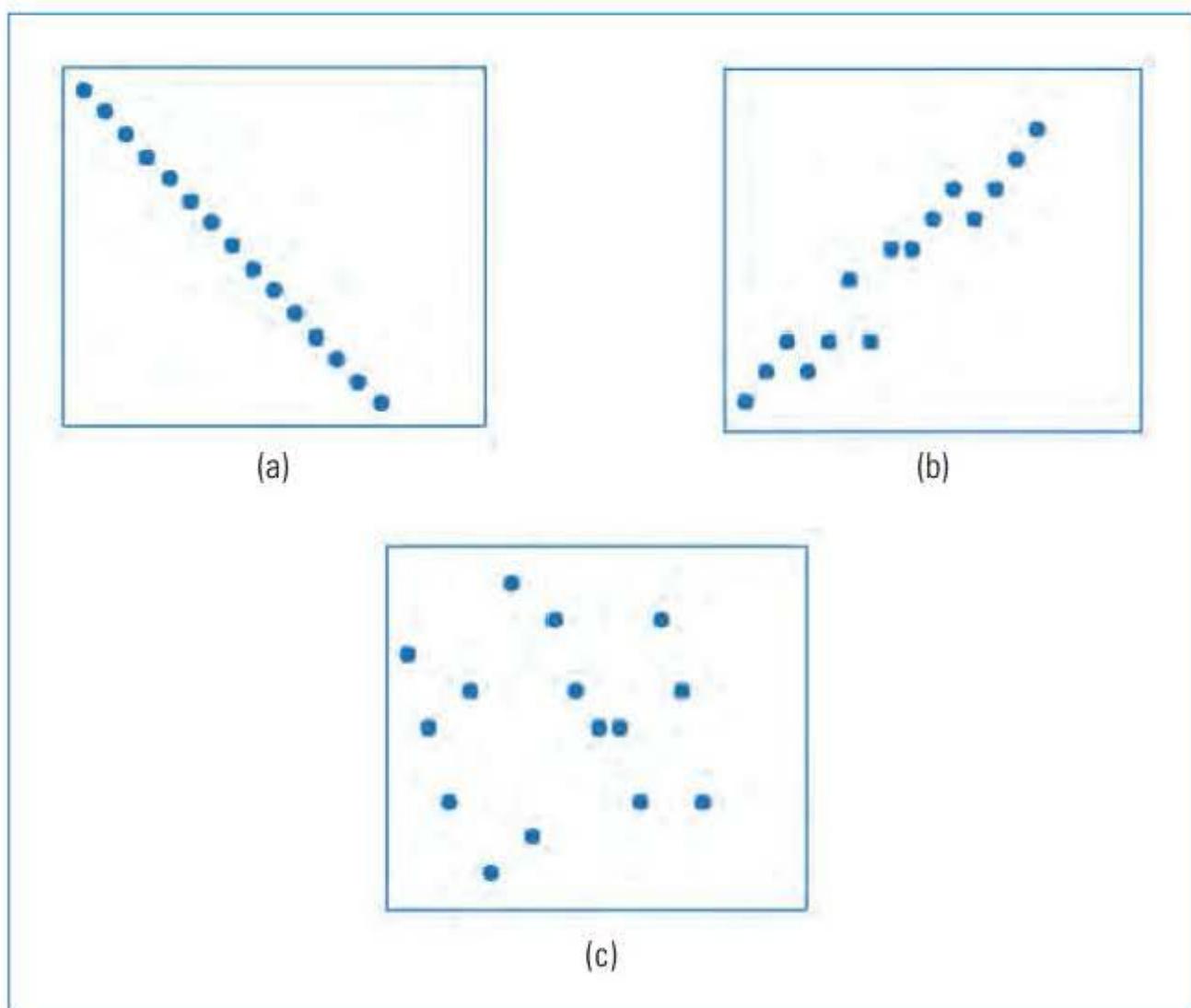


FIGURA 6.8 Diagramas de dispersão.

6.6.6 – Preencha os vazios:

O maior valor possível para o coeficiente de correlação é _____. Se todos os pontos caírem exatamente sobre uma reta, o valor de r será _____ ou _____, dependendo de a correlação ser _____ ou _____. Se todos os pontos estiverem espalhados ao acaso no diagrama de dispersão, o coeficiente de correlação terá valor próximo de _____. Quanto mais próximos de uma reta estiverem todos os pontos, _____ será o valor absoluto de r .

6.6.7 – A correlação entre idade e expectativa de vida é:

- a) positiva
- b) nula
- c) negativa
- d) irregular

6.6.8 – O diagrama de dispersão dever ser feito para estabelecer:

- a) se as variáveis estão ou não correlacionadas
- b) se as variáveis são positivas
- c) se as variáveis são negativas
- d) a qualidade das variáveis.

6.6.9 – Faça um diagrama de dispersão e calcule o coeficiente de correlação para os dados apresentados na Tabela 6.12. Discuta o resultado.

TABELA 6.12
Dados relativos a duas variáveis X e Y .

X	Y
3	2
5	2
4	7
2	7
1	2

6.6.10 – Faça diagramas de dispersão e calcule os valores de r para os conjuntos de dados da Tabela 6.13.

TABELA 6.13
Dois conjuntos de pares de valores de duas variáveis.

<i>Conjunto A</i>		<i>Conjunto B</i>	
<i>X</i>	<i>Y</i>	<i>X</i>	<i>Y</i>
1	1	1	1
2	3	1,5	2
3	6	3	3
4	5	4,5	2
5	8	5	1

6.6.11 – Se todos os valores de Y forem iguais, qual será o valor de r ?

6.6.12 – Calcule o coeficiente de correlação para os dados apresentados na Tabela 6.14.

TABELA 6.14
Idade gestacional, em semanas, e peso ao nascer, em quilogramas, de recém-nascidos.

<i>Idade gestacional</i>	<i>Peso ao nascer</i>
28	1,25
32	1,25
35	1,75
38	2,25
39	3,25
41	3,25
42	4,25

6.6.13 – Calcule os coeficientes de correlação de Pearson para os dados dos dois conjuntos a seguir. Discuta a razão de os valores de r serem tão diferentes, embora os dados sejam tão semelhantes.

TABELA 6.15
Dois conjuntos de pares de valores de duas variáveis.

<i>Conjunto A</i>		<i>Conjunto B</i>	
<i>X</i>	<i>Y</i>	<i>X</i>	<i>Y</i>
1	2	1	2
2	4	2	4
3	6	3	6
4	8	4	8
5	10	5	0

6.6.14 – Suponha que os seguintes dados⁶ foram obtidos de pacientes com enfisema: *X* é o número de anos que o paciente fumou e *Y* é a avaliação (uma nota) do próprio médico do paciente sobre a diminuição da capacidade pulmonar (medida numa escala de zero a 100.). Os resultados para 10 pacientes estão na Tabela 6.16. Calcule o valor do coeficiente de correlação.

Saiba que: $\Sigma XY = 18.055$; $\Sigma X^2 = 11.053$; $\Sigma Y^2 = 30.600$.

TABELA 6.16
Tempo do hábito de fumar (*X*), em anos, e diminuição da capacidade pulmonar (*Y*), avaliada pelo médico do paciente.

<i>Número do paciente</i>	<i>X</i>	<i>Y</i>
1	25	55
2	36	60
3	22	50
4	15	30
5	48	75
6	39	70
7	42	70
8	31	55
9	28	30
10	33	35

⁶ OTT, L e MENDENHALL, W. *Understanding Statistics*. Belmont, Wadsworth, 6 ed. 1994. p. 487.

6.6.15 – O volume máximo de oxigênio inalado ($VO_2\text{máx}$) tem sido usado como medida da situação cardíaca tanto de indivíduos saudáveis como de pessoas que sofrem de doenças cardíacas. Os dados⁷ de $VO_2\text{máx}$ em mililitros por quilograma por minuto para 12 homens saudáveis depois de exercícios estão na Tabela 6.17. Desenhe um diagrama de dispersão. Olhando o diagrama, você diria que $VO_2\text{máx}$ diminui quando aumenta a atividade?

TABELA 6.17

Duração do exercício, em minutos, e $VO_2\text{máx}$, em mililitros por quilograma por minuto, para 12 homens saudáveis.

<i>Duração do exercício</i>	<i>VO₂máx</i>
10	82
9,5	73
10,2	68
10,5	74
11	66
11,3	63
11,6	58
12	54
12,1	56
12,5	51
12,8	55
13	44

⁷ OTT, L e MENDENHALL, W. *Understanding Statistics*. Belmont, Wadsworth, 6 ed. 1994. p. 503.

(página deixada intencionalmente em branco)

Noções sobre Regressão

7

(página deixada intencionalmente em branco)

O Capítulo 6 mostrou como se estuda a *relação entre duas variáveis*. Muitas vezes, porém, interessa estudar *como* uma variável varia em função da outra. Por exemplo, todos nós sabemos que as crianças crescem — as variáveis idade e altura têm correlação positiva — mas é preciso saber também *como* a altura de uma criança varia em função da idade. Todos nós sabemos que a população do Brasil aumentou nas últimas décadas. Mas *como* e *quanto*? Para dar uma primeira resposta a estas questões, é importante desenhar um gráfico de linhas.

7.1 – GRÁFICO DE LINHAS

Para aprender como se faz um gráfico de linhas, vamos pensar em duas variáveis numéricas e — como fizemos no Capítulo 6 — chamar uma delas de *X* e a outra de *Y*. Então cada unidade da amostra fornece dois valores, um para cada variável.

Quando se estuda a variação da variável *Y* em função da variável *X*, diz-se que *Y* é a *variável dependente* e que *X* é a *variável explanatória*. Por exemplo, altura de criança varia em função da idade. Então *altura* é a *variável dependente* e *idade* é a *variável explanatória*.

Quem trabalha na área de saúde costuma observar *como* uma variável evolui ao longo do tempo. Com os dados observados de *Y* ao longo do tempo *X*, é possível fazer um *gráfico de linhas*. Para fazer esse gráfico:

- Colete valores da variável *Y* nos tempos que você quer estudar.
- Trace um sistema de eixos cartesianos; represente o tempo (*X*) no eixo das abscissas e a variável *Y* no eixo das ordenadas.
- Estabeleça as escalas e faça, em cada eixo, as necessárias graduações.
- Escreva os nomes das variáveis nos respectivos eixos.
- Desenhe um ponto para representar cada par de valores (*X*, *Y*).
- Una os pontos por segmentos de reta.
- Escreva o título.

Exemplo 7.1: Gráfico de linhas.

Na Tabela 7.1 são dados pares de valores das variáveis *X* e *Y*. A variável *X* é o ano do Censo Demográfico do Brasil e a variável *Y* é a população residente. Veja a Figura 7.1: o gráfico de linhas mostra o crescimento no período de forma a complementar os dados da Tabela 7.1.

TABELA 7.1
População residente no Brasil, segundo o ano do censo demográfico.

<i>Ano do censo</i>	<i>População</i>
1940 ¹	41.236.315
1950 ¹	51.944.397
1960 ¹	70.191.370
1970	93.139.037
1980	119.002.706
1991	146.815.796
2000	169.799.170

Fonte: IBGE (2003)¹

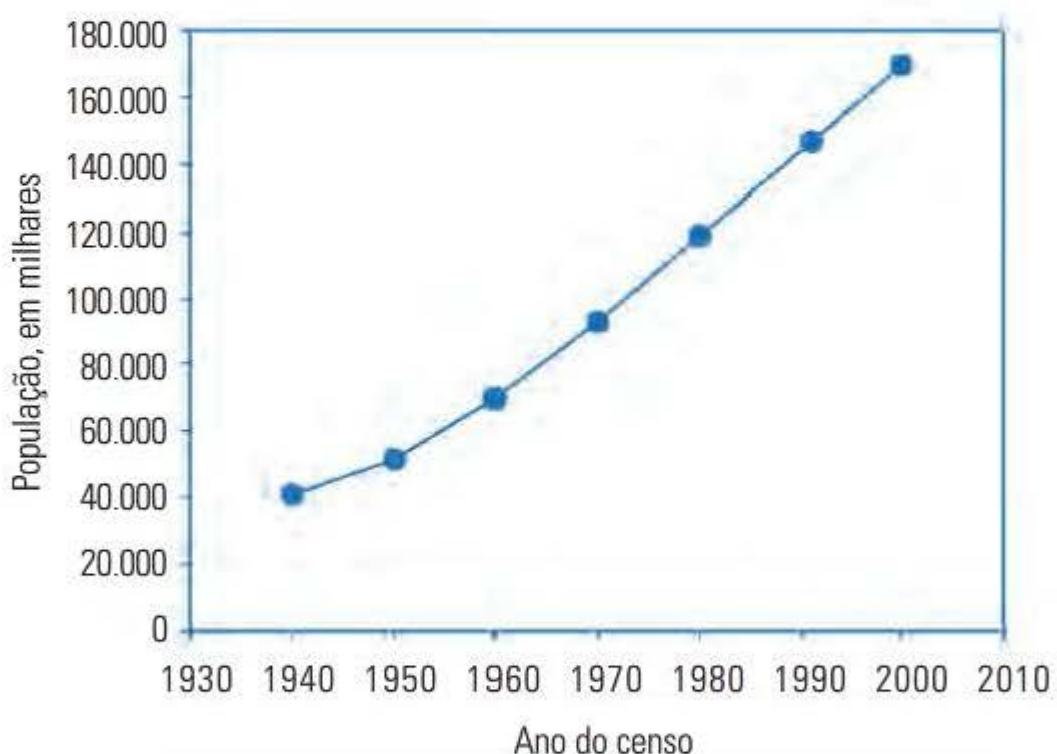


FIGURA 7.1 População residente no Brasil, segundo o ano do censo demográfico.

¹IBGE. Censo 2000: um retrato do Brasil na década de 90. Disponível em:< <http://www.ibge.gov.br>>. Acesso em: abr. 2003.

7.2 – RETA DE REGRESSÃO

A variação de Y em função de X deve ser observada no gráfico de linhas. Se os pontos ficam dispersos em torno de uma reta, é razoável traçar uma reta no meio desses pontos. A *melhor* reta (melhor, no sentido que tem propriedades estatísticas desejáveis) recebe o nome de *reta de regressão*². São dadas, nesta seção, as fórmulas para obter essa reta.

Exemplo 7.2: A idéia de regressão.

Observe os dados apresentados na Tabela 7.2. Foi colocada a mesma quantidade de plasma humano em oito tubos de ensaio e depois se ajuntou, em cada tubo, uma quantidade fixa de procaína (anestésico local). Mediu-se então, em tempos diferentes, a quantidade de procaína que já havia se hidrolisado. O diagrama de dispersão apresentado na Figura 7.2 mostra que a quantidade de procaína hidrolisada varia em função do tempo decorrido após sua administração.

TABELA 7.2
Quantidade de procaína hidrolisada, em 10 moles/litro, no plasma humano, em função do tempo, em minutos, decorrido após sua administração.

<i>Tempo</i>	<i>Quantidade hidrolisada</i>
2	3,5
3	5,7
5	9,9
8	16,3
10	19,3
12	25,7
14	28,2
15	32,6

²Muitos autores referem-se à reta de regressão como reta de mínimos quadrados porque esse é o método estatístico utilizado para chegar às fórmulas dadas nesta Seção.

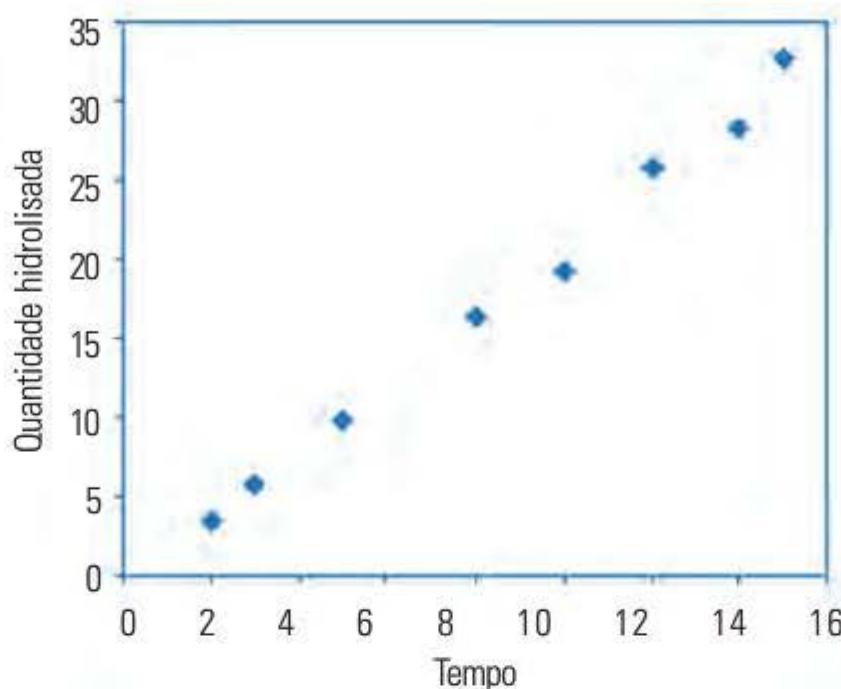


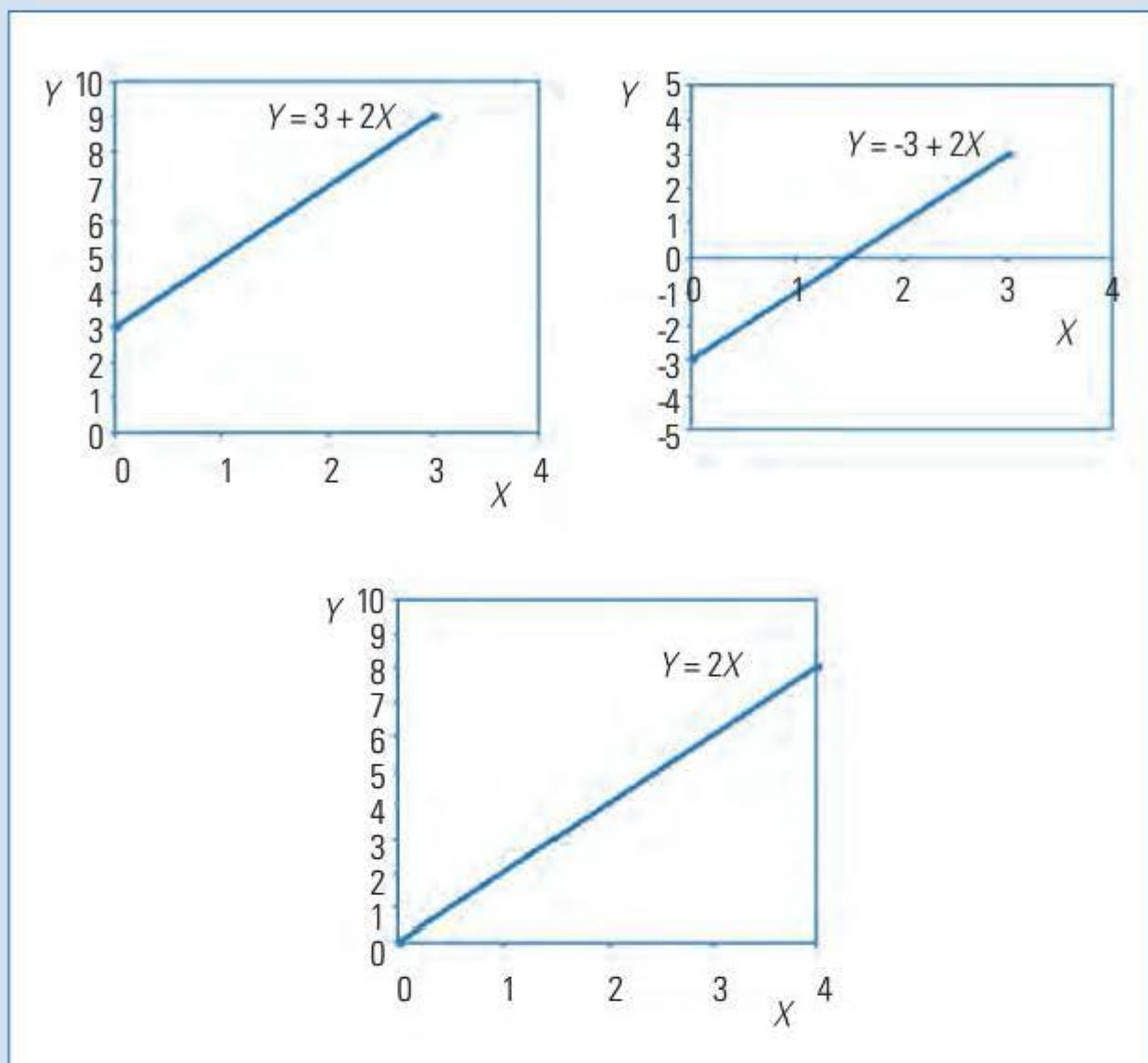
FIGURA 7.2 Quantidade de procaína hidrolisada, em 10 moles/litro, no plasma humano, em função do tempo, em minutos, decorrido após sua administração.

Vamos discutir um pouco mais o Exemplo 7.2. Parece razoável concluir, observando a Figura 7.2, que a variação da quantidade de procaína hidrolisada no plasma humano em função do tempo decorrido após sua administração pode ser descrita por meio de *uma reta de regressão*.

Para ajustar *uma reta de regressão* (isto é, estabelecer a equação da reta) aos dados apresentados na Tabela 7.2, é preciso obter o coeficiente linear e o coeficiente angular da reta, também chamados *coeficientes de regressão*. Convém lembrar o que são esses coeficientes.

No sistema de eixos cartesianos, a equação $Y = a + bX$ é uma reta. O *coeficiente linear* da reta, indicado neste livro por a , dá a *altura* em que a reta corta o eixo das ordenadas. Se a for um número:

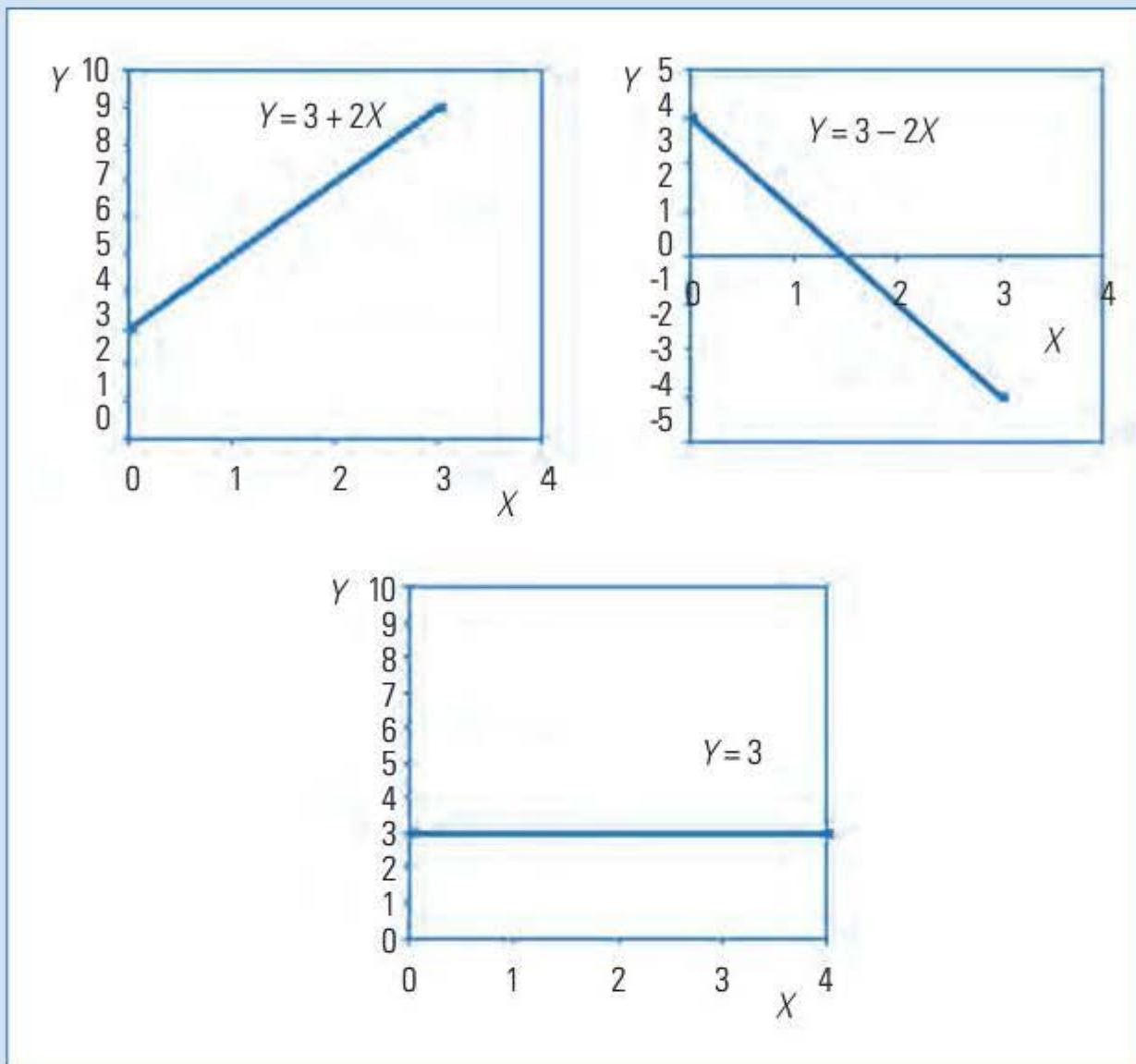
- *positivo*, a reta corta o eixo das ordenadas *acima* da origem;
- *negativo*, a reta corta o eixo das ordenadas *abaixo* da origem.
- *zero*, a reta passa na origem do sistema de eixos cartesianos.

Exemplo 7.3: Equação da reta: coeficientes lineares diferentes.**FIGURA 7.3** Apresentação gráfica de retas com diferentes coeficientes lineares.

O *coeficiente angular da reta*, indicado neste livro por b , dá a inclinação da reta³. Se b for um número:

- *positivo*, a reta é ascendente;
- *negativo*, a reta é descendente;
- *zero*, a reta é paralela aos eixos das abscissas.

³ O coeficiente angular, chamado neste livro de b , é a tangente trigonométrica do ângulo formado pelo eixo das abscissas e pela reta de equação $Y = a + bX$.

Exemplo 7.4: Equação da reta: coeficientes angulares diferentes.**FIGURA 7.4** Apresentação gráfica de retas com diferentes coeficientes angulares.

Em Estatística, o coeficiente angular da reta é obtido por meio da fórmula:

$$b = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sum X - \frac{(\sum X)^2}{n}}$$

e o coeficiente linear é obtido por meio da fórmula:

$$a = \bar{Y} - b \bar{X}$$

em que \bar{Y} e \bar{X} são as médias de Y e X , respectivamente. Veja o Exemplo 7.5.

Exemplo 7.5: Cálculo dos coeficientes de regressão.

Calcule a reta de regressão para o problema apresentado no Exemplo 7.2.

TABELA 7.3
Cálculos intermediários para a obtenção de a e de b .

X	Y	XY	X ²
2	3,5	7	4
3	5,7	17,1	9
5	9,9	49,5	25
8	16,3	130,4	64
10	19,3	193	100
12	25,7	308,4	144
14	28,2	394,8	196
15	32,6	489	225
69	141,2	1.589,2	767

Aplicando as fórmulas, obtém-se:

$$b = \frac{1589,2 - \frac{69 \times 141,2}{8}}{767 - \frac{69^2}{8}} = \frac{371,35}{171,875} = 2,16$$

$$a = \frac{141,2}{8} - 2,16 \times \frac{69}{8} = -0,98$$

Para traçar a *reta de regressão* é preciso dar valores arbitrários para X e depois calcular os valores de Y . Indicam-se os valores calculados de Y por \hat{Y} .

Fazendo $X = 5$, tem-se que:

$$\hat{Y} = -0,98 + 2,16 \times 5 = 9,82$$

e fazendo $X = 15$, tem-se que:

$$\hat{Y} = -0,98 + 2,16 \times 15 = 31,42.$$

Os dois pares de valores ($X = 5$ e $\hat{Y} = 9,82$) e ($X = 15$ e $\hat{Y} = 31,42$) permitem traçar a reta de regressão.

Exemplo 7.6: Traçado da reta de regressão.

Apresente, no diagrama de dispersão da Figura 7.2, a reta de *equação*

$$\hat{Y} = -0,98 + 2,16 X.$$

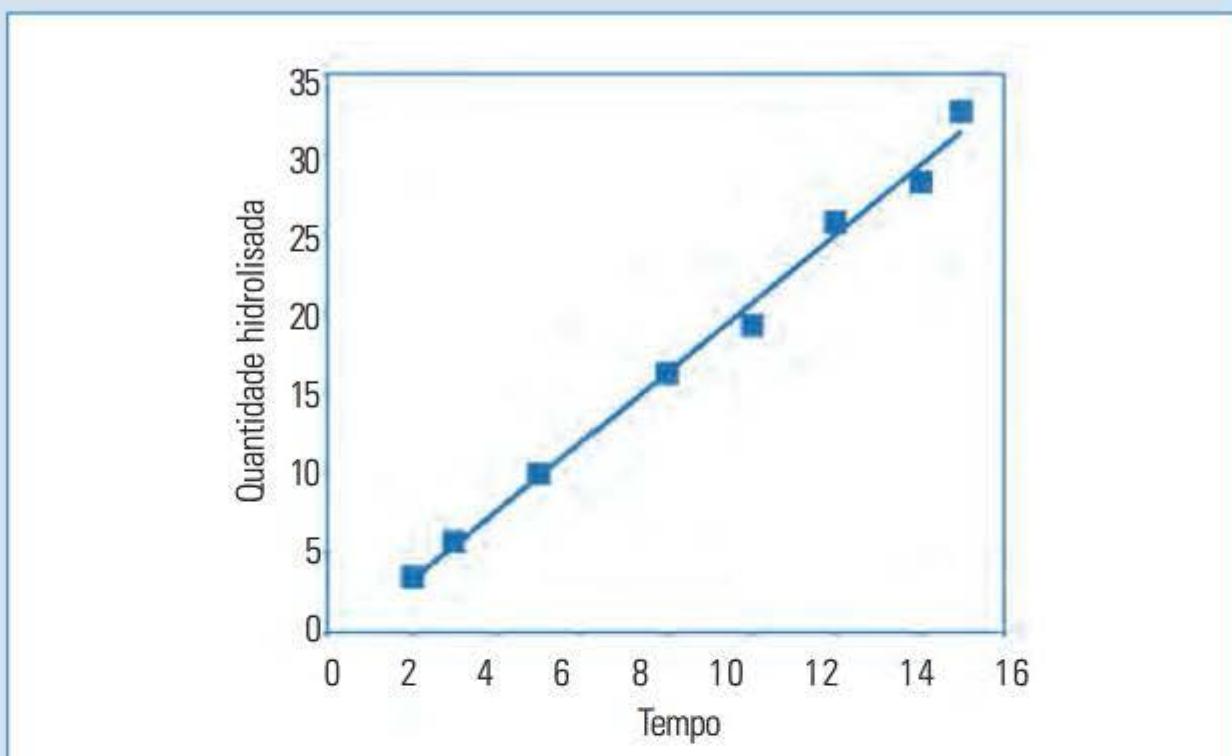


FIGURA 7.5 Reta de regressão: quantidade de procaína hidrolisada, em 10 moles/litro, no plasma humano, em função do tempo, em minutos, decorrido após sua administração.

A equação da reta de regressão permite *estimar* valores de Y para quaisquer valores de X dentro do intervalo estudado, mesmo que tais valores não existam na amostra. Observe os dados apresentados na Tabela 7.2. Não existe o valor $X = 13$, mas é possível estimar o valor de Y para $X = 13$. Basta fazer:

$$\hat{Y} = -0,98 + 2,16 \times 13 = 27,10$$

O valor $\hat{Y} = 27,10$ é uma *previsão*, feita com base na equação da reta de regressão, para a quantidade de procaína que deve estar hidrolisada 13 minutos após sua administração.

Dada a reta de regressão, fica fácil calcular o valor de Y para qualquer valor de X . No entanto, o bom senso deve fazer com que você *não* estime valores de Y para valores de X muito além do intervalo estudado: a *extrapolação* pode levar ao absurdo, porque a relação entre X e Y , linear no intervalo estudado, pode não ser linear fora desse intervalo.

É verdade que as pessoas tendem a prever, com base no que se observou em determinado período, o que acontecerá em outro período, próximo ou longínquo. A *extrapolação* é, geralmente, incorreta ou até desastrosa. Por exemplo, por volta dos 6 anos começam a irromper dentes permanentes.

tes em crianças, mas isso só acontece até certa idade. Ninguém espera, pelo fato de terem irrompido quatro dentes numa criança entre os 7 e os 8 anos, que isso ocorra entre 30 e 31 anos de idade.

Exemplo 7.7: A extração indevida.

A Tabela 7.4 apresenta as temperaturas médias mensais, nos primeiros sete meses do ano, de uma cidade do sul do Brasil. Esses dados estão no diagrama de dispersão da Figura 7.6. Se alguém ajustar uma reta como a mostrada no diagrama e quiser usar essa reta para “prever” a temperatura na cidade em dezembro (mês 12), chegará a um valor absurdo, menor do que 2 graus negativos. A razão disso é óbvia: o fenômeno não é linear além do período estudado.

TABELA 7.4
Temperaturas médias segundo o mês, de uma cidade do sul do Brasil.

Mês	Número do mês	Temperatura média no mês
Janeiro	1	23
Fevereiro	2	22
Março	3	20
Abril	4	18
Maio	5	15
Junho	6	12
Julho	7	9

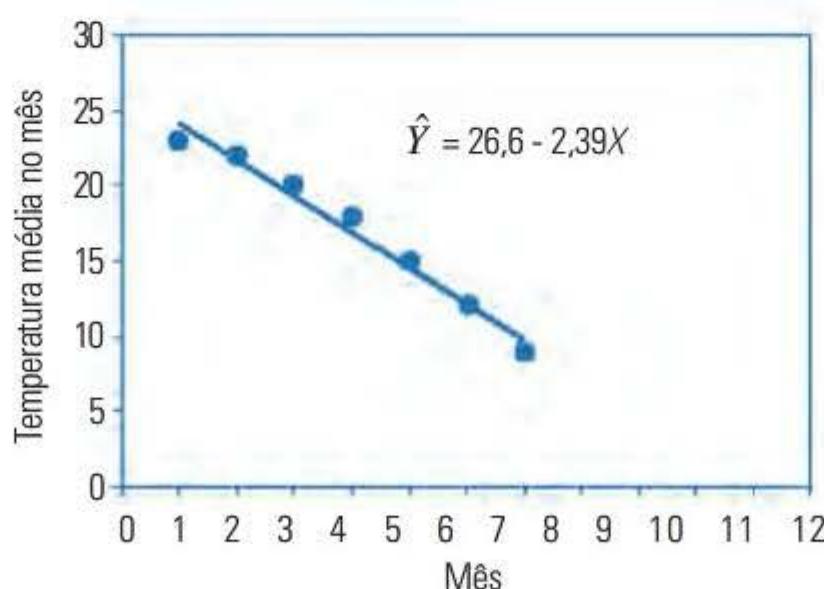


FIGURA 7.6 Reta ajustada às temperaturas médias de uma cidade do sul do Brasil, segundo o mês.

7.3 – ESCOLHA DA VARIÁVEL EXPLANATÓRIA

Quando os valores de X são fixados antes do início da coleta dos dados, ajusta-se a regressão de Y contra X . No Exemplo 7.2, o pesquisador fixou os tempos em que iria observar a quantidade de procaína que estaria hidrolisada no plasma, antes de iniciar a pesquisa. Então, a quantidade de procaína hidrolisada *depende* do tempo em que foi medida — não o contrário.

Nem sempre os valores de X são fixados *antes* do início dos trabalhos. Nesses casos, tanto se pode ajustar a regressão de Y contra X , como a regressão de X contra Y , mas recomenda-se identificar a variável que *deve ser prevista*, conhecido o valor da outra variável e ajustar a regressão de Y contra X toda vez que se pretende estudar a variação de Y (prever Y) em função da variação de X .

Exemplo 7.8: A escolha da variável explanatória.

Calcule a reta de regressão para os dados apresentados na Tabela 7.5.

É razoável estudar a variação da pressão arterial (Y) em função do peso (X), porque é o peso que pode explicar (explanar) a pressão arterial — e não o contrário. Então se deve ajustar uma regressão da pressão arterial (Y) contra o peso (X).

TABELA 7.5
Pressão arterial (PA), em milímetros de mercúrio e peso de homens adultos, em quilogramas.

Peso	PA	Peso	PA	Peso	PA
14	105	18	113	21	127
14	102	19	107	22	125
15	111	19	125	22	116
15	104	19	130	23	130
15	107	19	110	23	107
16	90	19	107	23	103
16	105	20	102	24	135
16	102	20	116	24	143
16	126	21	135	28	121
17	134	21	100	28	135

Foram calculados:

$$b = \frac{\frac{271159 - 3624 \times 2238}{30}}{167386 - \frac{2238^2}{30}} = 1,88$$

$$a = \frac{3624}{30} - 1,88 \times \frac{2238}{30} = -19,1$$

A reta de regressão

$$\hat{Y} = -19,1 + 1,88X$$

apresentada na Figura 7.7 mostra a *tendência* de ocorrer aumento de pressão arterial quando aumenta o peso, mas convém observar que os pontos estão *muito dispersos* em torno da reta. Isso significa que a *previsão* da pressão arterial de um homem adulto em função de seu peso tem grande margem de erro.

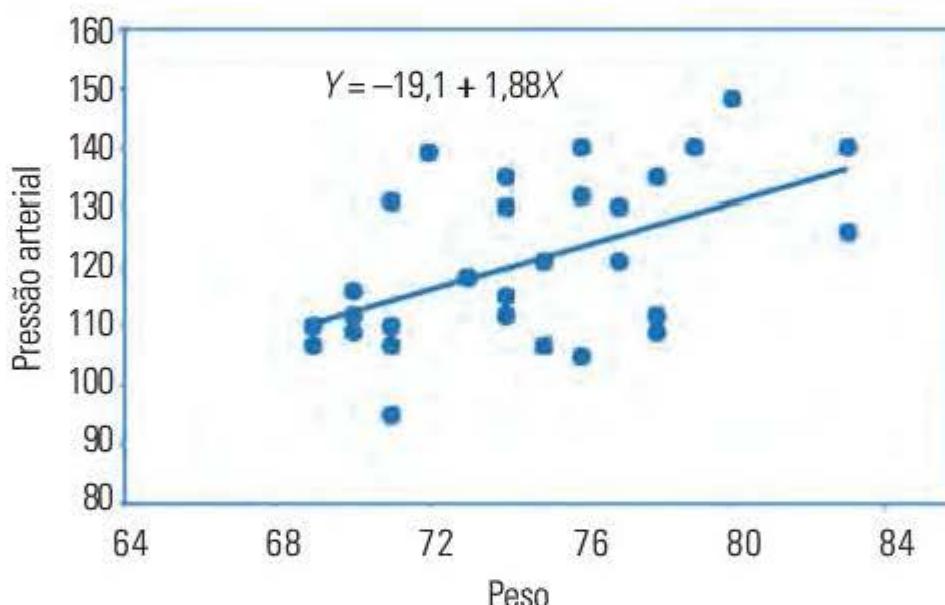


FIGURA 7.7 Reta de regressão para pressão arterial em função do peso.

7.4 – COEFICIENTE DE DETERMINAÇÃO

Antes de aprender o que é coeficiente de determinação, vamos entender o que é uma relação matemática e o que é uma relação estatística. Se você aumentar o lado de um quadrado em 1 cm, a área aumenta. E se você continuar aumentando o lado do quadrado de 1 cm em 1 cm, a área continuará aumentando. Você sabe dizer *exatamente* a área do quadrado para cada tamanho de lado porque a relação entre a área de um quadrado e seus lados é matemática: área = lado × lado.

Pense agora em uma pessoa que quer diminuir o peso porque — seu médico lhe disse — os gordos têm tendência a ter pressão arterial alta. Sabe-se, portanto, que o aumento da pressão arterial é função do aumento de peso. Será que existe uma *relação exata* entre essas duas variáveis, isto é, para cada quilo a mais haverá um aumento fixo na pressão arterial? *Não* é assim. Existe tendência de a pressão arterial aumentar com o aumento de peso, mas a pressão arterial também aumenta em função de outros fatores como idade, vida sedentária, hereditariedade e certos hábitos, como o hábito de fumar e o consumo excessivo de sal. E mesmo que conhecêssemos muitas das causas que explicam o aumento da pressão arterial, ainda assim não saberíamos prever *exatamente* a pressão arterial de uma pessoa. A relação entre pressão arterial e peso é probabilística e, portanto, sujeita a erro.

Com estes exemplos queremos lembrar a você que existem *relações determinísticas* — como é a relação entre lado e área de um quadrado — e *relações probabilísticas* — como é a relação entre peso e pressão arterial. No primeiro caso, não existe erro na previsão, isto é, dado o lado de um quadrado você pode dizer *exatamente* qual é a área: está determinado. No segundo caso, a previsão é possível, mas dentro de certas margens de erro. Neste ponto, a pergunta é inevitável: qual é o “tamanho” desse erro?

Existe uma estatística chamada *coeficiente de determinação*, indicada por R^2 , que mede a *contribuição* de uma variável na *previsão* de outra. Parece complicado, mas tente entender este exemplo: imagine que você quer comprar uma camiseta para uma criança. Você chega na loja e pede ajuda à vendedora. O que primeiro ela pergunta? A idade da criança, claro. Por quê? Porque o tamanho de uma criança é função da idade. Boa parte da variação do tamanho das crianças é explicada pela variação de suas idades — o que é medido pelo R^2 . Portanto, saber a idade da criança ajuda na *previsão* do tamanho da sua camiseta⁴.

O coeficiente de determinação é a proporção da variação de Y explicada pela variação de X.

O *coeficiente de determinação* é dado pelo quadrado do coeficiente de correlação. *Não* pode, portanto, ser negativo. Varia entre zero e 1, inclusive. Para interpretar o coeficiente de determinação, é melhor transformá-lo em porcentagem, multiplicando o resultado obtido em seu cálculo por 100. Veja o Exemplo 7.9.

⁴A vendedora também pergunta se o presente é para menino ou menina. Essa informação também contribui, embora menos do que idade, para a escolha do tamanho (na primeira infância os meninos são maiores), mas ajuda na escolha do modelo.

Exemplo 7.9. Coeficiente de determinação.

Calcule o coeficiente de determinação para os dados apresentados na Tabela 7.2 e na Tabela 7.5 e discuta cada um deles.

Usando os cálculos intermediários já apresentados na Tabela 7.3, é possível obter $R^2 = 0,994$. Isto significa que 99,4% da variação da quantidade de procaína hidrolisada no plasma se explica pelo tempo decorrido após sua administração. Em outras palavras, se você souber o tempo que decorreu depois que a procaína foi colocada no plasma, poderá justificar 99,4% da variação de procaína que hidrolisou.

Para os dados da Tabela 7.5, com a ajuda de um computador (ou de seu professor) é possível obter, $R^2 = 0,282$, um valor baixo. Se fosse alto, a explicação seria de que, dado o peso de um homem, a pressão arterial seria altamente previsível. No entanto, fatores como idade, vida sedentária, hereditariedade e certos hábitos, como o hábito de fumar e consumo abusivo de sal devem ser, também, importantes.

7.5 – UMA PRESSUPOSIÇÃO BÁSICA

Para ajustar uma regressão linear simples de X contra Y , é preciso que os dados de X e Y tenham sido *obtidos independentemente*. Então, quando você for interpretar os resultados do ajuste de uma regressão, verifique como foram obtidos os dados de X e Y . Veja o Exemplo 7.7: a regressão obtida é uma *falácia* porque não se pode fazer uma regressão da diferença das variáveis contra o valor inicial.

Exemplo 7.10: Uma falácia.

Observe os dados da Tabela 7.6, que estão no diagrama de dispersão da Figura 7.8: *os pontos não sugerem correlação entre as variáveis*. O coeficiente de determinação é $R^2 = 0,030$. No entanto, se você fizer a diferença $Y-X$ e colocar a diferença como função do valor inicial (X), obterá o diagrama de dispersão da Figura 7.9, com $R^2 = 0,582$. Só que isso *não* pode ser feito: a regressão obtida é uma falácia.

TABELA 7.6
Notas de 10 alunos em duas provas.

1 ^a prova	2 ^a prova	Diferença = 2 ^a prova - 1 ^a prova
7	7	0
5	5	0
4	8	4
9	9	0
2	10	8
4	3	-1
8	4	-4
10	6	-4
6	4	-2
7	3	-4

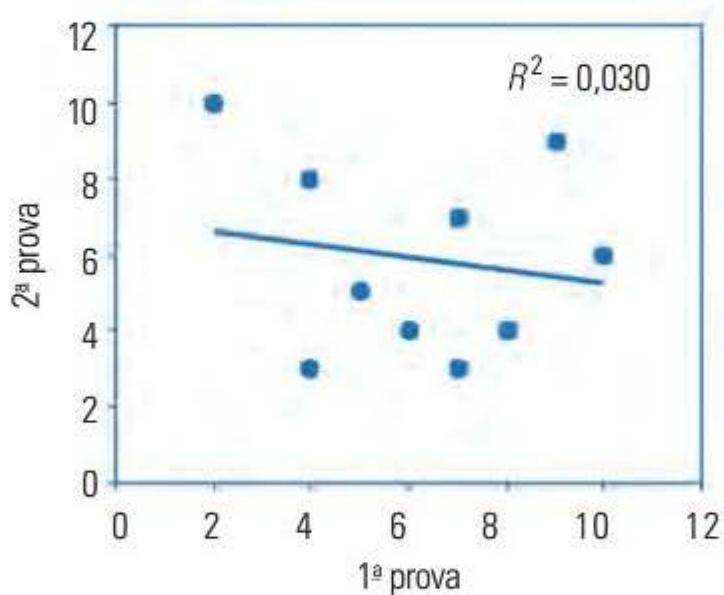


FIGURA 7.8 Nota na segunda prova em função da nota na primeira prova.

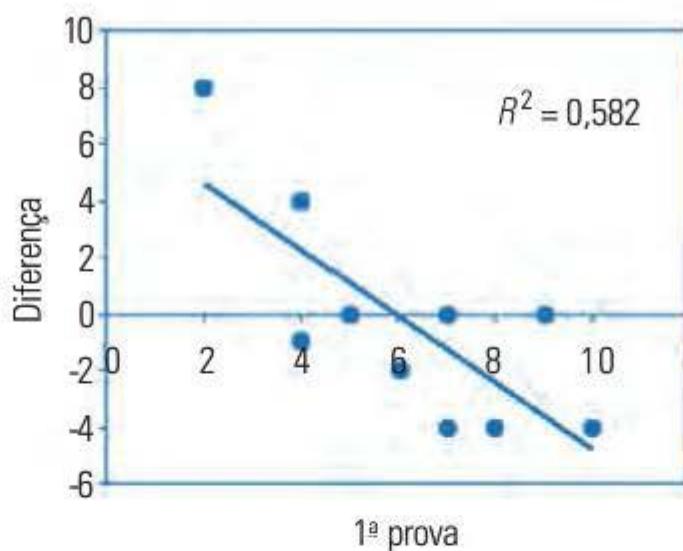


FIGURA 7.9 Diferença das notas de 10 alunos em duas provas em função da 1^a nota.

7.6 – OUTROS TIPOS DE REGRESSÃO

Existem situações em que os pares de valores das variáveis X e Y , apresentados em diagrama de dispersão, não se distribuem em torno de uma reta⁵. Veja o Exemplo 7.11.

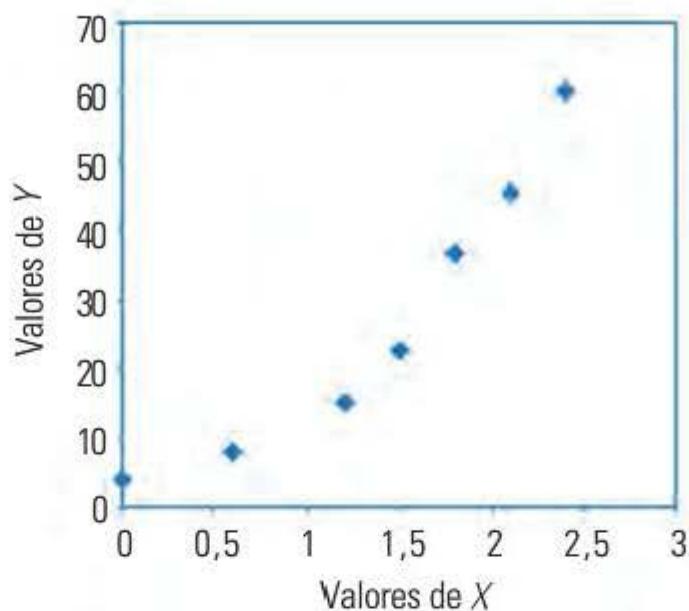
Exemplo 7.11: Uma regressão não-linear.

Observe os dados da Tabela 7.7, apresentados em diagrama de dispersão na Figura 7.10: os pontos estão dispersos em torno de uma curva.

TABELA 7.7
Valores de duas variáveis X e Y .

X	Y
0	4,0
0,6	8,0
1,2	15,0
1,5	22,6
1,8	36,4
2,1	45,3
2,4	60,0

⁵No programa EXCEL, você encontra as seguintes opções para ajuste de regressão: linear (que vimos até aqui), logarítmica, polinomial (que não será visto neste livro) potência, exponencial, média móvel (que não será visto neste livro).

**FIGURA 7.10** Diagrama de dispersão para os valores X e Y apresentados na Tabela 7.7.

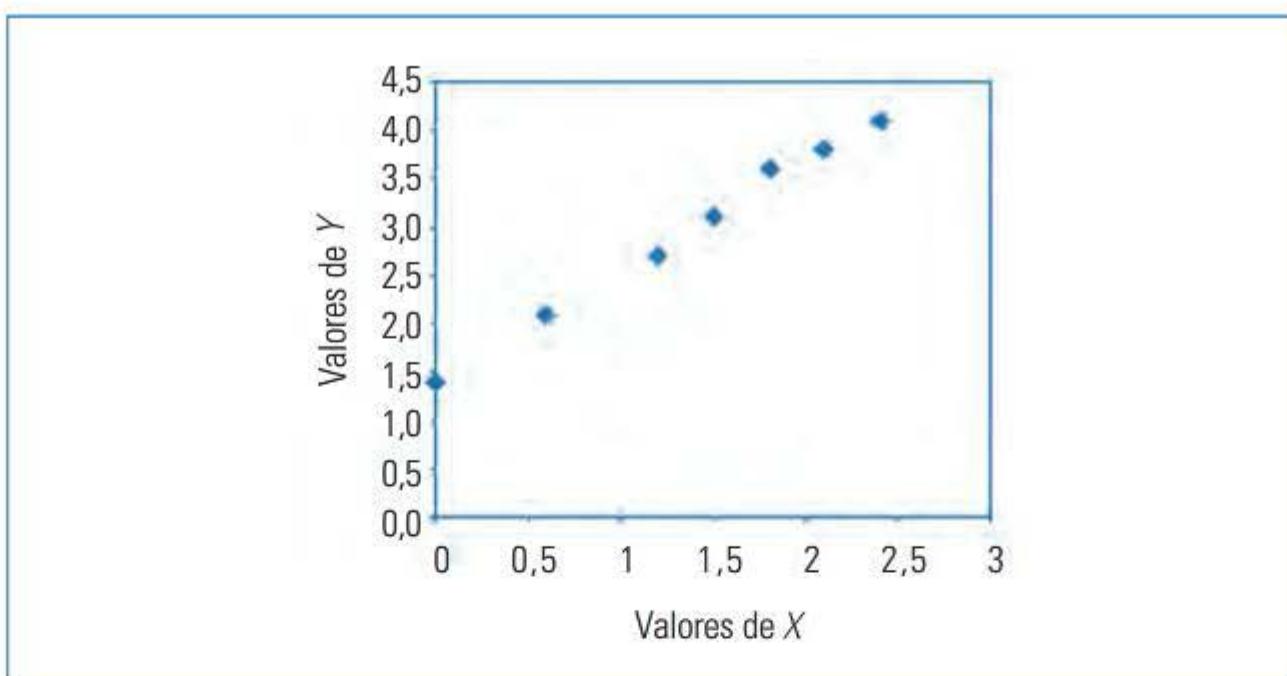
Quando os pontos apresentados em diagrama de dispersão *não* estão em torno de uma reta, devemos experimentar *transformar* a variável Y . Por exemplo, podemos experimentar fazer um diagrama de dispersão colocando, em lugar de valores de Y , os valores do logaritmo neperiano⁶ de Y .

Para os dados apresentados no Exemplo 7.11, os valores de X e dos logaritmos neperianos de Y estão apresentados na Tabela 7.8 e na Figura 7.11.

TABELA 7.8
Valores de X e valores dos logaritmos neperianos de Y .

X	$\ln Y$
0	1,3863
0,6	2,0794
1,2	2,7081
1,5	3,1179
1,8	3,5946
2,1	3,8133
2,4	4,0943

⁶No Excel, procure a opção *exponencial*.

**FIGURA 7.11** Diagrama de dispersão.

O diagrama de dispersão apresentado na Figura 7.11 mostra pontos praticamente sobre uma reta. Então é possível ajustar uma regressão linear de $\ln Y$ contra X . Para calcular a e b , são necessários os cálculos intermediários apresentados na Tabela 7.9.

TABELA 7.9
Cálculos intermediários para a obtenção de a e b .

X	$\ln Y$	$X \ln Y$	X^2
0	1,3863	0,0000	0
0,6	2,0794	1,2477	0,36
1,2	2,7081	3,2497	1,44
1,5	3,1179	4,6769	2,25
1,8	3,5946	6,4702	3,24
2,1	3,8133	8,0079	4,41
2,4	4,0943	9,8264	5,76
9,6	20,7940	33,4788	17,46

Com base nos cálculos apresentados na Tabela 7.9, é possível obter:

$$b = \frac{33,4788 - \frac{9,6 \times 20,7940}{7}}{17,46 - \frac{9,6^2}{7}} = 1,1554$$

$$a = \frac{20,7940}{7} - 1,1554 \times \frac{9,6}{7} = 1,3861$$

A equação de reta de regressão de $\ln \hat{Y}$ contra X é:

$$\ln \hat{Y} = 1,3861 + 1,1554X$$

Se você quiser voltar ao valor da variável Y , é preciso calcular o antilogaritmo da equação. Então, você obtém:

$$\hat{Y} = \text{antiln}(1,3861) e^{1,1554X}$$

ou:

$$\hat{Y} = 3,999 e^{1,1554X}$$

Esta equação é chamada de *exponencial* porque traz a variável explanatória no expoente.

Para que uma regressão linear possa ser ajustada aos dados, muitas vezes basta transformar uma das variáveis⁷. Outras vezes, é preciso transformar ambas as variáveis⁸. Também podem ser utilizadas outras transformações, além da *transformação logarítmica*, mostrada aqui. Assim, são também usadas a *extração de raiz quadrada* e a *inversão*, além de outras, mais complicadas.

As transformações são, em geral, *empíricas*, isto é, dados n pares de valores X e Y , é preciso fazer várias tentativas até achar a transformação que permita ajustar uma regressão linear aos pares de dados. Algumas vezes, porém, o modelo é *especificado teoricamente*. Por exemplo, a equação de Arrenhius dá a velocidade de uma reação química em função da temperatura em que a reação se processa. Se T é a temperatura em graus Kelvin em que ocorre a reação química, a equação de Arrenhius estabelece que a velocidade V é dada por:

$$\ln V = C - \frac{A}{R} \times \frac{1}{T}$$

em que $\ln V$ é o logaritmo neperiano da velocidade da reação química à temperatura T e R é uma constante (1,987 cal/grau/mol). Para ajustar a equação de Arrenhius aos dados de temperatura e de velocidade de uma reação química, é preciso calcular os valores das variáveis transformadas, isto é, o *logaritmo neperiano da velocidade* e o *inverso da temperatura*. Depois se ajusta uma regressão linear do logaritmo neperiano de V contra o inverso de T , isto é:

$$\ln V = a + b \frac{1}{T}$$

Então, $C = a$ e $A = -Rb$.

⁷Para ajustar uma regressão *logarítmica*, transforme X , isto é, ajuste a regressão dos logaritmos de X contra Y . Para ajustar uma regressão *potência*, transforme X e Y , isto é, ajuste a regressão dos logaritmos de X contra os logaritmos de Y .

⁸Veja mais sobre o assunto em VIEIRA, S. *Bioestatística: tópicos avançados*. 2 ed. Rio de Janeiro: Campus, 2004.

Uma regra, porém, é básica: antes de ajustar uma reta de regressão aos dados, devem-se colocar os pontos ($X; Y$) em um diagrama de dispersão e estudar o conhecimento disponível na literatura sobre o fenômeno. A inspeção dos dados numéricos é obrigatória. Às vezes, é possível ajustar mais de um modelo aos dados e depois escolher, com base nas estatísticas obtidas (coeficientes de determinação etc.), o modelo que melhor se ajusta aos dados.

Neste Capítulo vimos como se ajusta uma *regressão linear simples* aos dados: *linear*, porque é uma reta, e *simples*, porque está no plano, isto é, existe uma só variável dependente e uma só variável explanatória. Mas a variação da variável dependente pode serposta em função de diversas variáveis, isto é, podem existir diversas variáveis explanatórias. É o caso, por exemplo, da pressão arterial que depende não apenas de peso como mostrado no exemplo, mas da idade, de fatores hereditários, da alimentação etc. Nesses casos, ajusta-se aos dados uma *regressão multipla*, isto é, uma função com diversas variáveis explanatórias. Mas este tema não será tratado aqui.

7.7 – EXERCÍCIOS RESOLVIDOS

7.7.1 – Faça um gráfico de linhas para os dados apresentados na Tabela 7.10. Discuta.

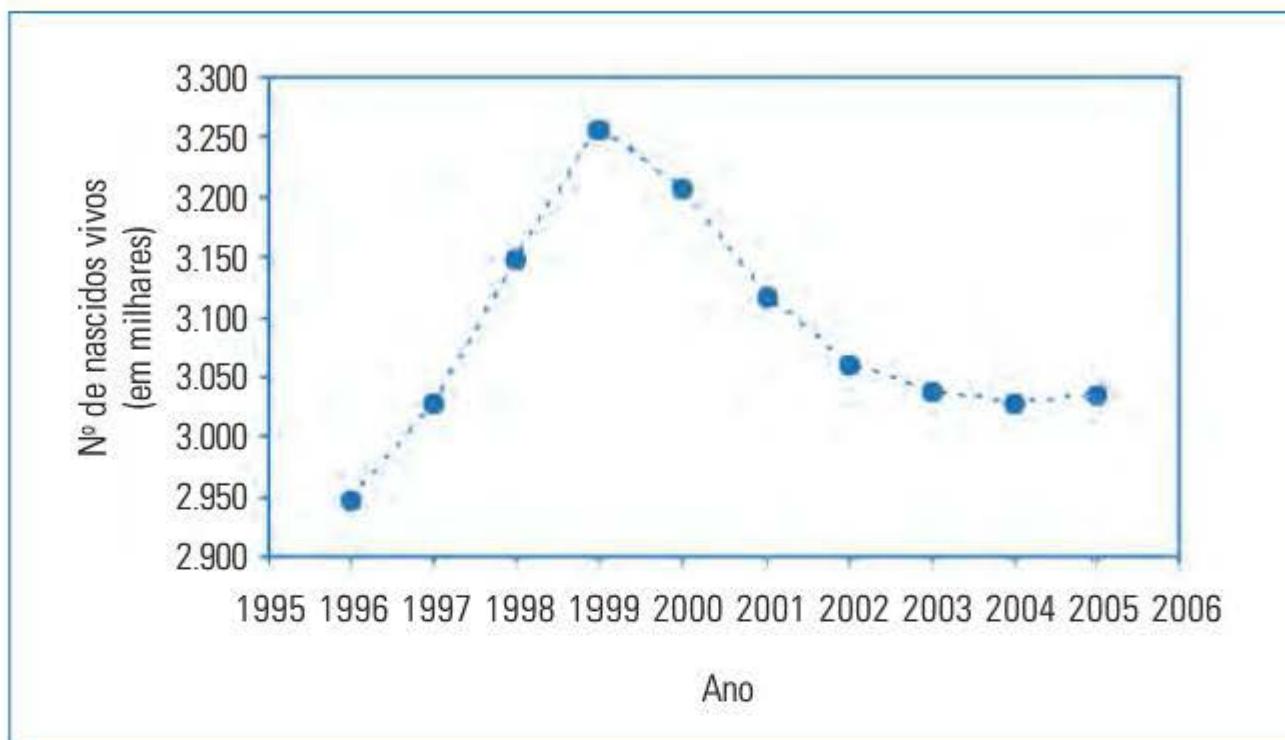
TABELA 7.10
Número de nascidos vivos no Brasil, no período de 1996 a 2005.

Ano	Número de nascidos vivos
1996	2.945.425
1997	3.026.658
1998	3.148.037
1999	3.256.433
2000	3.206.761
2001	3.115.474
2002	3.059.402
2003	3.038.251
2004	3.026.548
2005	3.035.096

Fonte: DATASUS (2008)⁹

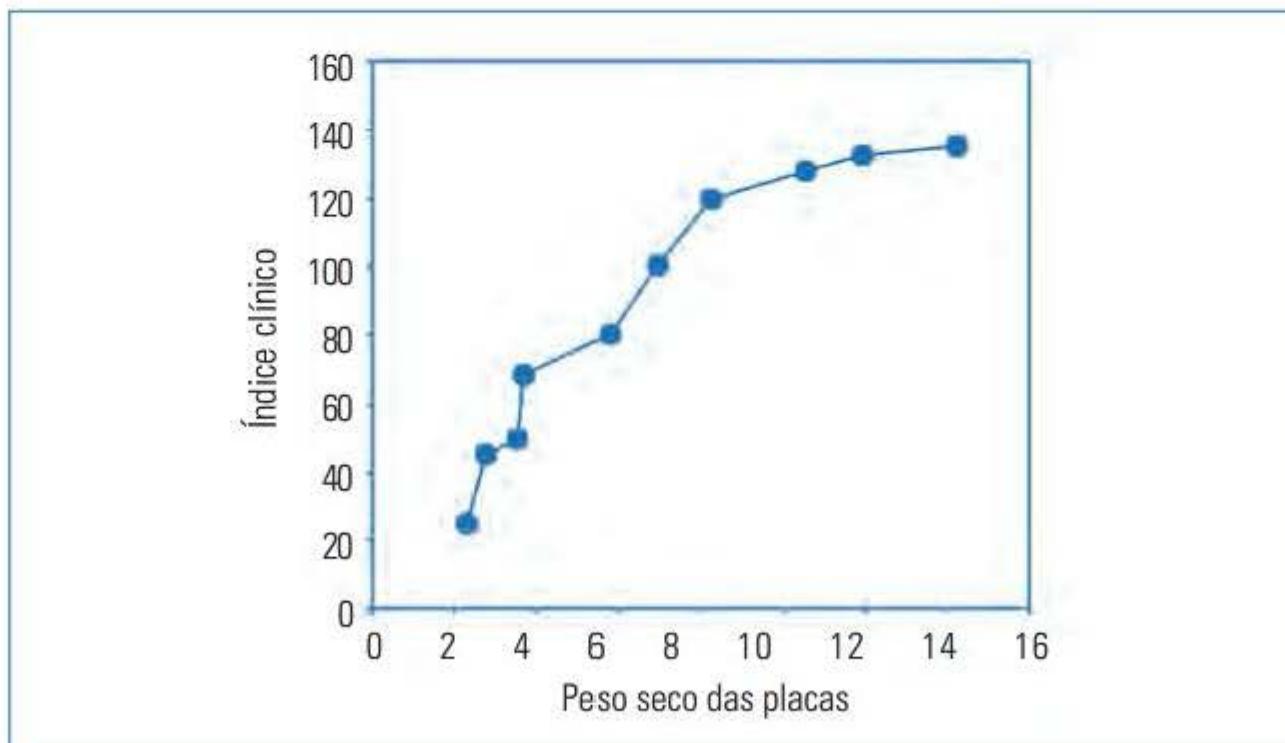
⁹Disponível em <http://tabnet.datasus.gov.br/cgi/deftohtm.exe?idb2006/a02.def> em 10 de abril de 2008.

Solução

**FIGURA 7.12** Número de nascidos vivos no Brasil, no período de 1996 a 2005.

O número de nascidos vivos no Brasil aumentou até 1999. De lá para 2006, observa-se decréscimo.

7.7.2 – Faça um gráfico de linhas para os dados apresentados no Exercício 6.5.2 do Capítulo 6, para mostrar como o índice clínico varia em função do peso seco das placas. Discuta.

**FIGURA 7.13** Índice clínico em função do peso seco das placas bacterianas.

A Figura 7.13 mostra que o índice clínico usado para medir a quantidade de placa aumenta linearmente (e aceleradamente) com o peso seco das placas, em miligramas, até cerca de 8 mg. Depois, tende a estabilizar. Isto talvez se explique pelo fato de o índice clínico medir a área dos dentes com placas bacterianas, mas não o volume. Ora, o peso leva em conta o volume, que aumenta quando o acúmulo de placas é grande.

7.7.3 – Ajuste uma reta de regressão aos dados apresentados no Exercício 6.5.3 do Capítulo 6, para estudar peso em função do comprimento dos recém-nascidos. Calcule o coeficiente de determinação.

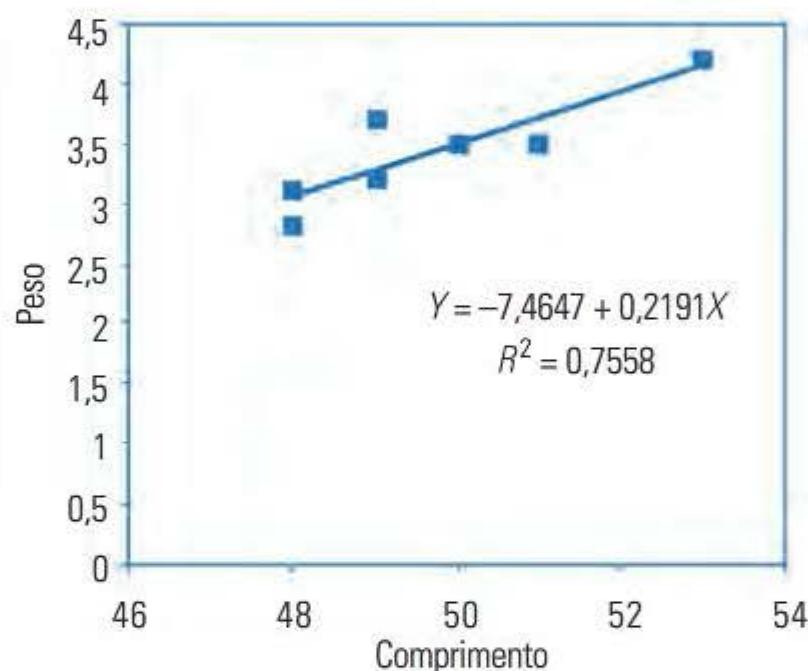


FIGURA 7.14 Reta de regressão para peso de recém-nascidos em função do comprimento.

7.7.4 – Ajuste uma reta de regressão aos dados apresentados no Exercício 6.5.4 do Capítulo 6, para estudar peso em função da altura. Calcule o coeficiente de determinação.

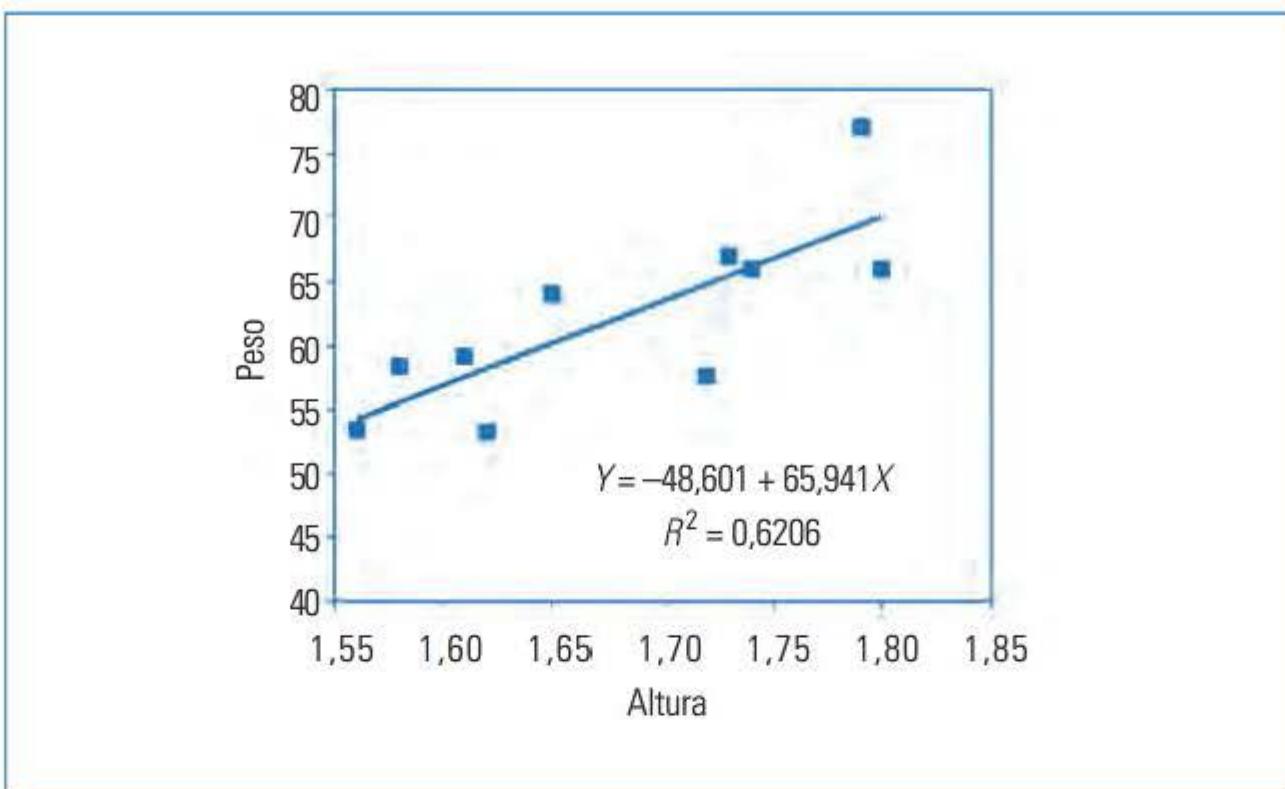


FIGURA 7.15 Reta de regressão para peso em função da altura.

7.8 – EXERCÍCIOS PROPOSTOS

7.8.1 – Faça um gráfico de linhas para os dados apresentados na Tabela 7.11. Discuta.

TABELA 7.11
Razão de sexos¹⁰ no Brasil, em 2005.

<i>Faixa etária</i>	<i>Razão de sexos</i>
Menos de 1 ano	104,36
De 1 a 4 anos	103,59
De 5 a 9 anos	103,49
De 10 a 14 anos	103,16
De 15 a 19 anos	102,29
De 20 a 24 anos	100,05
De 25 a 29 anos	97,57
De 30 a 34 anos	95,13
De 35 a 39 anos	94,41
De 40 a 44 anos	92,84
De 45 a 49 anos	92,61
De 50 a 54 anos	93,63
De 55 a 59 anos	90,40
De 60 a 64 anos	87,09
De 65 a 69 anos	81,49
De 70 a 74 anos	80,08
De 75 a 79 anos	77,81
80 e mais anos	64,49

Fonte: DATASUS¹¹ (2008)

¹⁰Razão de sexos: número de homens por 100 mulheres.

¹¹Disponível em <http://tabnet.datasus.gov.br/cgi/deftohtm.exe?idb2006/a02.def> em 10 de abril de 2008.

7.8.2 – Faça um gráfico de linhas para os dados apresentados na Tabela 7.12. Discuta.

TABELA 7.12
Coeficiente de mortalidade infantil¹² no Brasil, de 1889 a 1998

Ano	Coeficiente de mortalidade infantil
1889	52,02
1890	49,40
1891	46,99
1892	44,79
1893	42,80
1894	41,01
1895	39,40
1896	37,97
1897	36,70
1898	36,10

Fonte DATASUS (2008)¹³

7.8.3 – Ajuste uma reta de regressão aos dados apresentados na Tabela 7.13.

TABELA 7.13
Teor de vitamina C (mg de ácido ascórbico/100 ml de suco de maçã) em função do período de armazenamento em dias.

Período de armazenamento	Teor de vitamina C
1	4,09
45	3,27
90	2,45
135	3,27
180	1,64

¹²Taxa ou coeficiente de mortalidade infantil é a razão entre o total de óbitos de menores de 1 ano de idade (excluídos os nascidos mortos) e o total de nascidos vivos, em determinado período de tempo (normalmente 1 ano). Essa razão é multiplicada por 1.000. A taxa de mortalidade infantil estima o risco que um nascido vivo tem de morrer antes de completar 1 ano de idade. A Organização Mundial de Saúde considera *altas* as taxas de 50 por 1.000 ou mais, *médias* as que ficam entre 20 e 49 e *baixas* as menores do que 20.

¹³Disponível em <http://tabnet.datasus.gov.br/cgi/mortinf/mibr.htm#topo> em 10 de abril de 2008.

7.8.4 – A reta de regressão será a mesma se você trocar X por Y? O coeficiente de correlação muda?

7.8.5 – É preciso que X e Y tenham as mesmas unidades para poder se calcular a reta de regressão?

7.8.6 – Se os filhos fossem exatamente 5 cm mais altos do que seus pais, como ficaria a reta de regressão que daria a altura dos filhos em função da altura de seus pais?

7.8.7 – Como seria a reta de regressão se todos os pontos de X tivessem o mesmo valor?

7.8.8 – Os dados da Tabela 7.14 foram apresentados com a finalidade de mostrar que existe relação entre CPO-D médio (a média de um índice de cáries, ou seja, a média da soma do número de dentes afetados pela cárie em uma amostra de crianças: C = cariados; P = perdidos por cárie; O = obturados, ou seja, restaurados devido ao ataque de cárie) e a média do número de anos de estudo do responsável pelas crianças. O que você acha?

TABELA 7.14
Número médio de anos de estudo do responsável pelas crianças de uma amostra e CPO-D médio.

<i>Anos de estudo do responsável</i>	<i>CPO-D médio</i>
0	1,70
1 - 4	1,85
5 - 8	0,75
9 - 11	0,44

7.8.9 – Uma cadeia de padarias queria saber se a quantidade de dinheiro gasto em propaganda faz aumentar as vendas. Durante seis semanas fez, em ordem aleatória, gastos com propaganda de valores variados conforme mostra a Tabela 7.15 e anotou os valores recebidos nas vendas. Calcule a reta de regressão e coloque em gráfico. O que você acha?

TABELA 7.15
Gastos com propaganda, em reais, na semana e valores recebidos, em reais, nas vendas.

<i>Gastos</i>	<i>Valores recebidos</i>
100,00	1.020,00
150,00	1.610,00
200,00	2.030,00
250,00	2.560,00
300,00	2.800,00

7.8.10 – Com os dados¹⁴ apresentados no Exercício 6.6.14 do Capítulo 6, obtidos de pacientes com enfisema, calcule a reta de regressão.

7.8.11 – Com os dados¹⁴ apresentados no Exercício 6.6.15 do Capítulo 6 sobre o volume máximo de oxigênio inalado ($VO_2\text{máx}$), você diria que a variável diminui linearmente quando a atividade aumenta? Calcule a reta de regressão.

7.8.12 – Os dados¹⁵ apresentados na Tabela 7.16 referem-se à pressão sanguínea diastólica, em milímetros de mercúrio, quando a pessoa está em repouso. Os valores de X indicam o tempo em minutos desde o início do repouso e os valores Y são valores de pressão sanguínea. Desenhe um diagrama de dispersão. Por que não se deve ajustar uma reta de regressão aos dados?

TABELA 7.16
Tempo em minutos desde o início do repouso e pressão sanguínea diastólica, em milímetros de mercúrio.

<i>Tempo em minutos desde o início do repouso</i>	<i>Pressão sanguínea diastólica</i>
0	72
5	66
10	70
15	64
20	66

¹⁴OTT, L e MENDENHALL, W. *Understanding Statistics*. Belmont, Wadsworth, 6 ed. 1994. p. 487.

¹⁵SCHORK, M. A. e REMINGTON, R. D. *Statistics with applications to the biological and health sciences*. New Jersey, Prentice Hall, 3 ed. 2000. p. 297.

7.8.13 – Faça um diagrama de dispersão para apresentar os dados da Tabela 7.17. Calcule a reta de regressão. Coloque a reta no gráfico. Quanto devem pesar 10 ratos com 32 dias?

TABELA 7.17

Idade, em dias, e peso médio, em gramas, de 10 ratos machos da raça Wistar.

<i>Idade</i>	<i>Peso médio</i>
30	64
34	74
38	82
42	95
46	106

7.8.14 – Ajuste uma equação exponencial aos dados da Tabela 7.18.

TABELA 7.18

Dados de X e Y.

<i>X</i>	<i>Y</i>
28	1,25
32	1,25
35	1,75
38	2,25
39	3,25
41	3,25
42	4,25

(página deixada intencionalmente em branco)

Noções sobre Probabilidade

8

(página deixada intencionalmente em branco)

Você já sabe o que é *probabilidade*: se alguém perguntar qual é a probabilidade de sair cara no jogo de moeda, você responde: 1/2 ou 50%. A questão, aqui, é saber como se chega a esse resultado. Mas você deve ter pensado: quando se joga uma moeda, tanto pode sair cara como coroa; as duas faces não podem ocorrer ao mesmo tempo; logo, cara ocorre em metade das vezes.

Portanto, quando alguém diz que a probabilidade de sair cara num jogo de moedas é 1/2 — mesmo que esteja pensando em jogar a moeda uma única vez — está fornecendo, como resposta, a proporção de caras que obtaria se jogasse a moeda um grande número de vezes. E a pessoa não sabe o que vai acontecer em uma única jogada.

Neste exemplo, ficam claras duas características dos fenômenos probabilísticos:

- Não se pode antecipar um resultado.
- Existe um padrão de comportamento previsível no longo prazo.

Todo fenômeno probabilístico tem, como resultado, um *evento* (acontecimento) e o conjunto de eventos possíveis é chamado *espaço amostral*.

Exemplo 8.1: Espaço amostral.

Dê o espaço amostral do lançamento de duas moedas.

Solução

- cara-cara;
- cara-coroa;
- coroa-cara;
- coroa-coroa.

8.1 – DEFINIÇÃO CLÁSSICA DE PROBABILIDADE

Se forem possíveis n eventos mutuamente exclusivos e igualmente prováveis, se m desses eventos tiverem a característica que chamaremos A, a *probabilidade* de que ocorra um evento com a característica A é indicada por $P(A)$ e é dada pela razão m/n .

$$P(A) = \frac{m}{n}$$

Simplificando, você deve ter aprendido que a probabilidade de obter um evento favorável (um evento com uma característica que chamamos de A) é dada por:

$$P(A) = \frac{n^{\circ} \text{ de eventos favoráveis}}{n^{\circ} \text{ de eventos possíveis}}$$

Exemplo 8.2: Cálculo de probabilidade.

Qual é a probabilidade de ocorrer face 6, quando se joga um dado?

Solução

Quando se joga um dado pode ocorrer um dos seis ($n = 6$) eventos do espaço amostral: 1, 2, 3, 4, 5 ou 6.

Só existe um evento ($m = 1$) com a característica pedida: face 6. Então a probabilidade de ocorrer 6 é:

$$P(6) = \frac{1}{6} = 0,1667$$

Na prática, é comum que as pessoas falem em porcentagens quando tratam de probabilidades. Por exemplo, a maioria das pessoas diria que a probabilidade de sair “cara” quando se lança uma moeda é 50%. Os estatísticos preferem expressar valores de probabilidade por números entre zero e 1. Mas se você quiser expressar probabilidade em porcentagem, basta multiplicar o valor dado pela definição por 100.

Veja agora duas propriedades das probabilidades:

- A soma das probabilidades de todos os eventos possíveis (dados no espaço amostral) é obrigatoriamente 1 (ou 100%).
- A probabilidade varia entre zero e 1 (ou entre 0% e 100%), inclusive¹.

Exemplo 8.3: Extremos: zero ou 1.

Evento *certo* tem probabilidade 1 (ou 100%). Por exemplo, a probabilidade de que qualquer um de nós venha morrer um dia é 1 (ou 100%).

Evento *impossível* tem probabilidade zero. Por exemplo, a probabilidade de que qualquer um de nós seja imortal é zero.

8.2 – FREQÜÊNCIA RELATIVA COMO ESTIMATIVA DE PROBABILIDADE

O estudo de probabilidades tem enorme aplicação nas ciências em geral, mas começou com os jogos de azar. As pessoas queriam entender a “lei” que rege esses jogos para ganhar dinheiro nos cassinos². E os matemáticos acabaram estabelecendo a teoria das probabilidades.

¹Não existe, por exemplo, 200% de probabilidade. Expressões deste tipo aparecem na linguagem coloquial, na intenção de enfatizar uma certeza. Não têm lógica.

²Os jogos de azar são antiquíssimos e foram praticados não só como apostas, mas também para prever o futuro, decidir conflitos, dividir heranças. De qualquer modo, a teoria de probabilidades tem em Blaise Pascal, que viveu no século XVII, uma figura de destaque.

Mas a definição clássica de probabilidade, que se aplica bem aos jogos de azar é, de certa forma, uma definição “teórica”. Mesmo sem ter feito qualquer observação ou coleta de dados, construímos o espaço amostral e associamos um valor para a probabilidade de ocorrer cada evento. Na área de saúde, porém, é preciso dispor de dados para estimar probabilidades.

Perguntas como “qual é a probabilidade de um nascituro apresentar doença ou defeito sério?” ou “qual é a probabilidade de um recém-nascido chegar aos 90 anos?” ou “qual é a probabilidade de um fumante ter câncer do pulmão?” só podem ser respondidas com base em dados. Então é importante entender que, na área de saúde, as probabilidades são *estimadas* por freqüências relativas.

A freqüência relativa de um evento, obtida de *uma série de dados coletados nas mesmas condições*, estima a probabilidade de esse evento ocorrer.

As freqüências relativas são *empíricas* porque são calculadas com base nos *dados de uma amostra*. As amostras fornecem estimativas variáveis, mesmo que tais amostras tenham sido tomadas no mesmo local e na mesma época. As probabilidades são *teóricas* porque são construídas com base em *teoria* ou com base nos *dados de toda a população em estudo*.

Exemplo 8.4: Estimativa de probabilidade por freqüência relativa.

Foram examinadas³ 2.000 crianças em idade escolar e observou-se que 65 delas tinham ausência congênita de um ou mais dentes permanentes (anodontia parcial). Qual é a probabilidade de uma criança ter anodontia parcial?

Solução

Com base nos dados, podemos construir uma tabela.

TABELA 8.1
Distribuição dos escolares segundo o fato de terem ou não anodontia parcial.

<i>Anodontia parcial</i>	<i>Freqüência</i>	<i>Freqüência relativa</i>
Sim	63	0,0315
Não	1.937	0,9685
Total	2.060	1,0000

Com base na amostra, estima-se que a probabilidade de uma criança ter anodontia parcial é 0,0315 ou 3,15%.

³VEDOVELO FILHO, M. Prevalência de agenesias dentárias em escolares de Piracicaba, 1972. [Tese (mestrado) FOP-INICAMP].

8.3 – EVENTOS MUTUAMENTE EXCLUSIVOS E EVENTOS INDEPENDENTES

8.3.1 – Eventos mutuamente exclusivos

Dois eventos são *mutuamente exclusivos* quando não podem ocorrer ao mesmo tempo.

Exemplo 8.5: Eventos mutuamente exclusivos.

- Quando se joga uma moeda, ou sai cara, ou sai coroa. Os dois eventos não podem ocorrer ao mesmo tempo: a saída de cara *exclui* a possibilidade de ter saído coroa.
- Se a cirurgia foi um sucesso, fica excluída a possibilidade de ter sido um fracasso.
- Se o paciente tem IMC igual a 35, fica excluída a possibilidade de ter, naquele momento, IMC igual a 25 (pode até ser uma meta).

8.3.2 – Eventos independentes

8.3.2.1 – Conjuntos

Antes de definir eventos independentes, vamos lembrar um pouco da teoria dos conjuntos, que você já deve ter estudado.

União de dois conjuntos: na linguagem comum, usamos a expressão *ou* no sentido *exclusivo*, isto é, quando dizemos “João ou José” queremos dizer “um dos dois”, não ambos. Na linguagem dos conjuntos, que é a linguagem das probabilidades, “A ou B” significa “A ou B ou ambos”. Escrevemos:

$$A \cup B$$

e lê-se: “A união B”.

Exemplo 8.6: União de dois conjuntos ou a regra do “ou”.

Linguagem comum: quando você diz “quero sorvete de creme *ou* de chocolate”, significa que aceita qualquer um deles — e *não* que você aceita um deles, ou o outro, ou os dois!

Linguagem dos conjuntos: uma médica suspeita que sua paciente, que tem câncer de mama, tenha desenvolvido a doença na medula ou no fígado. Isto significa que a doença pode ter atingido a medula, ou o fígado, ou os dois.

Interseção de dois conjuntos: a idéia de *dois eventos que ocorrem juntos* é expressa pela conjunção “e”. Na linguagem dos conjuntos, que é a linguagem das probabilidades, escrevemos:

$$A \cap B$$

e lê-se: “A interseção B”; significa “A e B juntos”.

Exemplo 8.7: Interseção de dois conjuntos ou a regra do “e”.

Linguagem comum: quando você pede um sorvete e diz “quero de creme e chocolate”, significa que você quer os dois sabores.

Linguagem dos conjuntos: quando uma enfermeira diz à parturiente que ela acabou de dar à luz um menino e uma menina, isso significa gêmeos.

8.3.2.2 – Condição de independência

No nosso dia-a-dia, muitas vezes dizemos “uma coisa não tem nada a ver com outra”. Em linguagem técnica, queremos dizer que os eventos são *independentes*. O Exemplo 8.8 serve para ilustrar a *condição de independência*, que veremos em seguida. Mas você intui o resultado, mesmo sem ver os cálculos. Veja a pergunta: quando se jogam um dado e uma moeda, o que ocorre na moeda influí no que sai no dado ou “não tem nada a ver”?

Exemplo 8.8: Condição de independência.

Um dado e uma moeda são jogados ao mesmo tempo. Qual é a probabilidade de ocorrer cara na moeda e face 6 no dado?

Solução

Na Tabela 8.2 está o espaço amostral.

TABELA 8.2
Eventos possíveis no jogo de um dado e uma moeda.

<i>Dado</i>	<i>Moeda</i>	
	<i>Cara</i>	<i>Coroa</i>
1	1; Cara	1; Coroa
2	2; Cara	2; Coroa
3	3; Cara	3; Coroa
4	4; Cara	4; Coroa
5	5; Cara	5; Coroa
6	6; Cara	6; Coroa

A Tabela 8.2 mostra que seis dos 12 eventos do espaço amostral correspondem à saída de cara na moeda. Então a probabilidade desse evento é:

$$P(\text{cara}) = \frac{6}{12} = \frac{1}{2}$$

A Tabela 8.2 também mostra que dois dos 12 eventos correspondem à saída de seis no dado. A probabilidade é:

$$P(6) = \frac{2}{12} = \frac{1}{6}$$

Na mesma Tabela, você vê que apenas um dos 12 eventos corresponde ao que foi pedido: cara na moeda e 6 no dado — um conjunto interseção. A probabilidade é:

$$P(\text{cara} \cap 6) = \frac{1}{12}$$

Então, para este exemplo:

$$P(\text{cara} \cap 6) = P(\text{cara}) \times P(6) = \frac{1}{2} \times \frac{1}{6} = \frac{1}{12}$$

Dois eventos são independentes se a probabilidade de que ocorram juntos é igual ao produto das probabilidades de que ocorram em separado.

Escreve-se:

$$P(A \cap B) = P(A) \times P(B)$$

Esta é a *condição de independência* de dois eventos.

Exemplo 8.9: Eventos independentes na área da saúde.

Para determinar se existe associação entre implantes mamários e doenças do tecido conjuntivo e outras doenças⁴, foram observadas, durante vários anos, 749 mulheres que haviam recebido implante e exatamente o dobro de mulheres que não haviam recebido o implante. Verificou-se que cinco das mulheres que haviam recebido implantes e 10 das que não haviam recebido implante tiveram doenças do tecido conjuntivo. Você acha que ter doenças do tecido conjuntivo não depende de a mulher ter implantes mamários?

⁴GABRIEL SE et alii. Risk of connective tissues diseases and other disorders after breast implantation. *New Engl J Med* 330:1697-1702, 1994. Apud: MOTULSKY, H. **Intuitive Biostatistics**. Nova York, Oxford University Press, 1995. p.318.

Solução

Com base nos dados, podemos construir a Tabela 8.3.

TABELA 8.3

Distribuição de mulheres com implante mamário e o fato de terem ou não doenças do tecido conjuntivo e outras.

<i>Implante mamário</i>	<i>Doenças do tecido conjuntivo e outras</i>		<i>Total</i>	<i>Proporção que receberam implante mamário</i>
	<i>Sim</i>	<i>Não</i>		
Sim	5	744	749	$\frac{749}{2.247}$
Não	10	1.488	1.498	$\frac{1.498}{2.247}$
Total	15		2.232	2.247
Proporção de mulheres que tiveram doença	$\frac{15}{2.247}$	$\frac{2.232}{2.247}$		

A Tabela 8.3 mostra que 749 das 2.247 mulheres observadas receberam implante mamário. Então a probabilidade de, nessa amostra, uma mulher escolhida ao acaso ter implante mamário é:

$$\frac{749}{2.247}$$

A Tabela 8.3 também mostra que 15 das 2.247 mulheres observadas tiveram doenças do tecido conjuntivo e outras doenças. Então a probabilidade de, nessa amostra, uma mulher escolhida ao acaso ter doença do tecido conjuntivo e outras doenças é:

$$\frac{15}{2.247}$$

Como cinco das 2.247 mulheres observadas receberam implante mamário e tiveram doenças do tecido conjuntivo e outras doenças, a probabilidade de ter implante mamário e ter doença é:

$$\frac{5}{2.247}$$

Agora, é fácil verificar se ocorre a condição de independência:

$$P(A \cap B) = P(A) \times P(B)$$

Veja:

$$\frac{749}{2247} \times \frac{15}{2247} = \frac{1}{3} \times \frac{15}{2247} = \frac{5}{2247}$$

Logo, os eventos são independentes porque:

$$P(\text{implante} \cap \text{doença}) = P(\text{implante}) \times P(\text{doença})$$

8.3.2.3 – Diferença nos conceitos

É importante considerar aqui o perigo de *confundir* eventos independentes com eventos mutuamente exclusivos. Às vezes, as pessoas entendem que as duas expressões querem dizer a mesma coisa: que os eventos não se sobrepõem. No entanto, eventos mutuamente exclusivos — se um ocorre, o outro não pode ocorrer — *não* são independentes.

Pense no jogo de uma moeda: quando se joga uma moeda, não há como ocorrer cara e coroa ao mesmo tempo. Logo, esses eventos são *mutuamente exclusivos*. Eles são *independentes*? *Não*: a probabilidade de sair cara é $1/2$, mas dada a condição de que ocorreu coroa, é zero. Então a probabilidade de sair cara muda, se sair coroa. Pense nisso.

8.4 – PROBABILIDADE CONDICIONAL

Muitas vezes relatamos probabilidades que ocorrem sob uma dada condição. Por exemplo, a probabilidade de um universitário trabalhar bem em um computador é maior se estivermos nos referindo aos alunos de Ciências da Computação e não a todos os universitários do Brasil.

Denomina-se *probabilidade condicional* à probabilidade de ocorrer determinado evento sob uma dada condição. Indica-se a probabilidade condicional de ocorrer o evento A sob a condição de B ter ocorrido por $P(A|B)$, que se lê: “probabilidade de A dado B”.

Exemplo 8.10: Cálculo de probabilidade condicional.

Um dado foi lançado. Qual é a probabilidade de: a) ter ocorrido a face 5? b) ter ocorrido a face 5, sabendo que ocorreu face com número ímpar?

Solução

- a) Quando se joga um dado, pode ocorrer um dos eventos:
1, 2, 3, 4, 5 ou 6.

Só existe um evento com o atributo desejado (face 5). Então a probabilidade é:

$$\frac{1}{6}$$

- b) Dada a condição, de que ocorreu número ímpar, só podem ter ocorrido os números:
1, 3, ou 5.

Note que houve *redução do espaço amostral* — porque foi dada a condição “saiu número ímpar”. Como só existe um evento com o atributo desejado (face 5), a probabilidade é:

$$\frac{1}{3}$$

Vamos discutir um pouco mais o Exemplo 8.10. A probabilidade de ocorrer face 5 no dado foi modificada quando foi feita a *redução do espaço amostral*. Isto foi feito porque foi dada a *condição* em que o evento ocorreu: havia saído número ímpar.

Aprendemos que a probabilidade de ocorrer determinado evento depende, muitas vezes, das condições em que ocorre esse evento. Isto é conhecido na área de saúde: na condição de obeso, a probabilidade de doença cardíaca aumenta; na condição de chuva e vento fortes, a probabilidade de acidente automobilístico aumenta; em boas condições de higiene oral, a probabilidade de uma pessoa ter cáries diminui. Muitas pesquisas são feitas para estudar os *fatores que modificam as probabilidades*. Veja um exemplo em que o valor de probabilidade se modifica quando é imposta uma condição.

Exemplo 8.11: Probabilidade condicional na área de saúde.

Para verificar se a condição de hospital público ou privado modifica a probabilidade de cesarianas foram apresentados os dados que estão na Tabela 8.4, coletados em dois hospitais da mesma cidade.

TABELA 8.4
Número de cesarianas em dois hospitais, um público e um privado.

<i>Hospital</i>	<i>Cesariana</i>		<i>Total</i>	<i>Proporção de cesarianas</i>
	<i>Sim</i>	<i>Não</i>		
Privado	89	11	100	$\frac{89}{100} = 0,890$
Público	350	1.091	1.441	$\frac{350}{1.441} = 0,243$

Fonte: Fabri et alii (2002)⁵

A Tabela 8.4 mostra que, nos hospitais privados, 89 dos 100 partos foram por cesariana. Então a probabilidade estimada de cesariana em hospitais privados, com base nessa amostra, é 0,890.

A Tabela 8.4 também mostra que 350 dos 1.441 partos feitos em hospitais públicos foram por cesariana. Então a probabilidade estimada de cesariana em hospitais públicos, com base nessa amostra, é 0,243.

Veja a relação entre as duas estimativas de probabilidade:

$$\frac{0,890}{0,243} = 3,7$$

É fácil ver que a probabilidade estimada de cesariana é bem maior em hospitais privados (3,7 vezes maior). Então a probabilidade estimada de cesariana está condicionada à categoria do hospital, se público ou privado⁶.

⁵FABRI, RH et alii. Estudo comparativo das indicações de cesariana entre um hospital público-universitário e um hospital privado. *Rev. Bras. Saúde Mater. Infant.* v. 2 n. 1 Recife Jan./Abril, 2002.

⁶Os autores explicam que o aumento de cesarianas no hospital privado deve ser decorrente de iteratividade, distocia e a escolaridade mais elevada das pacientes.

*8.5 – TEOREMA DA SOMA OU A REGRA DO “OU”

A probabilidade de ocorrer A ou B é dada pela probabilidade de ocorrer A , mais a probabilidade de ocorrer B , menos a probabilidade de ocorrer A e B (porque a probabilidade de ocorrer A e B é contada duas vezes). Escreve-se:

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

No entanto, se A e B são *mutuamente exclusivos*, a probabilidade de ocorrer A e B é dada pela probabilidade de ocorrer A , mais a probabilidade de ocorrer B . Escreve-se:

$$P(A \cup B) = P(A) + P(B)$$

Exemplo 8.12: A ou B.

Uma carta será retirada ao acaso de um baralho. Qual é a probabilidade de sair uma carta de espadas ou um ás?

Solução

Como um baralho tem 52 cartas, das quais 13 são de espadas e quatro são ases, alguém poderia pensar que a probabilidade de sair uma carta de espadas ou um ás é dada pela soma:

$$\frac{13}{52} + \frac{4}{52}$$

mas esta resposta estaria errada, porque existe uma carta, o ás de espadas, que é tanto ás como espadas. Então o ás de espadas teria sido contado duas vezes. A probabilidade de sair uma carta de espadas ou um ás é dada por

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

$$= \frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{16}{52} = \frac{4}{13}$$

Exemplo 8.13: A ou B, disjuntos.

Uma urna contém quatro bolas: duas brancas, uma vermelha e uma azul. Retira-se uma bola da urna ao acaso. Qual a probabilidade de ter saído uma bola colorida, isto é, azul ou vermelha?

Solução

A probabilidade de sair bola azul é:

$$\frac{1}{4}$$

e a probabilidade de sair bola vermelha é:

$$\frac{1}{4}$$

Então a probabilidade de sair bola colorida, isto é, azul ou vermelha, é dada pela soma:

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) \\ &= \frac{1}{4} + \frac{1}{4} = \frac{2}{4} = \frac{1}{2} \end{aligned}$$

***8.6 – TEOREMA DO PRODUTO OU A REGRA DO “E”**

Muitas vezes queremos saber a probabilidade de dois eventos ocorrerem juntos, ou um em seguida do outro. Queremos, então, a probabilidade do conjunto interseção. Para resolver esse tipo de problema, existe a regra do *e* ou teorema do produto.

Se A e B são dependentes, a probabilidade de ocorrer A e B é dada pela probabilidade de ocorrer A, multiplicada pela probabilidade (*condicional*) de ocorrer B, dado que A tenha ocorrido. Escreve-se:

$$P(A \text{ e } B) = P(A) \times P(B | A).$$

Se A e B são eventos independentes, a probabilidade de ocorrer A e B é dada pela probabilidade de ocorrer A, multiplicada pela probabilidade de ocorrer B. Escreve-se:

$$P(A \text{ e } B) = P(A) \times P(B).$$

Exemplo 8.13: Teorema do produto: eventos independentes.

Uma moeda será jogada duas vezes. Qual é a probabilidade de ocorrer cara nas duas jogadas?

Solução

A probabilidade de ocorrer cara na primeira jogada é:

$$\frac{1}{2}$$

A probabilidade de ocorrer cara na segunda jogada também é:

$$\frac{1}{2}$$

porque ocorrer cara na primeira jogada não modifica a probabilidade de ocorrer cara na segunda jogada (os eventos são independentes). Para obter a probabilidade de ocorrer cara nas duas jogadas (primeira e segunda), faz-se o produto:

$$\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

Exemplo 8.14: Teorema do produto: eventos dependentes.

Uma urna contém três bolas: duas brancas e uma vermelha. Retiram-se duas bolas da urna, uma em seguida da outra e sem que a primeira tenha sido recolocada. Qual é a probabilidade de as duas serem brancas?

Solução

A probabilidade de a primeira bola ser branca é:

$$\frac{1}{3}$$

A probabilidade de a segunda bola ser branca depende do que ocorreu na primeira retirada. Se a bola branca saiu na primeira retirada, a probabilidade de a segunda também ser branca é:

$$\frac{1}{2}$$

Para obter a probabilidade de as duas bolas retiradas serem brancas, faz-se o produto:

$$\frac{1}{3} \times \frac{1}{2} = \frac{1}{6}$$

8.7 – EXERCÍCIOS RESOLVIDOS

8.7.1 – De uma classe com 30 alunos, dos quais 14 são meninos, um aluno é escolhido ao acaso. Qual é a probabilidade de: a) o aluno escolhido ser um menino? b) o aluno escolhido ser uma menina?

A classe tem 30 alunos ($n = 30$) e todos têm a mesma probabilidade de serem escolhidos. Como 14 são meninos ($m = 14$):

- a) a probabilidade de o aluno escolhido ser menino é $14/30$ ou $7/15$.
- b) a probabilidade de o aluno escolhido ser menina é $16/30$ ou $8/15$.

8.7.2 – Uma pessoa comprou um número de uma rifa que tem 100 números e irá sortear cinco prêmios. Qual é a probabilidade de essa pessoa: a) ganhar um prêmio? b) de não ganhar?

Todos os 100 números ($n = 100$) da rifa têm igual probabilidade de serem sorteados. Serão sorteados números ($m = 5$). Então:

- a) a probabilidade de uma pessoa que comprou um número ser sorteada é $5/100$ ou $1/20$.
- b) a probabilidade de a pessoa não ser sorteada é $95/100$ ou $19/20$.

8.7.3 – Uma urna tem 10 bolas brancas e quatro pretas. Retira-se uma bola ao acaso. Qual é a probabilidade de essa bola: a) ser branca? b) ser preta?

A urna tem 10 bolas brancas e quatro pretas ($n = 14$). Retira-se uma bola ao acaso. A probabilidade de essa bola:

- a) ser branca ($m = 10$) é $10/14$ ou $5/7$;
- b) ser preta ($m = 4$) é $4/14$ ou $2/7$.

8.7.4 – Joga-se um dado. Qual é a probabilidade de sair: a) o número 3? b) número maior do que 3? c) número menor do que 3? d) número par?

Quando se joga um dado, pode ocorrer um dos eventos: 1, 2, 3, 4, 5 ou 6.

- a) Apenas um ($m = 1$) dos seis eventos ($n = 6$) é igual a 3. Então a probabilidade de ocorrer 3 é $1/6$.
- b) Dos seis eventos, três ($m = 3$) são maiores do que 3 (4; 5; 6). Então a probabilidade de ocorrer número maior do que 3 é $3/6$ ou $1/2$.
- c) Dos seis eventos, dois ($m = 2$) são menores do que 3 (1; 2). Então a probabilidade de ocorrer número menor do que 3 é $1/3$.
- d) Dos seis eventos, três ($m = 3$) são números pares (2; 4; 6). Então a probabilidade de ocorrer número par é $1/2$.

- 8.7.5 – Jogam-se duas moedas. Qual é a probabilidade de saírem: a) duas caras? b) duas coroas? c) uma cara e uma coroa?**

Para resolver este problema, é conveniente escrever todos os eventos que podem ocorrer quando se joga uma moeda. Veja a Tabela 8.5.

TABELA 8.5
Resultados possíveis no jogo de duas moedas.

Evento	1 ^a moeda	2 ^a moeda
1	cara	coroa
2	coroa	cara
3	cara	cara
4	coroa	coroa

A Tabela 8.4 mostra $n = 4$ eventos mutuamente exclusivos e igualmente prováveis. A probabilidade de saírem:

- a) duas caras (evento 3 na Tabela) é $1/4$;
- b) duas coroas (evento 4 na Tabela) é $1/4$;
- c) uma cara e uma coroa (eventos 1 e 2 na Tabela) é $2/4$.

- 8.7.6 – Em uma família com três filhos, qual é a probabilidade de os três serem homens? Suponha que a probabilidade de nascer menino é $1/2$.**

Como o sexo de um filho não depende do sexo do anterior, a probabilidade de o primeiro filho ser homem e de o segundo filho ser homem e de o terceiro filho ser homem é, pelo teorema do produto:

$$\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}$$

- 8.7.7 – Em uma família com três filhos, qual é a probabilidade de: a) dois serem homens? b) um ser homem? c) nenhum ser homem? Suponha que meninos e meninas têm a mesma probabilidade de nascer.**

Para resolver este problema, é conveniente escrever todas as possibilidades em uma família com três filhos. Veja a Tabela 8.6.

TABELA 8.6
Resultados possíveis em uma família com três filhos.

Evento	1º filho	2º filho	3º filho
1	Homem	Homem	Homem
2	Homem	Homem	Mulher
3	Homem	Mulher	Homem
4	Homem	Mulher	Mulher
5	Mulher	Homem	Homem
6	Mulher	Homem	Mulher
7	Mulher	Mulher	Homem
8	Mulher	Mulher	Mulher

A probabilidade de:

- a) dois serem homens (eventos 2; 3 e 5 na Tabela) é 3/8;
- b) de um ser homem (eventos 4; 6 e 7 na Tabela) é 3/8
- c) nenhum ser homem (evento 8 na Tabela) é 1/8.

8.7.8 – Um casal tem dois filhos. Qual é a probabilidade de: a) o primogênito ser homem? b) os dois filhos serem homens? c) pelo menos um filho ser homem?

Suponha que a probabilidade de nascer menino é 1/2 e que o sexo do segundo filho não depende do sexo do primeiro. Então:

- a) a probabilidade de o primogênito ser homem é

$$1/2;$$

- b) a probabilidade de os dois filhos serem homens pode ser obtida pelo teorema do produto (o primeiro ser homem e o segundo ser homem):

$$\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

- c) a probabilidade de ser homem pelo menos um dos filhos pode ser obtida pelo teorema da soma (o primeiro ser homem, ou o segundo ser homem, ou os dois serem homens):

$$\frac{1}{4} + \frac{1}{4} + \frac{1}{4} = \frac{3}{4}$$

8.7.9 – No cruzamento de ervilhas amarelas homozigotas (AA) com ervilhas verdes homozigotas (aa) ocorrem ervilhas amarelas heterozigotas (Aa). Se estas ervilhas forem cruzadas entre si, ocorrem três ervilhas amarelas para cada ervilha verde (a proporção é de três para um). Suponha que foram pegas, ao acaso, três ervilhas resultantes do cruzamento de ervilhas amarelas heterozigotas. Qual a probabilidade de as três serem verdes?

A probabilidade de uma ervilha resultante do cruzamento Aa x Aa ser verde é 1/4. Logo, a probabilidade de as três ervilhas, pegas ao acaso, serem verdes é:

$$\frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} = \frac{1}{64}$$

8.7.10 – Qual é a probabilidade de o filho de um homem normal (XY) e de uma filha de hemofílico (X_hX) ser hemofílico (X_hY)?

Um homem normal (XY) não transmite a hemofilia para gerações seguintes. Uma mulher portadora do gene X_h tem 50% de probabilidade de ter um filho hemofílico. O filho será normal (XY) ou hemofílico (X_hY), com a mesma probabilidade, isto é, 1/2.

8.7.11 – Jogam-se duas moedas ao mesmo tempo. Os eventos “cara na primeira moeda” e “faces iguais nas duas moedas” são independentes?

Veja o espaço amostral:

TABELA 8.7
Resultados possíveis no jogo de duas moedas

Evento	1 ^a moeda	2 ^a moeda
1	Cara	Cara
2	Cara	Coroa
3	Coroa	Cara
4	Coroa	Coroa

Os eventos possíveis são quatro. Só um deles (cara-cara) atende “cara na primeira moeda” — que chamaremos de A — e “faces iguais nas duas moedas” — que chamaremos B. Então a probabilidade pedida é:

$$P(A \cap B) = \frac{1}{4}$$

Mas a probabilidade de “cara” na primeira moeda é:

$$P(A) = \frac{2}{4} = \frac{1}{2}$$

e a probabilidade de “faces iguais nas duas moedas” é:

$$P(B) = \frac{2}{4} = \frac{1}{2}$$

Então:

$$P(A \cap B) = P(A) \times P(B)$$

A condição de independência foi, portanto, satisfeita. Os eventos “cara na primeira moeda” e “faces iguais nas duas moedas” são independentes.

8.8 – EXERCÍCIOS PROPOSTOS

8.8.1 – Uma carta é retirada ao acaso de um baralho bem embaralhado. Qual é a probabilidade de: a) ser um ás? b) ser uma carta de ouro? c) ser um ás de ouro?

8.8.2 – Uma urna tem 10 bolas numeradas de 1 a 10. Retira-se uma bola ao acaso. Qual é a probabilidade de essa bola: a) ter número maior do que 7? b) ter número menor do que 7? c) ter número 1 ou 10?

8.8.3 – Uma urna tem 15 bolas numeradas de 1 a 15. Retira-se uma bola ao acaso. Qual é a probabilidade de essa bola: a) ter número par? b) ter número ímpar? c) ter número maior do que 15?

8.8.4 – Para melhorar as condições de pacientes com determinada doença crônica, existem cinco drogas: A, B, C, D e E. Um médico tem verba para comparar apenas três delas. Se ele escolher três drogas ao acaso para comparar, qual é a probabilidade de: a) a droga A ser escolhida? b) as drogas A e B serem escolhidas?

8.8.5 – Dois dados, um vermelho, outro azul, são lançados ao mesmo tempo e se pergunta: a) qual é a probabilidade de ocorrer face 6 no dado vermelho? b) qual é a probabilidade de ocorrer face 6 no dado vermelho, sabendo que saiu face 6 no dado azul?

8.8.6 – Um exame feito em jovens que terminaram o curso fundamental mostrou que 20% foram reprovados em Matemática, 10% foram reprovados em Português e 5% foram reprovados tanto em Matemática como em Português. Os eventos “ser reprovado em Matemática” e “ser reprovado em Português” são independentes?

8.8.7 – Um casal tem dois filhos. Qual é a probabilidade de: a) o segundo filho ser homem? b) o segundo filho ser homem, dado que o primeiro é homem?

8.8.8 – A probabilidade de determinado teste para a AIDS dar resultado negativo em portadores de anticorpos contra o vírus (falso-negativo) é 10%. Supondo que falsos-negativos ocorrem independentemente, qual é a probabilidade de um portador de anticorpos contra o vírus da AIDS, que se apresentou três vezes para o teste, ter tido, nas três vezes, resultado negativo?

8.8.9 – Uma pessoa normal, filha de pais normais, tem um avô albino (aa). Se os outros avós não forem portadores do gene para albinismo (AA), qual é a probabilidade de essa pessoa ser portadora do gene para albinismo (Aa)?

8.8.10 – Suponha que a probabilidade de uma pessoa ser do tipo sanguíneo O é 40%, ser A é 30% e ser B é 20%. Suponha ainda que o fator Rh não dependa do tipo sanguíneo e que a probabilidade de Rh^+ é de 100%. Nestas condições, qual é a probabilidade de uma pessoa tomada ao acaso da população ser: a) O, Rh^+ ? b) AB, Rh^- ?

(página deixada intencionalmente em branco)

Distribuição Binomial

9

(página deixada intencionalmente em branco)

A Estatística formaliza o que nós, muitas vezes, já sabemos. Por exemplo, você sabe que as idades das pessoas da sua família variam. Portanto, você tem consciência da *variabilidade*. Você sabe que no Nordeste faz calor o ano todo, o que não acontece no Sul. Então você tem consciência de que, no decorrer de um ano, as temperaturas dos estados nordestinos são, em *média*, mais altas do que as temperaturas dos estados do sul do país. E se você acha que o peso de uma pessoa depende da altura está mostrando que sabe o que é *correlação*. Ainda, todos nós sabemos que ganhar na loteria não é fácil. Temos, portanto, percepção sobre *probabilidade*. Vamos agora definir o que é variável aleatória — que você, intuitivamente, talvez já conheça.

9.1 – VARIÁVEL ALEATÓRIA

Quando você joga uma moeda, ou sai cara, ou sai coroa. O acaso determina o resultado. Quando, num jogo de baralho, você tira uma carta, pode sair carta de paus, de ouros, de espadas ou de copas. O acaso determina o resultado. Mas não é apenas nos jogos de azar que os resultados ocorrem ao acaso.

Imagine que uma casa foi escolhida por sorteio de uma comunidade de 5.000 domicílios. Todas as casas tiveram, portanto, igual probabilidade de serem amostradas. Um entrevistador vai, então, até a casa selecionada e pergunta gênero, idade e renda de todos os moradores. As respostas estão, evidentemente, associadas à casa escolhida. Se a casa sorteada tivesse sido outra, provavelmente o conjunto de respostas seria diferente. Logo, as respostas coletadas pelo entrevistador foram determinadas pelo acaso, uma vez que a casa foi escolhida por *processo aleatório*.

Uma variável é aleatória quando o acaso tem influência em seus valores.

As variáveis aleatórias são indicadas por *números*. Se um jogador ganha quando sai cara, associamos o número 1 à saída de cara e o número zero à saída de coroa. Se a pessoa entrevistada numa pesquisa disser que tem 42 anos, a variável aleatória que representa idade de pessoas assumiu, nesse caso, valor 42.

As variáveis aleatórias são, portanto, *numéricas*. Logo, podem ser *discretas* e *contínuas*. Neste Capítulo vamos estudar as variáveis aleatórias discretas.

9.1.1 – Variável aleatória binária

Alguns experimentos só podem resultar em uma de duas possibilidades: o evento no qual estamos interessados, que é denominado “sucesso” e o evento contrário, chamado de “fracasso”. O exemplo mais conhecido é o jogo de moedas. Quando se joga uma moeda, ou sai cara ou sai coroa — as duas faces não podem ocorrer ao mesmo tempo. Dizemos então que a variável aleatória é *binária*.

Na área de saúde, encontramos muitas variáveis binárias. Veja alguns exemplos:

- um exame laboratorial pode dar resultado positivo ou negativo;
- um nascituro pode ser menino ou menina;
- um medicamento pode surtir ou não o efeito esperado;
- um doador de sangue pode ser Rh+ ou Rh-;
- a dieta pode ser adequada ou não-adequada;
- determinado material pode estar contaminado ou não.

Variável aleatória binária é aquela que resulta em um de dois eventos mutuamente exclusivos — ou é “sucesso”, ou é “fracasso”. Associamos o valor 1 ao “sucesso” e valor zero ao “fracasso”.

9.1.2 – Variável aleatória binomial

Muitas vezes contamos o número de vezes que ocorre o evento de interesse (ou sucesso), em uma série de tentativas ou de experimentos. Por exemplo:

- Um jogador conta *quantas* caras saem quando lança 10 moedas.
- Um pesquisador conta *quantos*, dos 500 chefes de família que entrevistou, eram mulheres.
- Um médico conta *quantos*, dos 100 pacientes que tratou com uma nova droga, ficaram curados.
- Um biomédico conta *quantos*, dos 32 hemogramas que fez no dia, indicaram doença contagiosa.
- Uma enfermeira conta *quantos*, dos nascidos vivos durante determinado ano em uma maternidade, tinham doença ou defeito sério.

A variável que resulta da soma dos resultados de uma variável aleatória binária em n tentativas é uma variável aleatória binomial.

Exemplo 9.1: Variável aleatória binomial.

Escreva os eventos que podem ocorrer quando se lança uma moeda duas vezes. Conte o número X de caras em cada um desses eventos. Apresente os resultados em uma tabela.

Solução

TABELA 9.1
Eventos possíveis e número de caras quando uma moeda é lançada duas vezes.

<i>Eventos possíveis</i>	<i>Valor de X</i>
coroa e coroa	0
coroa e cara	1
cara e coroa	1
cara e cara	2

9.2 – DISTRIBUIÇÃO DE PROBABILIDADES

Os valores observados da variável aleatória X são indicados por x_1, x_2, \dots, x_k e as respectivas probabilidades por $P(x_1), P(x_2), \dots, P(x_k)$. Obrigatoriamente:

1. A soma das probabilidades de ocorrerem todos os valores possíveis de X é 1.
2. A probabilidade de ocorrer qualquer valor de X é igual ou maior que zero — *não* pode ser negativa.

Exemplo 9.2: Distribuição de probabilidades.

A variável X representa o número de caras que se obtêm quando se lança uma moeda duas vezes. Apresente a distribuição de probabilidades de X em tabela e em gráfico.

Solução

Quando se joga uma moeda duas vezes, os eventos possíveis são:

coroa, coroa;
coroa, cara;
cara, coroa;
cara, cara.

Se saírem duas coroas, a variável X assume valor zero. A probabilidade de isso acontecer é:

$$P(\text{coroa}) \times P(\text{coroa}) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} = 0,25$$

Se saírem uma coroa e uma cara, a variável X assume valor um. A probabilidade de isso acontecer é:

$$P(\text{coroa}) \times P(\text{cara}) + P(\text{cara}) \times P(\text{coroa}) = \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} + \frac{1}{4} = 0,50$$

Se saírem duas caras, a variável X assume valor dois. A probabilidade de isso acontecer é:

$$P(\text{cara}) \times P(\text{cara}) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} = 0,25$$

A Tabela 9.2 e a Figura 9.1 apresentam um resumo destes cálculos, ou seja, apresentam a distribuição de probabilidades de X . A soma das probabilidades é 1.

TABELA 9.2
Distribuição de probabilidades do número de caras em dois lançamentos de uma moeda.

Evento	Valor de X	$P(X)$
Coroa e Coroa	0	$\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$
Coroa e Cara ou Cara e Coroa	1	$\frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} = \frac{2}{4}$
Cara e Cara	2	$\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$
Total		1

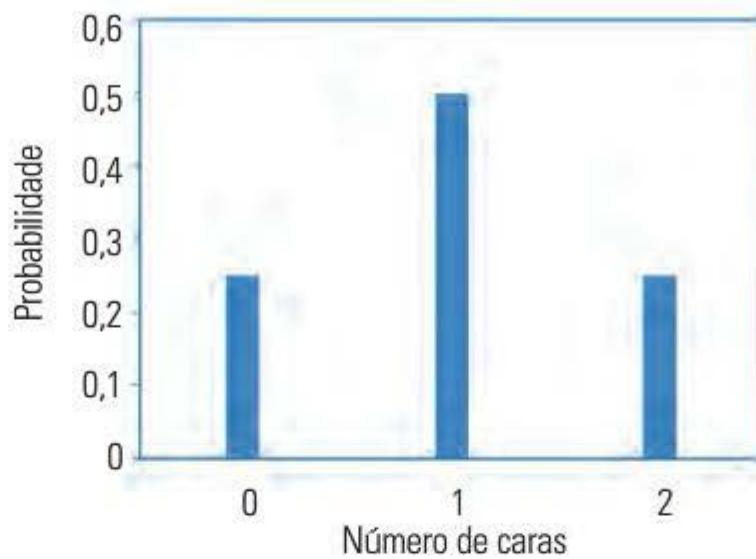


FIGURA 9.1 Distribuição de probabilidades do número de caras em dois lançamentos de uma moeda.

Neste ponto, é importante deixar claro que existe diferença entre *distribuição de probabilidades* e *distribuição de freqüências*. As distribuições de freqüências, tratadas no Capítulo 2, são *empíricas* porque são construídas com base nos dados de amostras. As amostras variam, mesmo que sejam tomadas no mesmo local e na mesma época. A distribuição de probabilidades é *teórica* porque é construída com base em teoria ou com base nos dados de toda a população em estudo. A distribuição de probabilidades é estável.

9.3 – DISTRIBUIÇÃO BINOMIAL

Uma distribuição de probabilidades bem conhecida é a *distribuição binomial*, que estuda o número X de sucessos em n tentativas e as suas respectivas probabilidades.

Para aprender a trabalhar com a distribuição binomial, imagine que em determinada maternidade nasceram três bebês em um dia. Vamos estudar a distribuição de meninos em três nascimentos.

Fazendo A indicar menina e 0 indicar menino, os eventos possíveis são os seguintes:

AAA	AA0	A00	000
	AOA	0AO	
	OAA	00A	

O número de meninos que pode ocorrer em três nascimentos é uma *variável aleatória binomial*, que indicaremos por X . A Tabela 9.3 apresenta os valores possíveis de X e o número de vezes que cada um deles ocorre, conforme mostrado no esquema.

TABELA 9.3
Números possíveis de meninos em três nascimentos.

Valor de X	Freqüência
0	1
1	3
2	3
3	1

Seja p a probabilidade de nascer menino e q a probabilidade de nascer menina. Evidentemente, $p + q = 1$.

Se nascerem três meninas, isto é, se ocorrer o evento AAA, a variável aleatória X assume valor zero, com probabilidade:

$$P[X = 0] = P[A] \times P[A] \times P[A] = q \times q \times q = q^3$$

Se nascerem duas meninas e um menino, X assume valor 1. Mas duas meninas e um menino podem ocorrer de três maneiras diferentes. Veja as probabilidades:

$$P[A] \times P[A] \times P[O] = q \times q \times p = pq^2$$

$$P[A] \times P[O] \times P[A] = q \times p \times q = pq^2$$

$$P[O] \times P[A] \times P[A] = p \times q \times q = pq^2$$

Então:

$$P[X = 1] = 3pq^2$$

Se nascerem uma menina e dois meninos, X assume valor 2. Mas uma menina e dois meninos podem ocorrer de três maneiras diferentes. Veja as probabilidades:

$$P[A] \times P[O] \times P[O] = q \times p \times p = p^2q$$

$$P[O] \times P[A] \times P[O] = p \times q \times p = p^2q$$

$$P[O] \times P[O] \times P[A] = p \times p \times q = p^2q$$

Então:

$$P[X = 2] = 3p^2q$$

Se nascerem três meninos, isto é, se ocorrer o evento 000, a variável aleatória X assume valor 3, com probabilidade:

$$P[X = 3] = P[0] \times P[0] \times P[0] = p \times p \times p = p^3$$

A distribuição binomial do número X de meninos em $n = 3$ nascimentos está na Tabela 9.4. São dados os resultados possíveis de X e suas respectivas probabilidades.

TABELA 9.4
Distribuição de probabilidades do número de meninos em três nascimentos.

Valor de X	Probabilidade
0	q^3
1	$3pq^2$
2	$3p^2q$
3	p^3

Vamos considerar, por facilidade, que a probabilidade de nascer menino é $p = 0,5$ e que a probabilidade de nascer menina é $q = 0,5$, embora se saiba que a probabilidade de nascer menino é ligeiramente maior do que 0,5. Estamos, também, ignorando nascimentos de gêmeos e nascimentos múltiplos. Considerando: $p = 0,5$ e $q = 0,5$ obtemos a distribuição de probabilidades do número de meninos em três nascimentos, apresentada na Tabela 9.5 e na Figura 9.2.

TABELA 9.5
Distribuição de probabilidades do número de meninos em três nascimentos ($p = q = 0,5$).

Valor de X	$P(X)$
0	$\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8} = 0,125$
1	$3 \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{3}{8} = 0,375$
2	$3 \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{3}{8} = 0,375$
3	$\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8} = 0,125$
Total	1

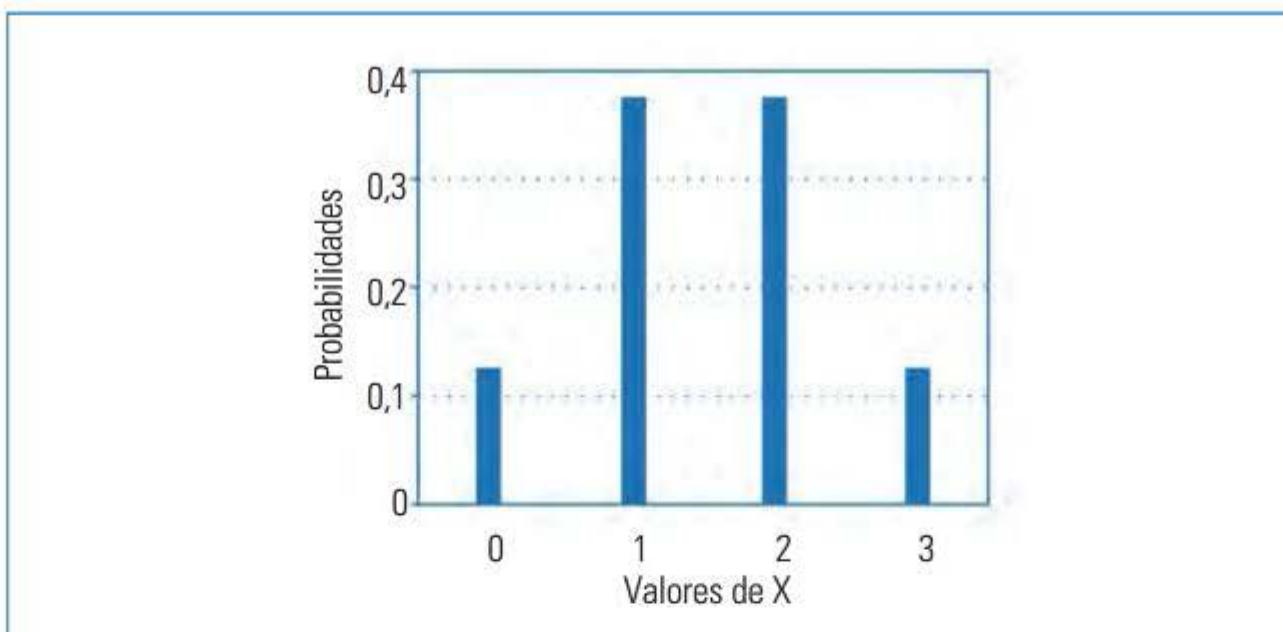


FIGURA 9.2 Distribuição de probabilidades do número de meninos em três nascimentos.

9.3.1 – Caracterização da distribuição binomial

Uma distribuição binomial tem as seguintes características:

- Consiste de n ensaios, ou n tentativas, ou n eventos idênticos.
- Cada ensaio só pode resultar em um de dois resultados, identificados como “sucesso” e “fracasso” — com valores 1 e zero, respectivamente.
- A variável aleatória X é o número de sucessos em n ensaios.
- A probabilidade de sucesso (ocorrer o evento de interesse) é p e o valor de p permanece o mesmo em todos os ensaios.
- Os ensaios são independentes: o resultado de um ensaio não tem efeito sobre o resultado de outro.

A distribuição binomial fica, portanto, definida quando são dados *dois parâmetros*:

- n , isto é, o *número de ensaios* (p. ex., se uma moeda for lançada 10 vezes);
- p , isto é, a probabilidade de sucesso em uma tentativa (por exemplo, a probabilidade de sair cara quando se joga uma moeda).

*9.3.2 – Função de distribuição na distribuição binomial

Vamos aceitar, sem demonstração, que, dada uma distribuição binomial de parâmetros n e p , a probabilidade de ocorrerem x eventos favoráveis é dada pela fórmula:

$$\binom{n}{x} p^x x^{(n-x)}$$

em que $\binom{n}{x}$ é a combinação¹ de n , x a x . Portanto, a probabilidade de ocorrerem x eventos favoráveis em n tentativas é dada pela fórmula:

¹Uma rápida revisão sobre análise combinatória está inserida no final deste Capítulo.

$$P(X = x) = \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

Veja, agora, um exemplo que ajuda a entender como trabalhamos com a distribuição binomial.

Exemplo 9.3: Eventos em uma distribuição binomial.

Um dentista vai examinar uma amostra de quatro crianças de 6 anos de idade para saber se elas têm (Sim, indicado por S) ou não (Não, indicado por N) cárie. Quais são os eventos possíveis?

Solução

Os eventos possíveis são os que seguem:

NNNN	NNNS	NNSS	NSSS	SSSS
	NNSN	NSNS	SNSS	
	NSNS	NSSN	SSNS	
	SNNN	SNNS	SSSN	
		SNSN		
		SSNN		

Exemplo 9.4: Distribuição binomial.

Reveja o Exemplo 9.3. Faça X indicar o número de crianças com cárie, p indicar a probabilidade de uma criança ter cárie e q indicar a probabilidade de uma criança não ter cárie. Escreva a distribuição.

Solução

TABELA 9.6
Distribuição de probabilidades do número de crianças com cárie em quatro crianças.

Evento	Valor de X	$P(X)$
Nenhuma criança com cárie	0	q^4
Uma criança com cárie	1	$4pq^3$
Duas crianças com cárie	2	$6p^2q^2$
Três crianças com cárie	3	$4p^3q$
Quatro crianças com cárie	4	p^4

Exemplo 9.5: Distribuição binomial ($n = 4$; $p = 0,4$).

Reveja o Exemplo 9.4. Considere que, na população estudada, a probabilidade de uma criança de 6 anos ter cárie é $p = 0,4$ (ou seja, 40%). Qual é a probabilidade de duas das quatro crianças examinadas terem cárries?

Solução

A Tabela 9.6 mostra a probabilidade de a variável X assumir valor 2. Se a probabilidade de uma criança dessa população ter cárie é $p = 0,4$, então:

$$P(X = 2) = 6p^2q^2 = 6 \times 0,4^2 \times 0,6^2 = 6 \times 0,16 \times 0,36 = 0,3456$$

Exemplo 9.6: Cálculo de probabilidades na distribuição binomial.

Reveja o Exemplo 9.4. A probabilidade de uma criança de 6 anos ter cárie é $p = 0,4$ (ou 40%). Calcule a probabilidade de duas ($X = 2$) das quatro (n) crianças examinadas terem cárries aplicando a fórmula:

$$P(X = 2) = \binom{4}{2} \times 0,4^2 \times 0,6^2 = 0,3456$$

A probabilidade de o dentista encontrar duas de quatro crianças com cárries, nessa população, é 0,3456.

***9.3.3 – Média e variância na distribuição binomial**

A média μ (lê-se: mi) de uma distribuição binomial é dada pela fórmula:

$$\mu = np$$

e a variância σ^2 (lê-se: sigma ao quadrado) é dada pela fórmula:

$$\sigma^2 = npq$$

Exemplo 9.7: Média e variância da distribuição binomial.

A probabilidade de nascer um menino é $p = 0,5$ (ignorando nascimentos de gêmeos e nascimentos múltiplos). Calcule a média e a variância do número de meninos em 1.000 nascituros.

Solução

A média é:

$$\mu = np = 1.000 \times 0,5 = 500 \text{ meninos},$$

e a variância é:

$$\sigma^2 = npq = 1.000 \times 0,5 \times 0,5 = 250.$$

9.4 – REVISÃO SOBRE ANÁLISE COMBINATÓRIA

Se n é um número inteiro positivo maior do que zero, por definição, *fatorial de n* , que se indica por $n!$ é dado por:

$$n! = n(n-1)(n-2)\dots1.$$

O fatorial de 5 é, portanto:

$$5! = 5 \times 4 \times 3 \times 2 \times 1 = 120.$$

O desenvolvimento de um fatorial pode ser interrompido antes de chegar ao número 1, desde que se coloque o símbolo $!$, que indica o fatorial, logo após o último número. Escreve-se:

$$5! = 5 \times 4 \times 3!$$

porque

$$3! = 3 \times 2 \times 1.$$

O fatorial de zero, que se indica por $0!$, é, por definição, igual a 1.

Dado um conjunto de n elementos, onde $n > 0$, e dado o número $x \leq n$, *combinação de n , x a x* , é indicada por:

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

Esta fórmula dá o número de diferentes conjuntos de x elementos que podem ser formados com n elementos distintos.

Seja $n = 5$ e $x = 3$. Então a combinação de 5, 3 a 3 é:

$$\binom{5}{3} = \frac{5!}{3!(5-3)!} = \frac{5!}{3!2!} = 10$$

Convém observar que, para todo n :

$$\binom{n}{n} = \frac{n!}{n!(n-n)!} = \frac{n!}{n!0!} = 1$$

9.5 – EXERCÍCIOS RESOLVIDOS

9.5.1 – Ache o erro nas duas afirmativas feitas em seguida:

- a) A probabilidade de você ser aprovado em Estatística é 2 e de ser reprovado é 0,2.
- b) A probabilidade de chover amanhã é 20%, de ficar nublado sem chuva é 10% e de ter sol é 80%.

A soma de probabilidades deve ser 1 ou 100%. Nas duas afirmativas, as somas excedem o valor 1 ou 100%.

9.5.2 – Numa prova², o aluno deve assinalar a resposta que fornece as datas, na ordem em que estão mencionadas, de três acontecimentos históricos: Descoberta do Brasil, Descoberta da América, Independência do Brasil. As opções são:

- a) 1492, 1822, 1500
- b) 1822, 1492, 1500
- c) 1492, 1500, 1822
- d) 1822, 1500, 1492
- e) 1500, 1492, 1822
- f) 1500, 1822, 1492

Um aluno, que nada sabe sobre a matéria, tenta adivinhar. Qual é distribuição de probabilidades do número de acertos?

A resposta correta é a resposta e: Descoberta do Brasil (1500), Descoberta da América (1492), Independência do Brasil (1822). Mas outras respostas têm as datas de um, ou dois acontecimentos, na ordem correta. Veja o número de acertos em cada resposta.

Resposta	Probabilidade	Nº de acertos na resposta
a	1/6	0
b	1/6	1
c	1/6	1
d	1/6	0
e	1/6	3
f	1/6	1

TABELA 9.7
Distribuição de probabilidade do número de acertos

Acertos	Probabilidade
0	2/6
1	3/6
2	0
3	1/6
Total	1

²Adaptado de MOSTELLER, F. ROURKE, R. E. K., THOMAS JR, G. B. **Probability and Statistics**. Reading, Addison-Wesley, 1961. p. 160.

9.5.3 – Na população branca do Brasil, 85% têm Rh⁺. Três pessoas são amostradas ao acaso dessa população. Construa a distribuição binomial e faça um gráfico.

No problema:

n é o número de pessoas: $n = 3$

X é o número de pessoas com Rh⁺ na amostra

p é a probabilidade de Rh⁺: $p = 0,85$

q é a probabilidade de Rh⁻: $q = 0,15$.

TABELA 9.8
Cálculos intermediários para obter a distribuição binomial.

Eventos	Valores possíveis de X	Cálculos	Probabilidade
Rh ⁺ , Rh ⁺ , Rh ⁺	3	$0,85 \times 0,85 \times 0,85$	0,614125
Rh ⁺ , Rh ⁺ , Rh ⁻	2	$0,85 \times 0,85 \times 0,15$	0,108375
Rh ⁺ , Rh ⁻ , Rh ⁺	2	$0,85 \times 0,15 \times 0,85$	0,108375
Rh ⁻ , Rh ⁺ , Rh ⁺	2	$0,15 \times 0,85 \times 0,85$	0,108375
Rh ⁺ , Rh ⁻ , Rh ⁻	1	$0,85 \times 0,15 \times 0,15$	0,019125
Rh ⁻ , Rh ⁺ , Rh ⁻	1	$0,15 \times 0,85 \times 0,15$	0,019125
Rh ⁻ , Rh ⁻ , Rh ⁺	1	$0,15 \times 0,15 \times 0,85$	0,019125
Rh ⁻ , Rh ⁻ , Rh ⁻	0	$0,15 \times 0,15 \times 0,15$	0,003375

Para construir a tabela de distribuição binomial, você soma as probabilidades dos eventos que levam ao mesmo valor de X . A distribuição é dada na Tabela 9.9.

TABELA 9.9
Distribuição de probabilidades do número de pessoas com Rh⁺, numa amostra de três pessoas.

Valores de X	Probabilidade
3	0,614125
2	0,325125
1	0,057375
0	0,003375



FIGURA 9.3 Distribuição de probabilidades do número de pessoas com Rh⁺, em três pessoas.

9.5.4 – Apresente, em tabela e em gráfico, a distribuição do número de meninos que podem ocorrer em uma família com seis crianças.

No problema, n é o número de crianças (6), p é a probabilidade de menino ($1/2$) e q é a probabilidade de menina ($1/2$). Para obter a probabilidade de X assumir o valor 0, ou seja, de não ocorrer nenhum menino, calcule:

$$\begin{aligned} P(X = 0) &= \binom{6}{0} \times \left(\frac{1}{2}\right)^0 \times \left(\frac{1}{2}\right)^6 = \\ &= \frac{6!}{1!(6-1)!} \times \frac{1}{2^0} \times \frac{1}{2^6} = \frac{1}{64} \end{aligned}$$

Para obter a probabilidade de X assumir o valor 1, isto é, de ocorrer um menino em uma família com seis crianças, calcule:

$$P(X = 1) = \binom{6}{1} \times \left(\frac{1}{2}\right)^1 \times \left(\frac{1}{2}\right)^5 = \frac{6}{64}$$

Para obter a probabilidade de x assumir o valor 2, isto é, de ocorrerem dois meninos em uma família com seis crianças, calcule:

$$P(X = 2) = \binom{6}{2} \times \left(\frac{1}{2}\right)^2 \times \left(\frac{1}{2}\right)^4 = \frac{15}{64}$$

Para obter a probabilidade de X assumir o valor 3, calcule:

$$P(X = 3) = \binom{6}{3} \times \left(\frac{1}{2}\right)^3 \times \left(\frac{1}{2}\right)^3 = \frac{20}{64}$$

Para obter a probabilidade de X assumir o valor 4, calcule:

$$P(X = 4) = \binom{6}{4} \times \left(\frac{1}{2}\right)^4 \times \left(\frac{1}{2}\right)^2 = \frac{15}{64}$$

Para obter a probabilidade de X assumir o valor 5, calcule:

$$P(X = 5) = \binom{6}{5} \times \left(\frac{1}{2}\right)^5 \times \left(\frac{1}{2}\right)^1 = \frac{6}{64}$$

Para obter a probabilidade de X assumir o valor 6, calcule:

$$P(X = 6) = \binom{6}{6} \times \left(\frac{1}{2}\right)^6 \times \left(\frac{1}{2}\right)^0 = \frac{1}{64}$$

Com os valores de X e as respectivas probabilidades, podemos construir a Tabela 9.10, que apresenta uma distribuição binomial para $n = 6$ e $p = 0,5$. O gráfico de barras está na Figura 9.4.

TABELA 9.10
Distribuição do número de meninos em uma família com seis crianças.

Evento	X	$P(X)$
Nenhum menino	0	1/64
1 menino	1	6/64
2 meninos	2	15/64
3 meninos	3	20/64
4 meninos	4	15/64
5 meninos	5	6/64
6 meninos	6	1/64

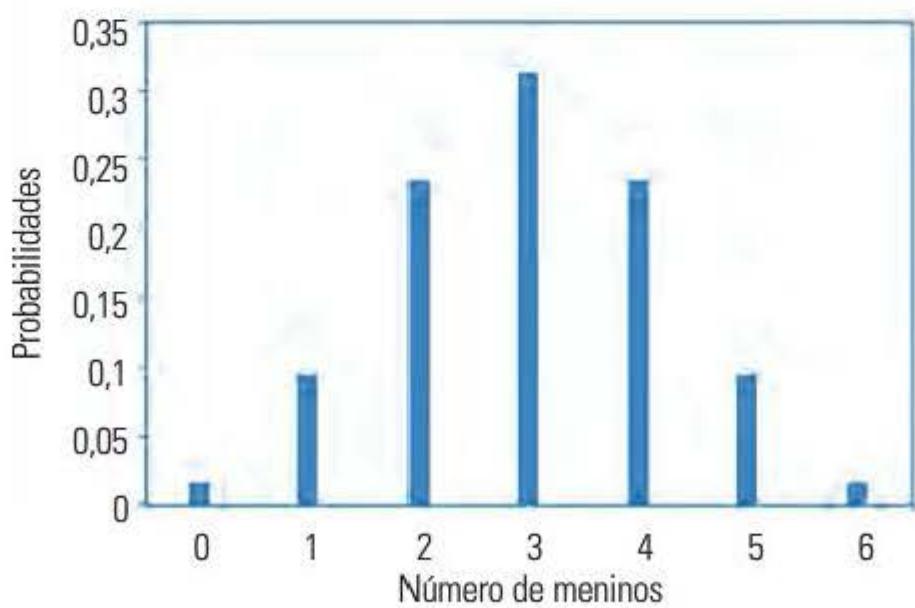


FIGURA 9.4 Distribuição do número de meninos em uma família com seis crianças.

9.5.5 – A probabilidade de um menino ser daltônico é 8%. Qual é a probabilidade de serem daltônicos todos os quatro meninos que se apresentaram, em determinado dia, para um exame oftalmológico?

No problema, $p = 0,08$. Então $q = 1 - 0,08 = 0,92$. O número de meninos é $n = 4$. Para obter a probabilidade de X assumir valor 4, aplica-se a fórmula:

$$P(X = x) = \binom{n}{x} p^x q^{(n-x)}$$

Então:

$$\begin{aligned} P(X = 4) &= \binom{4}{4} \times 0,8^4 \times 0,92^0 = \\ &= 0,00004096 \text{ ou } 0,004096\% \end{aligned}$$

9.5.6 – O resultado do cruzamento de ervilhas amarelas homozigotas (AA) com ervilhas verdes homozigotas (aa) são ervilhas amarelas heterozigotas (Aa). Se estas ervilhas forem cruzadas entre si, ocorrem ervilhas amarelas e verdes na proporção de 3 para 1. Portanto, a probabilidade de, num cruzamento desse tipo, ocorrer ervilha amarela é $p = 3/4$ e a probabilidade de ocorrer ervilha verde é $q = 1/4$. Logo, o número de ervilhas amarelas em um conjunto de n ervilhas é uma variável aleatória com distribuição binomial de parâmetros n e $p = 3/4$. Foram pegas, ao acaso, quatro ervilhas resultantes do cruzamento de ervilhas amarelas heterozigotas. Qual é a probabilidade de duas dessas quatro ervilhas serem de cor amarela?

A probabilidade de duas das quatro ervilhas serem amarelas é dada por:

$$\begin{aligned} P(X = 2) &= \binom{4}{2} \times \left(\frac{3}{4}\right)^2 \left(\frac{1}{4}\right)^2 \\ &= 0,2109 \text{ ou } 21,09\% \end{aligned}$$

9.5.7 – Considere novamente o cruzamento de ervilhas amarelas e verdes, descrito no Exercício 9.5.6. Qual é a média de ervilhas amarelas, considerando uma amostra de $n = 100$ ervilhas? Qual é a variância?

Um conjunto de $n = 100$ ervilhas tem, em média:

$$\mu = 100 \times \frac{3}{4} = 75 \text{ ervilhas amarelas}$$

e variância:

$$\sigma^2 = 100 \times \frac{3}{4} \times \frac{1}{4} = 18,75$$

9.5.8 – Um exame é constituído de 100 testes com cinco opções, onde apenas uma é correta. Um aluno que nada sabe sobre a matéria do exame acerta, em média, quantos testes? Qual é a variância da distribuição?

A probabilidade de um aluno acertar uma resposta por acaso é $p = 1/5$. Existem $n = 100$ testes. Então, aplicando a fórmula, vem:

$$\mu = 100 \times \frac{1}{5} = 20$$

ou seja, um aluno que nada sabe sobre a matéria acerta em média 20 testes. A variância da distribuição é:

$$\sigma^2 = 100 \times \frac{1}{5} \times \frac{4}{5} = 16$$

9.5.9 – Um pesquisador de mercado quer saber a proporção de consumidores que preferem café sem cafeína. Se ele perguntar a 500 pessoas que tipo de café adquiriu em sua última compra, como ele estimaria a média e a variância da distribuição?

O pesquisador terá respostas “Sim” e “Não”, além de outras, como “Não sei”, “Não me lembro”, “Não tenho tempo para responder questionários”. Se as respostas do tipo “Sim” e “Não” chegarem a 70%, isto é, se a taxa de resposta for de 70% (quando a quantidade de não-respondentes é grande, a pesquisa não tem validade), terá uma distribuição binomial. A média será obtida pela fórmula:

$$\mu = np$$

e a variância σ^2 pela fórmula:

$$\sigma^2 = npq$$

O valor de p é obtido dividindo o número de consumidores que prefere café sem cafeína pelo número n de respondentes.

9.5.10 – Numa cirurgia experimental, uma cobaia pode sobreviver (S) ou morrer (M). O pesquisador não sabe (é isto que ele está pesquisando), mas considere que a probabilidade de uma cobaia sobreviver na cirurgia é 0,25. A cirurgia será feita em duas cobaias. Se ambas sobreviverem, operam-se mais duas. Se só uma sobreviver, outra é operada. Se as duas morrerem, o pesquisador pára o experimento. Qual é a probabilidade de não se fazer uma segunda seqüência de cirurgias (as duas primeiras cobaias operadas morrerem)? Qual é a probabilidade de quatro cobaias ser operadas e as quatro sobreviverem?

As respostas são dadas na Tabela 9.11. Se as duas cobaias morrerem (sobrevida zero), o pesquisador pára o experimento. A probabilidade de isso ocorrer é 0,5625. Se as duas cobaias sobreviverem (sobrevida 2), o pesquisador opera mais duas. A probabilidade de isso ocorrer é:

$$0,0625 \times 0,0625 = 0,0039.$$

TABELA 9.11
Probabilidade de sobrevida de cobaias submetidas a uma cirurgia experimental.

1ª seqüência			2ª seqüência			Total		
Operadas	Vivas	P(vivas)	Operadas	Vivas	P(vivas)	Operadas	Vivas	P(vivas)
2	0	0,5625	0			2	0	0,5625
2	1	0,3750	1	0	0,7500	3	1	0,2813
				1	0,2500		2	0,0938
2	2	0,0625	2	0	0,5625	4	2	0,0352
				1	0,3750		3	0,0234
				2	0,0625		4	0,0039

9.6 – EXERCÍCIOS PROPOSTOS

9.6.1 – Há três bolas numeradas em uma caixa, cada uma com um número diferente. Os números são 1, 2 e 3. Tira-se uma bola da caixa e, em seguida, outra. Forma-se, então, um número de dois dígitos com os números das bolas retiradas. Por exemplo, se saiu 3 e depois 2, foi formado o número 32. Um jogador ganha se sair número par. Nesse jogo, se ganha mais do que se perde ou é justamente o contrário?

9.6.2 – Seja X a variável aleatória que indica o número de meninos em uma família com cinco crianças. Apresente a distribuição de X em uma tabela. Faça um gráfico.

9.6.3 – Um exame é constituído de 10 testes tipo certo-errado. Um aluno que nada sabe sobre a matéria do exame, quantos testes, em média, acerta? Qual é a variância da distribuição?

- 9.6.4 – Um exame é constituído de 10 testes com cinco opções, das quais apenas uma é correta. Um aluno que nada sabe sobre a matéria do exame acerta, em média, quantos testes? Qual é a variância da distribuição?
- 9.6.5 – Suponha que determinado medicamento usado para o diagnóstico precoce da gravidez é capaz de confirmar casos positivos em apenas 90% das gestantes muito jovens. Isto porque, em 10% das gestantes muito jovens, ocorre uma escamação do epitélio do útero, que é confundida com a menstruação. Nestas condições, qual é a probabilidade de duas, de três gestantes muito jovens que fizeram uso desse medicamento, não terem confirmado precocemente a gravidez?
- 9.6.6 – A probabilidade de um casal heterozigoto para o gene da fenilcetonúria ($Aa \times Aa$) ter um filho afetado (aa) é $1/4$. Se o casal tiver três filhos, qual é a probabilidade de ter um filho com a doença?
- 9.6.7 – A probabilidade de um indivíduo ter sangue Rh^- é 10%, na população brasileira toda. Qual é a possibilidade de se apresentarem, em determinado dia em um banco de sangue, cinco doadores de sangue, todos Rh^- ?
- 9.6.8 – Foi feito um levantamento da opinião de 1.000 enfermeiras que trabalhavam em determinado hospital sobre determinada questão que tinha duas alternativas: “Sim” e “Não”. As respostas têm distribuição binomial? Algumas enfermeiras não responderam ao questionário. Que efeito isso pode ter sobre as respostas?
- 9.6.9 – A experiência demonstra que um detector de mentiras dá resposta positiva (indicando mentira) 10% das vezes em que uma pessoa está dizendo a verdade e 95% das vezes em que a pessoa está mentindo. Imagine que seis suspeitos de um crime são submetidos ao detector de mentiras. Todos os suspeitos se dizem inocentes e estão dizendo a verdade. Qual é a probabilidade de ocorrer uma resposta positiva?
- 9.6.10 – O diretor de uma grande empresa está preocupado com a questão de acidentes e quer fazer um levantamento da situação. Existem os registros do número de acidentes por dia na empresa. Essa variável tem distribuição binomial?

(página deixada intencionalmente em branco)

Distribuição
Normal

10

(página deixada intencionalmente em branco)

No Capítulo 3 deste livro você aprendeu a apresentar dados contínuos em histogramas ou em polígonos de freqüências. Esses gráficos mostram a configuração de *distribuições empíricas*, isto é, de distribuições obtidas com base em dados observados. Veja o Exemplo 10.1.

Exemplo 10.1: Uma distribuição empírica.

Um matemático belga do século XIX pôs na cabeça a idéia de descrever o "homem médio" e, por conta disso, mediu muitas e muitas variáveis¹. A Tabela 10.1 mostra a distribuição do perímetro torácico² que esse matemático mediu em nada menos do que 5.732 soldados escoceses. As medidas estão em polegadas. Como uma polegada vale 2,54 cm, você vê que as medidas variaram entre 83,82 cm e 121,92 cm.³ Veja o histograma apresentado na Figura 10.1.

TABELA 10.1
Distribuição de freqüências para perímetro torácico de homens adultos, em polegadas.

Perímetro torácico	Freqüência	Freqüência relativa
33	3	0,00052
34	19	0,00331
35	81	0,01413
36	189	0,03297
37	409	0,07135
38	753	0,13137
39	1062	0,18528
40	1082	0,18876
41	935	0,16312
42	646	0,11270
43	313	0,05461
44	168	0,02931
45	50	0,00872
46	18	0,00314
47	3	0,00052
48	1	0,00017

Fonte: Daly, F.; Hand, D; Jones, C; Lunn, AD (1995).

¹Adolphe Quetelet. 1796-1874.

²DALY, F.; HAND, D; JONES, C; LUNN, AD: **Elements of Statistics**: Addison Wesley, 1995.

³Os homens eram, em média, menores do que são hoje.

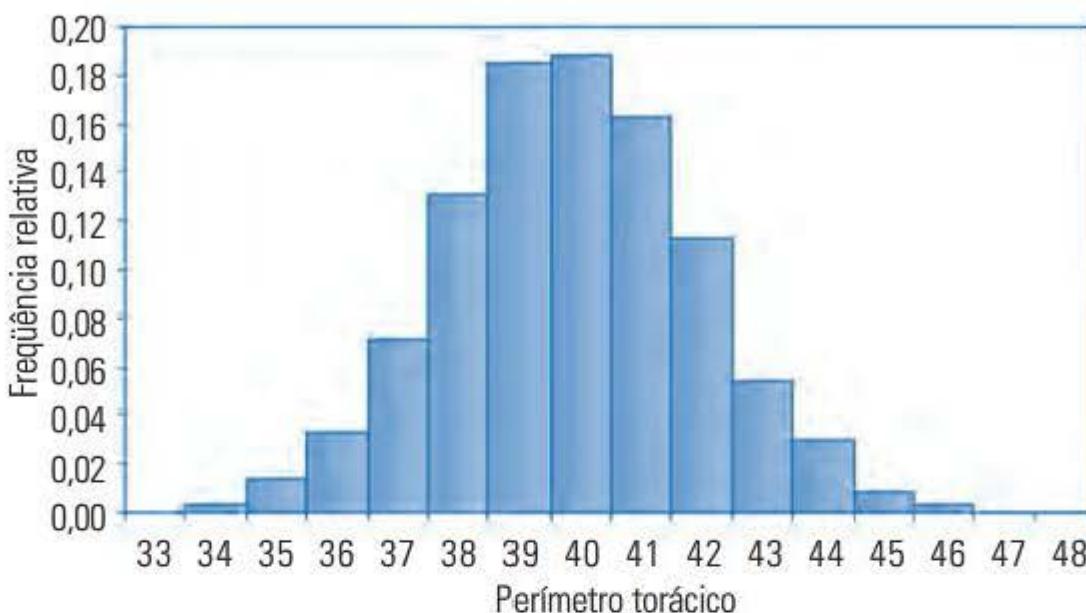


FIGURA 10.1 Histograma para a distribuição de freqüências do perímetro torácico de homens adultos, em polegadas.

Muitas distribuições de freqüências têm a aparência da distribuição da Figura 10.1. Todas elas se aproximam de uma *distribuição teórica* chamada *distribuição normal* (também conhecida como distribuição de Gauss), apresentada em gráfico na Figura 10.2. Nenhuma distribuição empírica, no entanto, tem todas as características da distribuição normal. Mas o fato de pressupor que uma variável tem distribuição normal permite resolver muitos problemas em Estatística.

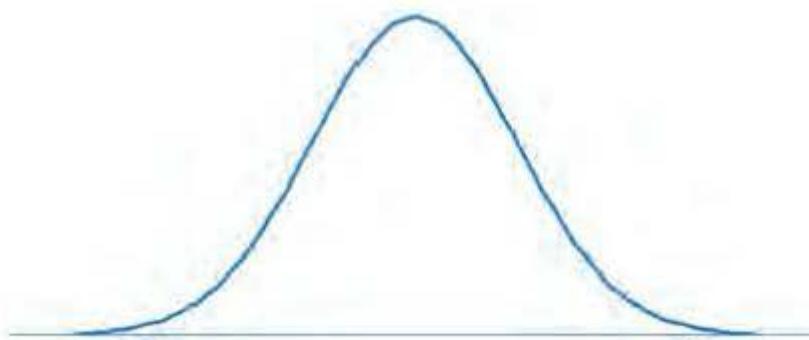


FIGURA 10.2 Gráfico da distribuição normal.

10.1 – CARACTERÍSTICAS DA DISTRIBUIÇÃO NORMAL

Os gráficos apresentados nas Figuras 10.1 e 10.2 têm configuração semelhante. Mas o primeiro é empírico e o segundo é teórico, o que os tornam diferentes. Observe novamente o histograma da Figura 10.1: a freqüência relativa de unidades em cada intervalo é dada pela altura (medida no eixo das ordenadas) do retângulo que representa o intervalo. Então, a proporção de homens adultos com perímetro torácico igual a 37 polegadas, por exemplo, é dada no eixo das ordenadas (aproximadamente 0,07). Essas proporções são *estimativas de probabilidade*.

A distribuição teórica dada na Figura 10.2 representa uma *população infinita*. Logo, o eixo das ordenadas *não* mostra a proporção de indivíduos em cada categoria porque não há como calcular proporções sobre um total que é infinito. Mas a curva abriga toda a população em estudo. Então a área total sob a curva é 1, ou seja, 100%, porque toda a população está sob a curva.

A distribuição normal fica definida quando são dados *dois parâmetros*: a média, que se representa pela letra grega μ (lê-se: mi) e o desvio padrão, que se representa pela letra grega σ (lê-se: sigma).

Algumas características da distribuição normal são bem conhecidas:

- a média, a mediana e a moda coincidem e estão no centro da distribuição;
- o gráfico da distribuição normal tem aspecto típico: é uma curva em forma de sino, simétrica em torno da média;
- como a curva é simétrica em torno da média, 50% dos valores são iguais ou maiores do que a média e 50% dos valores são iguais ou menores do que a média.

Exemplo 10.2: Uma distribuição normal.

Um teste de inteligência⁴ foi idealizado pressupondo que quociente de inteligência tem distribuição normal de média $\mu = 100$ e desvio padrão $\sigma = 15$. Veja a Figura 10.3 e note que, de acordo com esse teste:

- As pessoas têm, em média, QI igual a 100.
- Metade das pessoas tem QI igual ou maior do que 100 e metade tem QI igual ou menor do que 100.
- Pessoas com QI muito alto (na cauda à direita da curva) são raras, como também são raras pessoas com QI muito baixo (na cauda à esquerda da curva).

⁴Existem muitas maneiras de “medir” inteligência (embora nenhuma delas explique, exatamente, o que está sendo medido). Mas um dos testes (Weschler) foi idealizado pressupondo que inteligência tem distribuição normal, como mostrado no exemplo. In: MOTULSKY, H. **Intuitive Biostatistics**. Nova York, Oxford Press, 1995. p.38.

A grande vantagem de pressupor que uma variável tem distribuição normal é o fato de ser possível — porque a distribuição é conhecida — calcular as probabilidades relacionadas a essa variável. Essas probabilidades são dadas pelas *áreas sob a curva*. Mas como isso é feito? Você já sabe a relação entre a *área sob a curva* e a média: metade das observações é maior do que a média (e, obviamente, metade das observações é menor do que a média). Mas também existem relações entre a *área sob a curva* e o *desvio padrão* da variável. Veja:

- Prova-se, teoricamente⁵, que se a variável tem distribuição normal, 34,13% da área sob a curva estão entre a média (μ) e um ponto de abscissa igual à média mais um desvio padrão ($\mu + \sigma$).
- A curva é simétrica em torno da média. Segue-se daí que 34,13% da área sob a curva está entre a média (μ) e um ponto de abscissa igual à média menos um desvio padrão ($\mu - \sigma$).
- Se você somar as porcentagens, terá 68,26%. Então, entre ($\mu - \sigma$) e ($\mu + \sigma$) estão 68,26% da área da curva, como mostra a Figura 10.3.

A proporção da área sob a curva dá a probabilidade de ocorrerem casos no mesmo intervalo. Veja o Exemplo 10.3.

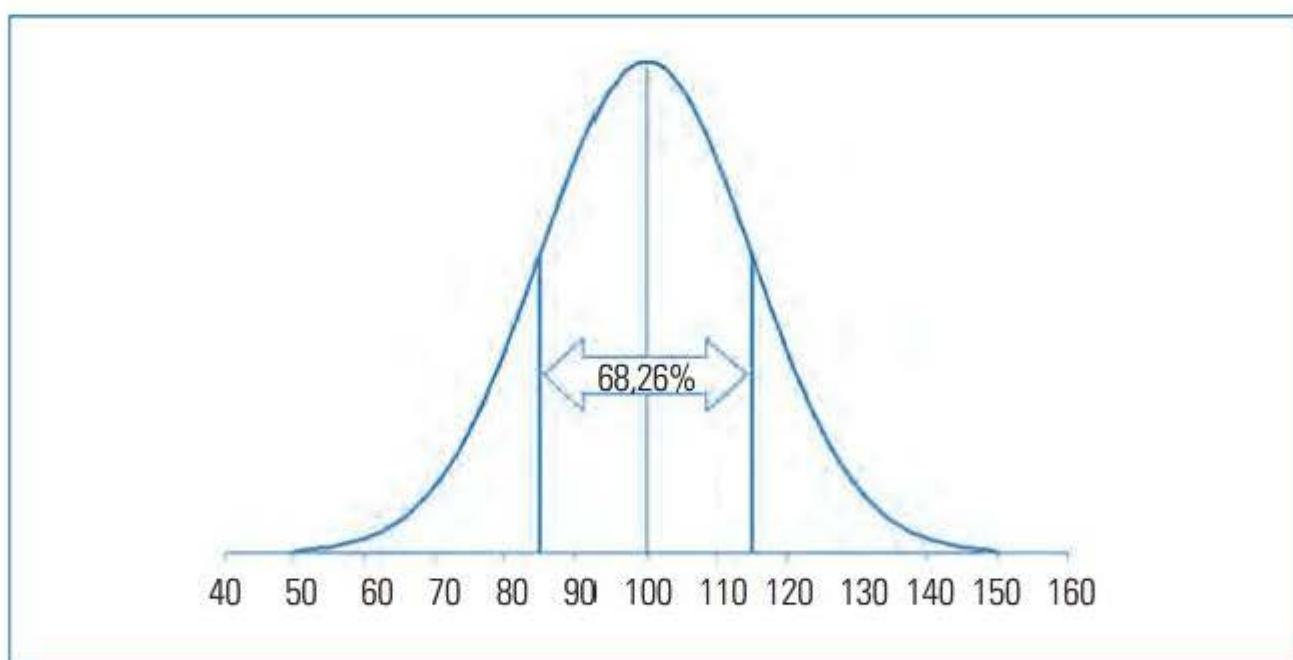


FIGURA 10.3 Distribuição normal: 68,26% dos casos estão entre a média \pm 1 desvio padrão.

⁵Neste livro você aprende como usar as tabelas prontas. A teoria é encontrada em textos teóricos de Estatística.

Exemplo 10.3: Média \pm desvio padrão.

Reveja o Exemplo 10.2. Pressupondo que quociente de inteligência tem distribuição normal de média $\mu = 100$ e desvio padrão $\sigma = 15$, então:

- 34,13% das pessoas, segundo o teste, têm quociente de inteligência entre $\mu = 100$ e $\mu + \sigma = 100 + 15 = 115$, ou seja, entre 100 e 115.
- 34,13% das pessoas, segundo o teste, têm quociente de inteligência entre $\mu = 100$ e $\mu - \sigma = 100 - 15 = 85$, ou seja, entre 100 e 85
- 68,26% das pessoas, segundo o teste, têm quociente de inteligência entre 85 e 115.

Olhe novamente a Figura 10.2: as áreas sob a curva diminuem à medida que os valores de X se afastam da média. Prova-se teoricamente que se a variável tem distribuição normal:

- 13,59% da área sob a curva estão entre a média mais um desvio padrão ($\mu + \sigma$) e um ponto de abscissa igual à média mais dois desvios padrões ($\mu + 2\sigma$).
- A curva é simétrica em torno da média. Segue-se daí que 13,59% da área sob a curva estão entre a média menos um desvio padrão ($\mu - \sigma$) e um ponto de abscissa igual à média menos dois desvios padrões ($\mu - 2\sigma$). Veja a Figura 10.4.

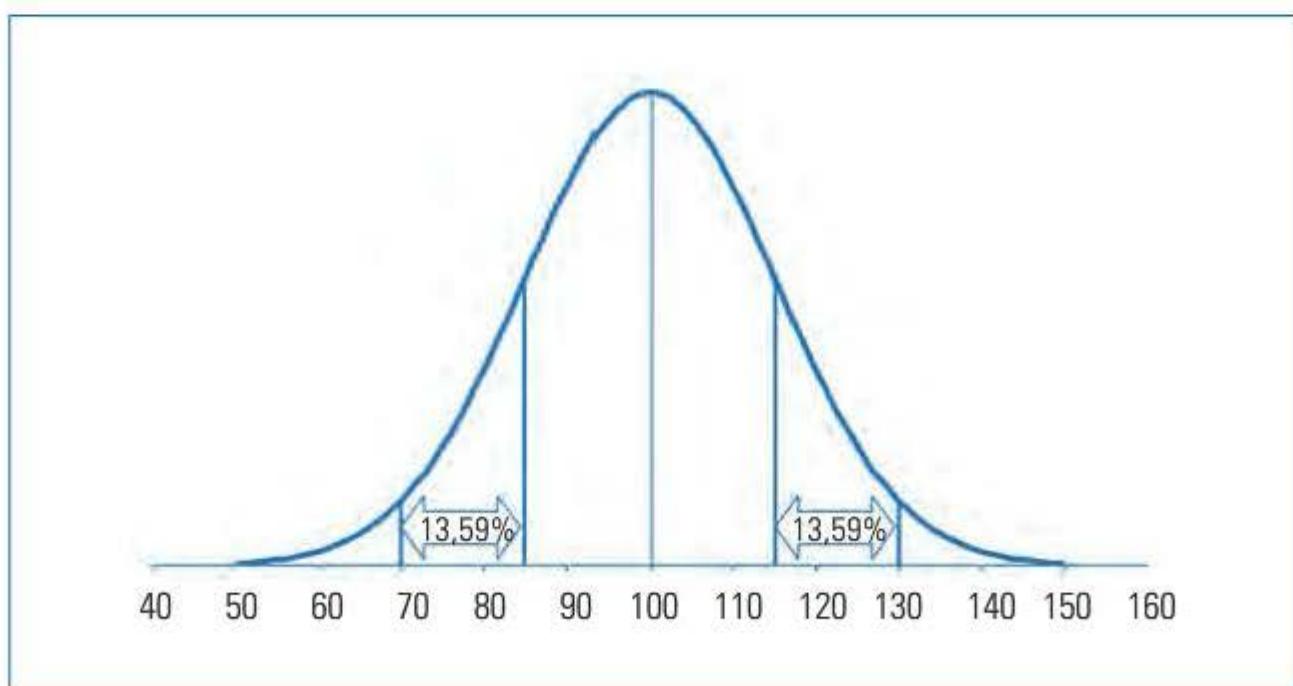


FIGURA 10.4 Distribuição normal: 13,59% dos casos entre $\mu + \sigma$ e $\mu + 2\sigma$ e 13,59% dos casos entre $\mu - \sigma$ e $\mu - 2\sigma$.

Exemplo 10.4: Outros dois intervalos.

Reveja o Exemplo 10.2. Pressupondo que quociente de inteligência tem distribuição normal de média $\mu = 100$ e desvio padrão $\sigma = 15$, então:

- 13,59% das pessoas, segundo o teste, têm quociente de inteligência entre $\mu + \sigma = 100 + 15 = 115$ e $\mu + 2\sigma = 100 + 30 = 130$, ou seja, entre 115 e 130.
- 13,59% das pessoas, segundo o teste, têm quociente de inteligência entre $\mu - \sigma = 100 - 15 = 85$ e $\mu - 2\sigma = 100 - 30 = 70$, ou seja, entre 70 e 85.

Vamos, agora, reunir as informações das duas últimas figuras. Isso significa calcular a probabilidade de uma observação cair no intervalo $\mu \pm 2\sigma$ ou — o que é o mesmo — estar entre $(\mu - 2\sigma)$ e $(\mu + 2\sigma)$. Escrevemos:

$$P[(\mu - 2\sigma) \leq X \leq (\mu + 2\sigma)]$$

Lembrando os valores apresentados nas figuras 10.3 e 10.4, podemos escrever:

$$P[(\mu - 2\sigma) \leq X \leq (\mu + 2\sigma)] = 13,59\% + 34,13\% + 34,13\% + 13,59\% = 95,44\%$$

Logo, o intervalo $\mu \pm 2\sigma$ engloba 95,44% da área sob a curva.

Exemplo 10.5: Média ± 2 desvios padrões.

Reveja o Exemplo 10.2. Pressupondo que quociente de inteligência tem distribuição normal de média $\mu = 100$ e desvio padrão $\sigma = 15$, então: 95,44% das pessoas, segundo o teste, têm quociente de inteligência entre 70 e 130, isto é, entre

$$\mu - 2\sigma = (100 - 2 \times 15) = 70$$

e

$$\mu + 2\sigma = (100 + 2 \times 15) = 130.$$

Agora, olhe novamente a Figura 10.4: a área sob a curva depois do ponto de abscissa $(\mu + 2\sigma)$ é muito pequena. Do que foi visto, é fácil entender que essa área tem probabilidade:

$$50,0\% - 34,13\% - 13,59\% = 2,28\%.$$

Por similaridade, a área sob a curva antes do ponto de abscissa $(\mu - 2\sigma)$ tem, como se vê na Figura 10.4, probabilidade:

$$50,0\% - 34,13\% - 13,59\% = 2,28\%.$$

Exemplo 10.6: Caudas da distribuição.

Reveja o Exemplo 10.2. Pressupondo que quociente de inteligência tem distribuição normal de média $\mu = 100$ e desvio padrão $\sigma = 15$, qual é o valor da abscissa (QI) que delimita os 2,28% de QI mais alto? E qual é o valor da abscissa (QI) que delimita os 2,28% de QI mais baixo?

Solução

Os 2,28% das pessoas com QI mais alto são os que estão acima de $(\mu + 2\sigma) = (100 + 2 \times 15) = 130$.

Os 2,28% das pessoas com QI mais baixo são os que estão abaixo de $(\mu - 2\sigma) = (100 - 2 \times 15) = 70$.

Portanto, pessoas com QI muito alto (na cauda à direita da curva) são raras, como também são raras pessoas com QI muito baixo (na cauda à esquerda da curva)

É importante lembrar que, no exemplo dado, os valores obtidos pressupõem *distribuição normal*. Na prática, encontramos distribuições *aproximadamente normais*. Então, os resultados obtidos são *aproximações*. De qualquer forma, na maioria das vezes, o intervalo $\bar{x} \pm s$ captura a maioria dos casos e o intervalo $\bar{x} \pm 2s$ engloba a grande maioria de casos.

*10.2. – DISTRIBUIÇÃO NORMAL REDUZIDA

Distribuição normal reduzida ou distribuição normal padronizada é a distribuição normal de média zero e variância 1.

A variável que tem distribuição normal reduzida ou distribuição normal padronizada é chamada *variável reduzida ou padronizada* e é indicada pela letra *z*.

A distribuição normal reduzida tem grande importância:

1. As *probabilidades associadas à distribuição normal reduzida* são *dadas em tabelas*, o que torna fácil saber as probabilidades associadas a essa distribuição. Basta procurar na tabela.
2. Podemos *transformar* qualquer variável aleatória *X* com distribuição normal de média e desvio padrão conhecidos numa distribuição normal reduzida.

3. Dos itens 1 e 2 segue-se que qualquer probabilidade associada a X pode ser obtida transformando X (distribuição normal) em z (distribuição normal reduzida).

Vamos aprender como se acham as probabilidades na distribuição normal reduzida. Por exemplo, qual é a probabilidade de ocorrer valor entre a média, zero, e o valor $z = 1,25$? Essa probabilidade é encontrada na *tabela de distribuição normal reduzida* que você acha neste livro, em Apêndice. Mas parte dessa tabela foi reproduzida neste Capítulo: é a Tabela 10.2.

Para aprender como se usa a tabela de distribuição normal reduzida, observe a Figura 10.5. A probabilidade de ocorrer valor entre a média zero e o valor $z = 1,25$ corresponde à área sombreada na Figura 10.5.

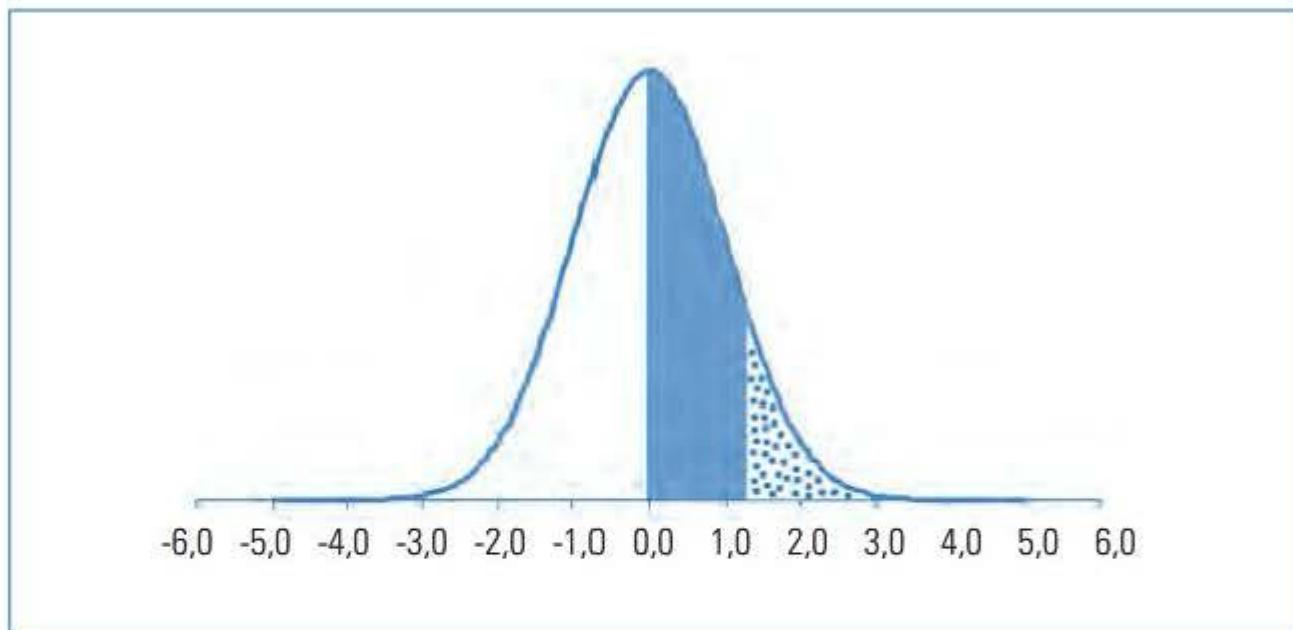


FIGURA 10.5 Probabilidade de ocorrer valor entre zero e $z = 1,25$.

Agora, olhe a Tabela 10.2: na primeira *coluna* está o valor 1,2 (negrito); na primeira *linha* da Tabela 10.2 está o valor 5 (negrito). O número 1,2 compõe, com o algarismo 5, o número $z = 1,25$. No *cruzamento* da linha 1,2 com a coluna 5 está o número 0,3944 (negrito). Esta é a probabilidade de ocorrer valor entre a média zero e o valor $z = 1,25$ (área sombreada na Figura 10.5).

TABELA 10.2
Tabela (parcial) de distribuição normal reduzida: probabilidade de valor entre zero e 1,25.

	0	1	2	3	4	5	6
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636
0,2	0,0793	0,0832	0,0871	0,0910	0,0946	0,0987	0,1026
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454
0,7	0,2580	0,2611	0,2642	0,2673	0,2703	0,2734	0,2764
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279

Exemplo 10.7: Probabilidade na distribuição normal reduzida.

Qual é a probabilidade de ocorrer valor maior do que $z = 1,25$?

Solução

A probabilidade de ocorrer valor entre a média zero e o valor $z = 1,25$ (área sombreada) é 0,3944, como foi visto anteriormente. Essa probabilidade corresponde à área pontilhada na Figura 10.5. A probabilidade de ocorrer valor maior do que a média zero é 0,5. Então a probabilidade pedida (área com hachuras) é:

$$0,5 - 0,3944 = 0,1056 \text{ ou } 10,56\%$$

Exemplo 10.8: Probabilidade na distribuição normal reduzida.

Qual é a probabilidade de ocorrer valor menor do que $z = -0,75$?

A probabilidade de ocorrer valor menor do que $z = -0,75$ é dada pela área com hachuras na Figura 10.6. Observe: a área pontilhada entre zero e $z = -0,75$ é igual à área sombreada entre zero e $z = 0,75$.

Para achar essa área procure na primeira coluna da tabela de distribuição normal reduzida o número 0,7 e, na primeira linha, o número 5. Você compõe o número $z = 0,75$. No cruzamento entre a coluna (0,7) e a linha (5) você lê 0,2734, que é a probabilidade de ocorrer valor entre zero e $z = 0,75$ (área pontilhada).

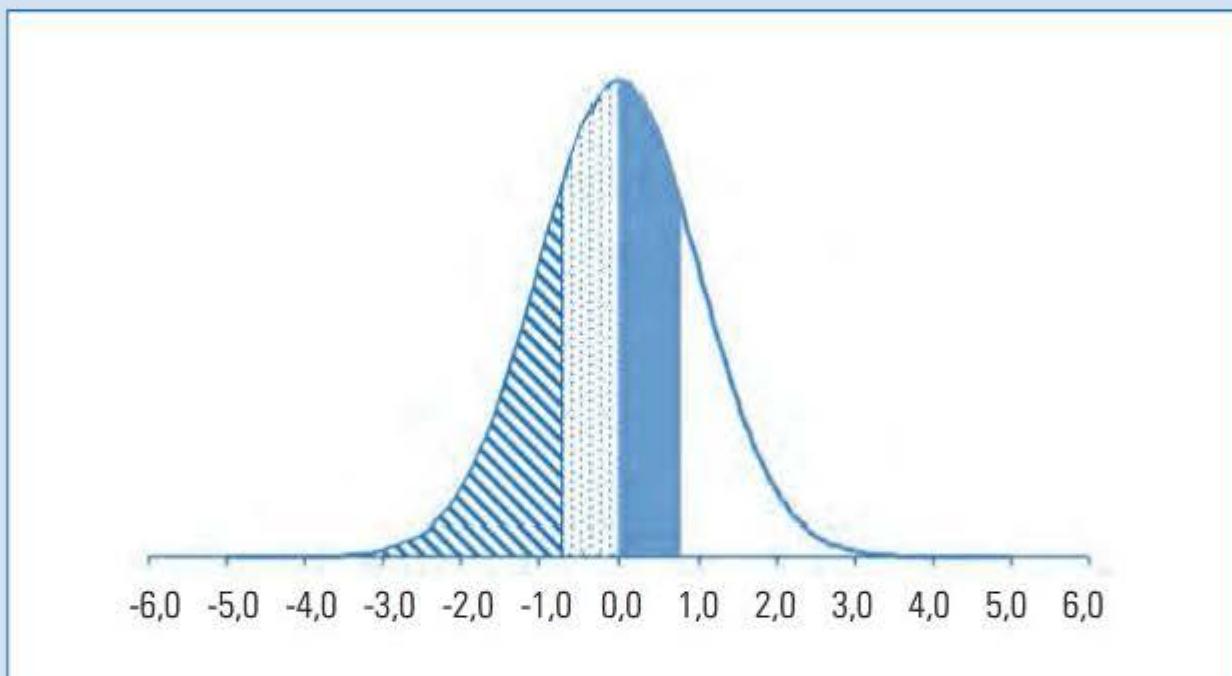


FIGURA 10.6 Probabilidade de ocorrer valor menor do que $z = -0,75$.

A probabilidade de ocorrer valor menor do que $z = -0,75$ (área com hachuras) é igual à probabilidade de ocorrer valor maior do que $z = 0,75$ (área em branco). Como a probabilidade de ocorrer valor maior do que a média zero é 0,5, a probabilidade pedida é dada por: $0,5 - 0,2734 = 0,2266$ ou 22,66%.

*10.3 – PROBABILIDADES NA DISTRIBUIÇÃO NORMAL

Você aprendeu a trabalhar com a distribuição normal reduzida. Aprenda, agora, como trabalhar com a distribuição normal.

Mas como se transforma uma variável que tem distribuição normal com média μ e desvio padrão σ , em uma variável com distribuição normal reduzida de média zero e desvio padrão 1? Basta calcular:

$$z = \frac{X - \mu}{\sigma}$$

Com o valor de z , calculado pela fórmula dada, você procura a probabilidade pedida na tabela de distribuição normal reduzida, como mostra a Seção 10.2 deste Capítulo.

Exemplo 10.9: Probabilidade na distribuição normal.

A quantidade de colesterol em 100 ml de plasma sanguíneo humano tem distribuição normal com média 200 mg e desvio padrão 20 mg. Qual é a probabilidade de uma pessoa apresentar entre 200 e 225 mg de colesterol por 100 ml de plasma?

Solução

Observe a Figura 10.7. A probabilidade pedida corresponde à área sombreada. Como você acha o valor dessa área?

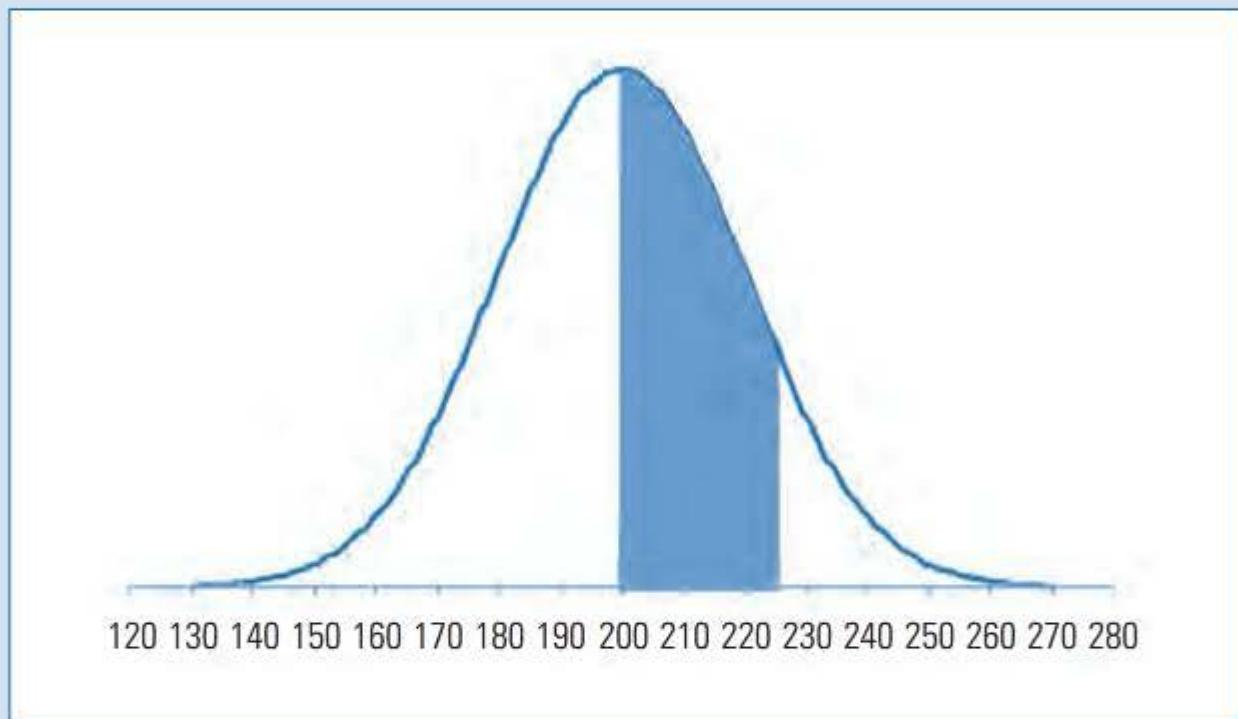


FIGURA 10.7 Probabilidade de taxa de colesterol entre 200 e 225 mg por 100 ml de sangue.

Para obter a probabilidade pedida, é preciso transformar a distribuição normal em *distribuição normal reduzida*.

Na distribuição normal reduzida, a média é zero. Para obter $X = 225$ na distribuição normal reduzida, calcule:

$$z = \frac{X - \mu}{\sigma}$$

$$\frac{225 - 200}{20}$$

$$= 1,25$$

A área sombreada na Figura 10.7 corresponde à área sombreada na Figura 10.5. Então a probabilidade de X assumir valor entre 200 e 225 é igual à probabilidade de Z assumir valor entre zero e $z = 1,25$ que, como se viu na Seção 10.2, é 0,3944 ou 39,44%. Logo, a probabilidade de uma pessoa apresentar taxa de colesterol entre 200 e 225 mg por 100 ml de plasma é 0,3944 ou 39,44%.

Exemplo 10.10: Probabilidade na distribuição normal.

A quantidade de colesterol em 100 ml de plasma sanguíneo humano tem distribuição normal com média 200 mg e desvio padrão 20 mg. Qual é a probabilidade de uma pessoa apresentar menos do que 195 mg de colesterol por 100 ml de plasma?

Solução

Essa probabilidade corresponde à área com hachuras na Figura 10.8.

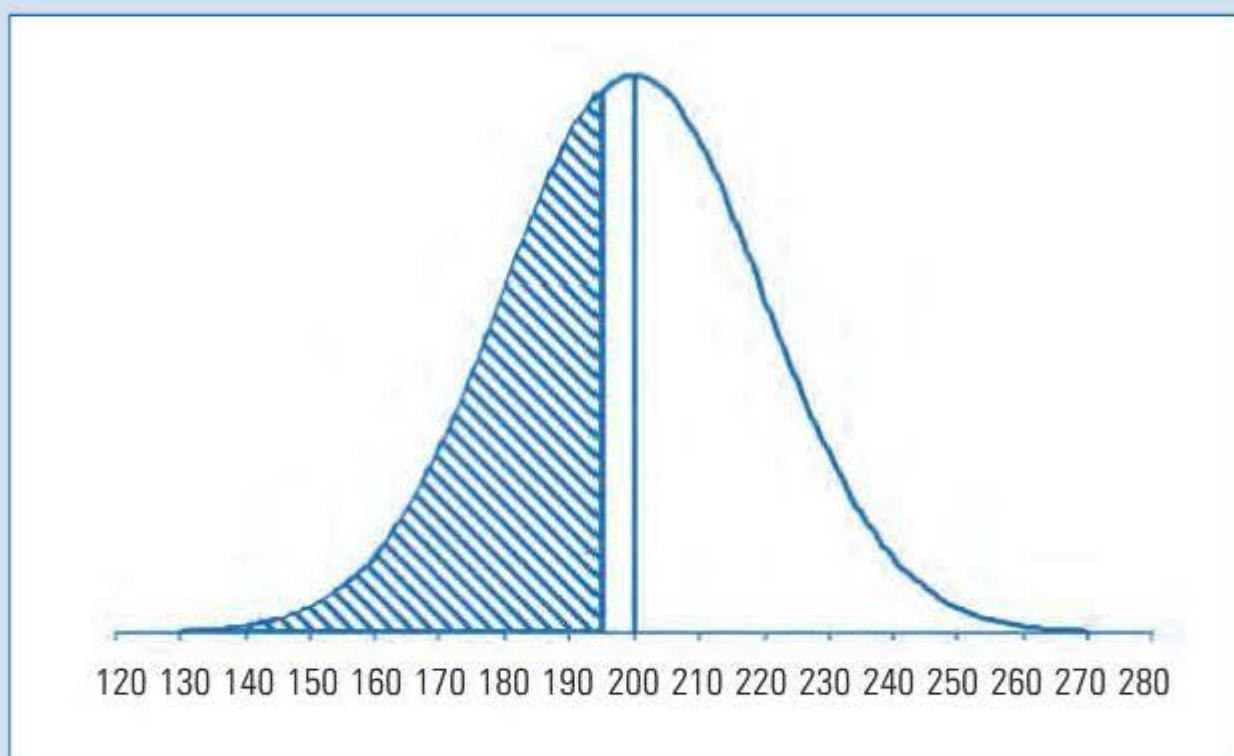


FIGURA 10.8 Probabilidade de taxa de colesterol menor do que 195 mg por 100 ml de sangue.

É preciso transformar o valor $X = 195$ em z . Obtém-se então:

$$z = \frac{195 - 200}{20} = -0,25$$

A probabilidade de ocorrerem valores de z iguais ou menores do que $-0,25$ é igual à probabilidade de valores de z iguais ou maiores do que $0,25$.

A probabilidade de ocorrerem valores de z entre a média zero e 0,25 você encontra na tabela de distribuição normal reduzida: é 0,0987 (no cruzamento da coluna 0,2 e da linha 5).

A probabilidade de valores de z iguais ou maiores do que 0,25 é, portanto: $0,5 - 0,0987 = 0,4013$ ou 40,13%

Então a probabilidade de uma pessoa apresentar 195 mg de colesterol por 100 ml de plasma ou menos é 0,4013 ou 40,13%.

10.4 – USOS DA DISTRIBUIÇÃO NORMAL

Imagine que você está lendo um artigo da área de Cardiologia. Nesse artigo você lê que a amostra de 100 pacientes forneceu, para pressão sistólica, a média $\bar{x} = 123,4$ mm de mercúrio e desvio padrão $s = 14,0$ mm de mercúrio. Esses valores *estimam* os parâmetros, isto é, a média μ e o desvio padrão σ da população de onde essa amostra provém. Por que essa informação é útil?

Primeiro, é razoável assumir que a pressão sistólica tem distribuição normal. Veja o gráfico da Figura 10.9. Depois, você já aprendeu que:

- A probabilidade de ocorrer valor de X no intervalo $\mu \pm \sigma$ é 0,6826 (34,13+0,3413)
- A probabilidade de ocorrer valor de X no intervalo $\mu \pm 2\sigma$ é 0,9544 (0,4772+0,4772).

No caso da amostra em discussão, temos que:

$$\bar{x} - s = 123,4 - 14,0 = 109,4$$

$$\bar{x} + s = 123,4 + 14,0 = 137,4$$

$$\bar{x} - 2s = 123,4 - 2 \times 14,0 = 95,4$$

$$\bar{x} + 2s = 123,4 + 2 \times 14,0 = 151,4$$

Considerando a média e o desvio padrão obtidos da amostra como boas estimativas de μ e σ , respectivamente, vem que:

- A probabilidade de encontrar pessoas na população de onde a amostra provém com pressão sistólica entre 109,4 e 137,4 mm de mercúrio é, aproximadamente (porque a distribuição é aproximadamente normal e os parâmetros estão estimados), 68,26%. Ou seja, cerca de 2/3 da população estudada deve ter pressão sistólica entre 109,4 e 137,4 mm de mercúrio.

- A probabilidade de encontrar pessoas na população de onde a amostra provém com pressão sistólica entre 95,4 e 151,4 mm de mercúrio é, aproximadamente (porque a distribuição é aproximadamente normal e os parâmetros estão estimados), 95,44%. Ou seja, a grande maioria da população estudada deve ter pressão sistólica entre 95,4 e 151,4 mm de mercúrio.

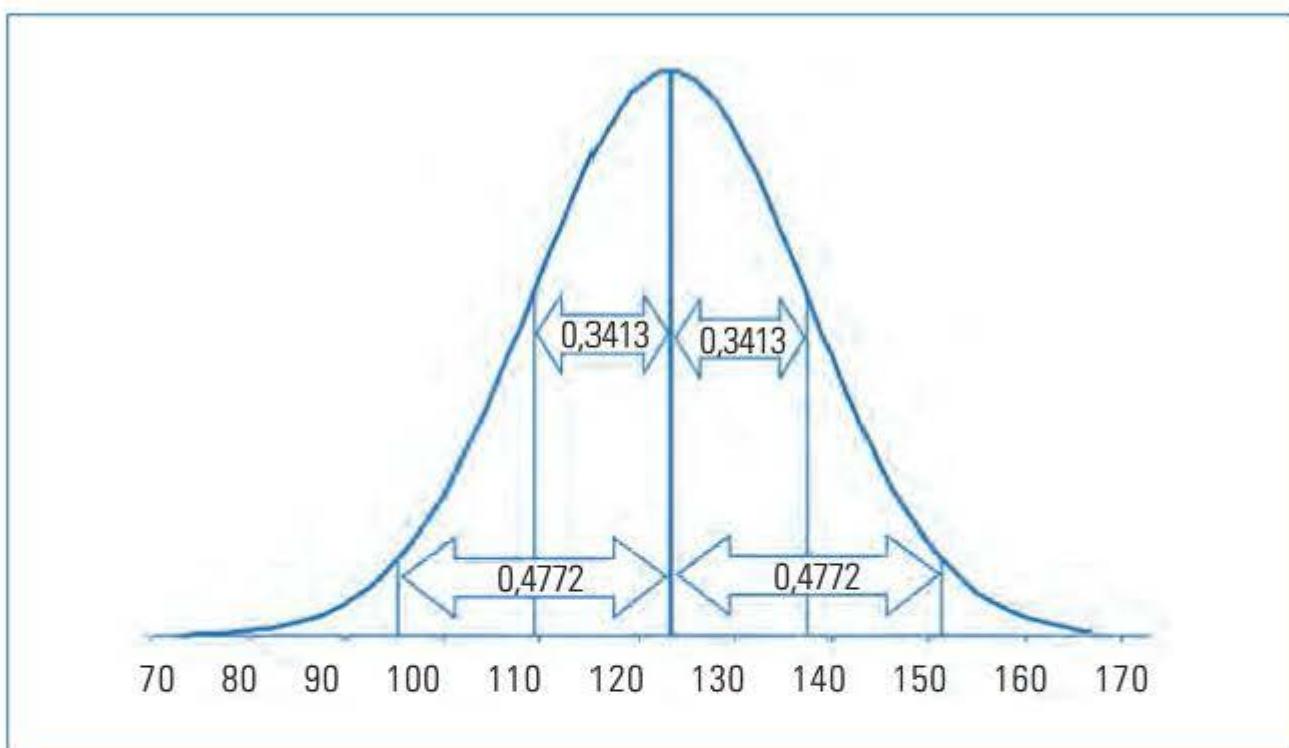


FIGURA 10.9 Distribuição da pressão sistólica.

A distribuição normal tem, ainda, outro uso importante em Estatística. Você já sabe que amostras tomadas ao acaso da mesma população são diferentes. Logo, as médias dessas amostras são diferentes. Pense no exemplo que acabamos de examinar. Foi medida a pressão sistólica de uma amostra de 100 pessoas, tomadas ao acaso da mesma população. A média calculada foi 123,4 mm de mercúrio. Se fossem obtidas outras 50 amostras dessa mesma população, as médias de pressão sistólica variariam e teriam uma distribuição. Mas qual seria essa distribuição?

Qualquer que seja a distribuição dos dados, as médias terão distribuição normal, de acordo com um teorema da Estatística (o teorema do limite central). Como consequência, se tomarmos amostras de centenas de observações, podemos ignorar a distribuição dos dados. A grande aplicação desta informação — o intervalo de confiança para uma média — será vista no Capítulo 11.

Mas o uso da distribuição normal vai mais além. Em exames radiológicos e laboratoriais, o uso da distribuição normal é comum. Veja como isto

é feito. Com base em grandes amostras, estimam-se μ e σ^2 . Depois, com base na distribuição normal, definem-se critérios de normalidade e não-normalidade. Por exemplo, para densidade mineral óssea (BMD, porque em inglês é: *bone mineral density*), que é medida em gramas por centímetro ao quadrado, a Organização Mundial de Saúde considera:

- Normal: qualquer valor mais alto que $\mu - \sigma$.
- Osteopenia ou osteoporose pré-clínica: valores entre $\mu - \sigma$ e $\mu - 2,5\sigma$.
- Osteoporose: valores abaixo de $\mu - 2,5\sigma$.

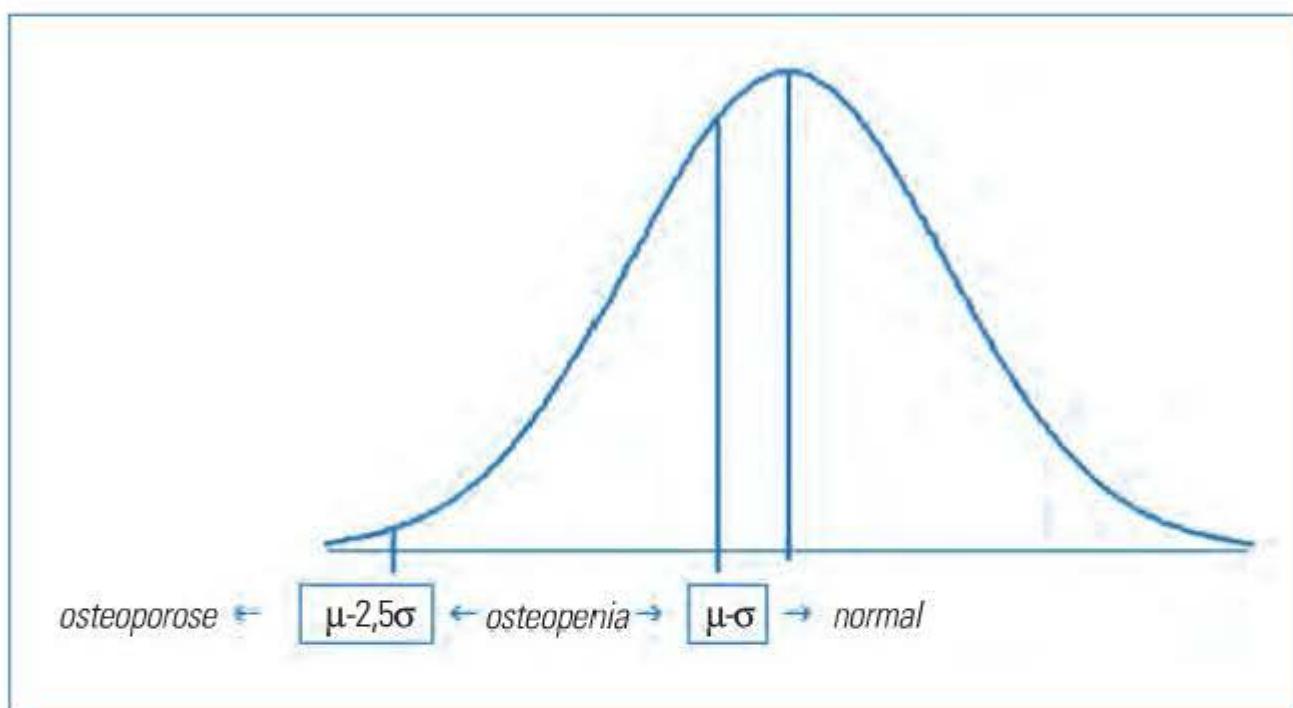


Figura 10.10 Distribuição de BMD.

Então, se for aceito que, para coluna lombar, o BMD médio é 1,061 com desvio padrão 1,0, a pessoa que tiver $BMD = 0,060$ é diagnosticada como tendo osteopenia.

10.5 – EXERCÍCIOS RESOLVIDOS

10.5.1 – Em uma distribuição normal reduzida, que proporção de casos cai: a) fora dos limites $z = 1$ e $z = -1$? b) fora dos limites $z = 1,96$ e $z = -1,96$?

a) A probabilidade de ocorrer valor maior do que a média zero é 0,5. A tabela de distribuição normal reduzida mostra que a probabilidade de ocorrer valor entre a média zero e $z = 1$ (procure $z = 1$ na tabela) é 0,3413. Então a probabilidade de ocorrer valor maior do que $z = 1$ é:

$$0,5000 - 0,3413 = 0,1587$$

Como a curva é simétrica, a probabilidade de ocorrer valor fora dos limites $z = 1$ e $z = -1$ é:

$$2 \times 0,1587 = 0,3174$$

- b) A probabilidade de ocorrer valor maior do que a média zero é 0,5. A tabela de distribuição normal reduzida mostra que a probabilidade de ocorrer valor entre a média zero e $z = 1,96$ (procure $z = 1,96$ na tabela) é 0,4975. Então a probabilidade de ocorrer valor maior do que $z = 1,96$ é:

$$0,5000 - 0,4975 = 0,0025.$$

Como a curva é simétrica, a probabilidade de ocorrer valor fora dos limites $z = 1,96$ e $z = -1,96$ é:

$$2 \times 0,0025 = 0,0500.$$

10.5.2 – Em homens, a quantidade de hemoglobina por 100 ml de sangue é uma variável aleatória com distribuição normal de média $\mu = 16$ g e desvio padrão $\sigma = 1$ g. Calcule a probabilidade de um homem apresentar de 16 a 18 g de hemoglobina por 100 ml de sangue.

Primeiro, é preciso calcular:

$$z = \frac{x - \mu}{\sigma} = \frac{18 - 16}{1} = 2$$

A probabilidade de X assumir valor entre a média 16 e o valor 18 corresponde à probabilidade de Z assumir valor entre a média zero e o valor 2 (área sombreada na Figura 10.11). Esta probabilidade, que pode ser encontrada na tabela de distribuição normal reduzida, é 0,4772. Então a probabilidade de um homem apresentar de 16 a 18 g de hemoglobina por 100 ml de sangue é 0,4772 ou 47,72%.

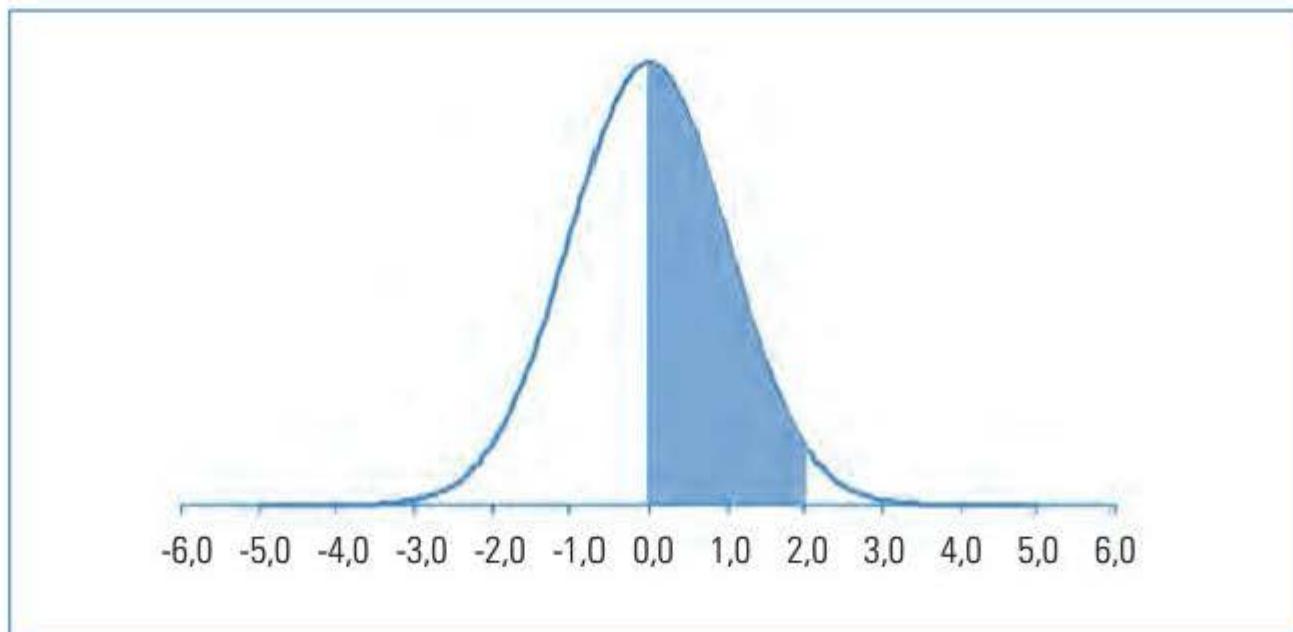


FIGURA 10.11 Probabilidade de taxa de hemoglobina entre 16 e 18 g de hemoglobina por 100 ml de sangue.

10.5.3 – No problema 10.5.2, qual é a probabilidade de um homem apresentar mais de 18 g de hemoglobina por 100 ml de sangue?

Como para $x = 18$ corresponde $z = 2$, e a probabilidade de Z assumir valor entre a média zero e o valor $z = 2$ é 0,4772, segue-se que a probabilidade de Z assumir valor maior do que 2 é:

$$0,5 - 0,4772 = 0,0228 \text{ ou } 2,28\%.$$

10.5.4 – Sabe-se que o tempo médio para completar um teste, feito para candidatos ao vestibular de uma escola, é de 58 minutos, com desvio padrão igual a 9,5 minutos. Se o responsável pelo vestibular quiser que apenas 90% dos candidatos terminem o teste, quanto tempo deve dar aos candidatos para que entreguem o teste?

Para resolver o problema, primeiro observe a Figura 10.12. Lembre que a média delimita 0,5 da distribuição. Então é preciso achar o valor de z que corresponde à probabilidade 0,4 (porque $0,4 + 0,5 = 0,9$, ou seja, os 90% pedidos). Na tabela de distribuição normal reduzida, você encontra, para 0,3997, que é o valor mais próximo de 0,4, o ponto $z = 1,28$. Como

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

$$x = \mu + z\sigma = 58 + 1,28 \times 9,5 = 70,16$$

ou seja, devem ser fixados 70 minutos (ou, mais exatamente, 70,16 minutos) para terminar o teste.

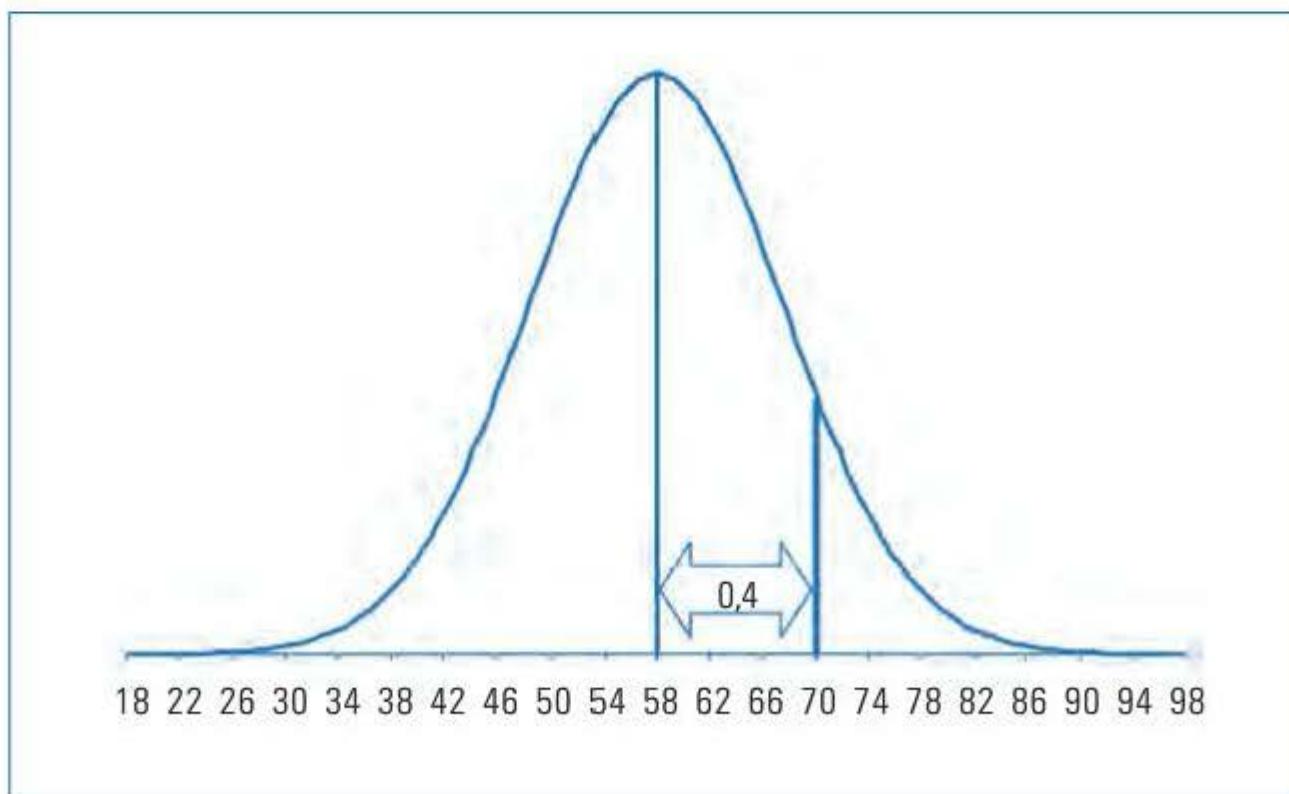


FIGURA 10.12 Distribuição do tempo despendido para completar o teste.

10.5.5 – Se X tem distribuição normal de média $\mu = 150$ e 97,5% dos valores de X são menores do que 210, qual é o desvio padrão da distribuição?

A média delimita 0,5 da distribuição. Observe a Figura 10.13: é preciso achar o valor de z que corresponde à probabilidade 0,475 (porque $0,475 + 0,5 = 0,975$, ou seja, 97,5%). Na tabela de distribuição normal reduzida, você encontra, para 0,475, o ponto $z = 1,96$. Como

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

$$\sigma = \frac{x - \mu}{z} = \frac{210 - 150}{1,96} = 30,61$$

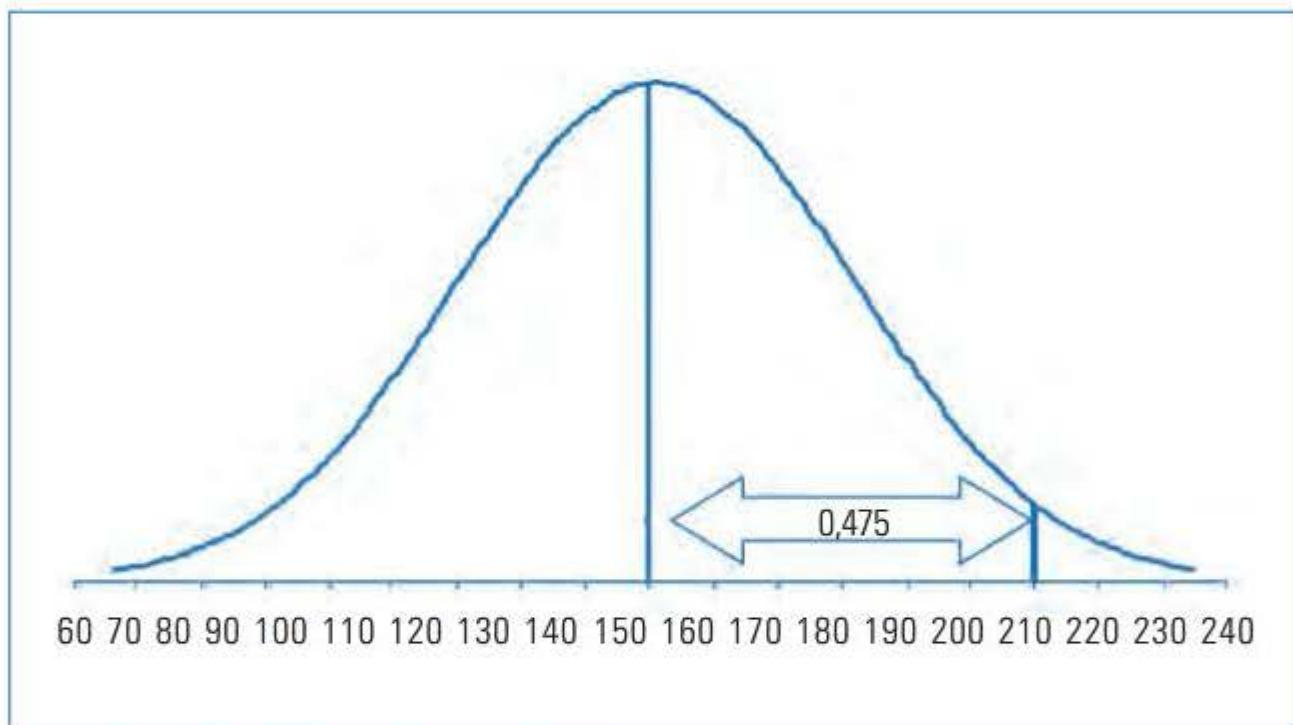


FIGURA 10.13 Distribuição da variável X

10.6 – EXERCÍCIOS PROPOSTOS

10.6.1 – O quociente de inteligência tem média 100 e desvio padrão 15. Qual é a proporção de pessoas com quociente de inteligência acima de 135?

10.6.2 – Em uma distribuição normal reduzida, que valores de z englobam: a) 50% dos casos que ficam no centro da distribuição? b) 90% dos casos que ficam no centro da distribuição? c) 95% dos casos que ficam no centro da distribuição?

10.6.3 – Suponha que a pressão sanguínea sistólica em indivíduos com idade entre 15 e 25 anos é uma variável aleatória com distribuição aproximadamente normal de média $\mu = 120\text{mmHg}$ e desvio padrão $\sigma = 8\text{mmHg}$. Nestas condições, calcule a probabilidade de um indivíduo dessa faixa etária apresentar pressão: a) entre 110 e 130mmHg; b) maior do que 130mmHg.

10.6.4 – A taxa de glicose no sangue humano é uma variável aleatória com distribuição normal de média $\mu = 100\text{ mg por 100 ml de sangue}$ e desvio padrão $\sigma = 6\text{ mg por 100 ml de sangue}$. Calcule a probabilidade de um indivíduo apresentar taxa: a) superior a 110 mg por 100 ml de sangue; b) entre 90 e 100 mg por 100 ml de sangue.

10.6.5 – Em um hospital psiquiátrico, os pacientes permanecem internados em média 50 dias, com um desvio padrão de 10 dias. Se for razoável pressupor que o tempo de permanência tem distribuição aproximadamente normal, qual é a probabilidade de um paciente permanecer no hospital: a) mais de 30 dias? b) menos de 30 dias?

10.6.6 – A estatura de recém-nascidos do sexo masculino é uma variável aleatória com distribuição aproximadamente normal de média $\mu = 50\text{ cm}$ e desvio padrão $\sigma = 2,50\text{ cm}$. Calcule a probabilidade de um recém-nascido do sexo masculino ter estatura: a) inferior a 48 cm; b) superior a 52 cm.

10.6.7 – A concentração de sódio no plasma tem média igual a 139,5 mEq/l de plasma, com desvio padrão igual a 3 mEq/L de plasma. Que valor você poria como ponto de corte para dizer que está alta a concentração de sódio no plasma de uma pessoa?

10.6.8 – Em uma distribuição normal reduzida, que proporção de casos cai: a) acima de $z = 1$? b) abaixo de $z = -2$? c) abaixo de $z = 0$? d) acima de $z = 1,28$?

10.6.9 – Na distribuição normal reduzida, a média é sempre zero. Isso sugere que metade dos escores é positiva e metade é negativa? Explique sua resposta.

10.6.10 – Em uma academia, os ginastas levantam, em média, 80 kg de peso, com desvio padrão de 12 kg. Pressupondo distribuição normal, que proporção dos ginastas levanta mais de 100 kg?

(página deixada intencionalmente em branco)

**Intervalo de
Confiança**

11

(página deixada intencionalmente em branco)

Os resultados das pesquisas são expressos de maneiras diferentes. A forma de apresentar os resultados depende, em muito, do tipo de variável e do delineamento do experimento. Neste Capítulo, vamos nos concentrar em duas formas de expressar resultados — por meio de uma *proporção* (nas pesquisas em que a variável é qualitativa) ou por meio de uma *média* (nas pesquisas em que a variável é quantitativa). Veja dois exemplos que tornam a situação mais concreta.

Exemplo 11.1: Uma proporção.

Um dentista examinou 100 crianças que ingressavam no ensino fundamental e verificou que 33 delas não tinham cárie. A proporção de crianças sem cárie na amostra é 33/100, ou seja, 0,33. Essa proporção é uma estimativa da probabilidade de uma criança, da mesma população de onde proveio a amostra, não ter cáries. Será uma boa estimativa?

Antes de responder à pergunta, é preciso saber se as crianças examinadas são realmente *representativas* da população em estudo. Se o dentista disser que sim, tomaremos isso como *pressuposição* porque, para saber se a amostra é representativa da população, são necessários conhecimentos na área em que a pesquisa se enquadra — não de Estatística.

Depois, é preciso pensar na *margem de erro* da estimativa fornecida pela pesquisa. Será que as crianças selecionadas para a amostra poderiam ter experiência de cárie mais alta (ou mais baixa) do que as crianças da população de onde a amostra foi retirada, por *simples acaso*? É preciso informar, de alguma maneira, a confiança que se pode ter na estimativa. É isto que veremos neste Capítulo.

Exemplo 11.2: A média.

Um professor de Fisioterapia obteve dados biométricos dos alunos que ingressaram na faculdade. A média da pressão sanguínea sistólica de 100 alunos foi 120,3mmHg com desvio padrão de 14,0mmHg. O professor considera que esses alunos constituem amostra representativa de outros alunos que ingressam em outros cursos da universidade em outros anos. Mas que confiança pode ter na estimativa da média que está fornecendo?

11.1 – INTERVALO DE CONFIANÇA PARA UMA PROPORÇÃO

O fato de sabermos a proporção de determinado evento em uma amostra *não* nos garante o conhecimento da proporção desse evento na população. O que podemos fazer, usando conhecimentos de Estatística, é calcular um intervalo que *possa incluir* a proporção do evento na população (o parâmetro).

A maioria dos pesquisadores considera aceitável um *intervalo de 95% de confiança*. Isto significa que o pesquisador terá 95% de probabilidade de obter, com base em uma amostra, um intervalo de confiança que venha a conter a proporção do evento na população (o parâmetro).

Entenda bem: se você calculou um intervalo de confiança com base em uma amostra, *não sabe* se o parâmetro (valor na população) está contido no intervalo que calculou. No entanto, você sabe que 95% dos intervalos construídos da mesma forma conterão o parâmetro.

11.1.1 – Cálculo do intervalo de confiança para uma proporção

Você viu, no Capítulo 9, o que é uma variável aleatória com distribuição binomial: são feitas n tentativas; cada tentativa só pode resultar em um de dois eventos possíveis; o número de vezes que ocorre o evento de interesse é a variável X . Agora, reveja o Exemplo 11.1: um dentista examinou 100 crianças. Cada criança foi classificada como tendo, ou não, experiência de cárie. Então o número de crianças sem experiência de cárie nas 100 examinadas é uma variável binomial.

A proporção de valores X , obtida com base em uma amostra, é:

$$\bar{p} = \frac{X}{n}$$

Essa proporção é uma estimativa da probabilidade de ocorrer o evento de interesse na população. Essa estimativa está associada a uma variabilidade. A variabilidade é medida pelo desvio padrão. O desvio padrão é dado por:

$$\sigma = \sqrt{\frac{\bar{p} \times \bar{q}}{n}}$$

O intervalo de 95% de confiança para a probabilidade p , obedecidas às condições apontadas na Seção 11.3, é dado por:

$$\bar{p} \pm 1,96 \sqrt{\frac{\bar{p} \times \bar{q}}{n}}$$

Os valores $\pm 1,96$ são obtidos da distribuição normal¹. Lembre que são esses os valores de z que englobam 95% dos casos que ficam no centro da distribuição. Esta fórmula vale para grandes amostras.

Exemplo 11.3: Intervalo de confiança para uma proporção.

Lembre o Exemplo 11.1: um dentista examinou 100 crianças e verificou que 33 delas não tinham cárie. A proporção de crianças sem cárie é 0,33. O dentista quer então saber se esse valor é uma boa estimativa da probabilidade de uma criança da mesma população de onde proveio a amostra não ter cárries.

O intervalo de confiança é dado por:

$$\bar{p} \pm 1,96 \sqrt{\frac{\bar{p} \times \bar{q}}{n}}$$

No exemplo, $p = 0,33$; $q = 1 - 0,33 = 0,67$; $n = 100$.

Logo:

$$\begin{aligned} & 0,33 \pm 1,96 \sqrt{\frac{0,33 \times 0,67}{100}} \\ & = 0,33 \pm 1,96 \times 0,047 \\ & = 0,33 \pm 0,092 \end{aligned}$$

Os limites do intervalo de 95% de confiança são, portanto, $0,33 - 0,092 = 0,238$ e $0,33 + 0,092 = 0,422$. Podemos então ter 95% de confiança de que a probabilidade de uma criança da população de onde proveio a amostra não ter cárries esteja entre 0,238 e 0,422 ou, em porcentagem, entre 23,8% e 42,2%.

11.1.2 – Pressuposições

Para construir um intervalo de confiança, algumas pressuposições precisam ser feitas. Primeiro, a amostra deve ser *representativa* da população. Por exemplo, se for pedido num show de televisão que os telespectadores telefonem dizendo se gostam ou não do programa, não tem sentido usar como indicador do grau de aprovação a proporção de pessoas que telefonaram dizendo que gostam do programa, pelo simples fato de que quem não gosta de um programa provavelmente não o assiste.

¹Essa fórmula considera que a distribuição da variável binomial aproxima-se de uma distribuição normal. Para que isso aconteça é preciso que a amostra seja grande. Use a fórmula se $np > 5$ e $nq > 5$ ou, pelo menos, que $0,3 < p < 0,7$.

Outra pressuposição importante é a de *independência* das observações. O fato de uma pessoa ter sido selecionada para a amostra não deve mudar a probabilidade de outra pessoa ser, também, selecionada. Por exemplo, não se deve entrevistar alguém e depois pedir para essa pessoa trazer outras para serem entrevistadas.

Finalmente, uma observação que não se refere às pressuposições, mas à interpretação de um intervalo de confiança. O intervalo que você construiu pode *conter, ou não conter*, o parâmetro. Sabe-se que, se você repetir o procedimento — da mesma maneira — muitas e muitas vezes, espere-se que 95% dos intervalos calculados contenham o parâmetro. Portanto, *não é correto* dizer que a probabilidade de o intervalo conter o parâmetro é de 95%.

11.1.3 – A margem de erro

A proporção de determinado evento na amostra *estima* a proporção desse evento na população de onde a amostra foi selecionada. O intervalo de confiança, na forma apresentada neste Capítulo, fornece a *margem de erro da estimativa*. Essa margem é dada pela amplitude do intervalo de confiança.

Exemplo 11.4: Margem de erro: amostra pequena.

Lembre o Exemplo 11.1: um dentista examinou 100 crianças e verificou que 33 delas não tinham cárie. A proporção de crianças sem cárie é 0,33. O dentista obteve o intervalo de 95% de confiança. Os limites desse intervalo são 0,238 e 0,422. Qual é a margem de erro?

A margem de erro é dada pela amplitude do intervalo, ou seja, pela diferença:
 $0,422 - 0,238 = 0,184$

Então o dentista está 95% seguro de que a proporção de crianças sem cárie na população estudada está entre 23,8 e 42,2%. A margem de erro é de 18,4%.

Para diminuir a margem de erro, é preciso aumentar a amostra. Daí a insistência dos estatísticos em dizer que a amostra deva ser tão grande quanto possível². Veja o Exemplo 11.5.

²No caso de estimativas de proporções (que em geral são transformadas em porcentagem), as amostras devem ser maiores do que 100. Se p for muito pequeno, as amostras devem ser ainda maiores.

Exemplo 11.5: Margem de erro: amostra grande.

Lembre o Exemplo 11.1. Imagine que o dentista examinou não 100, mas 1.000 crianças e verificou que 330 delas não tinham cárie. A proporção de crianças sem cárie é 0,33. Qual é a margem de erro?

O intervalo de confiança é dado por:

$$\bar{p} \pm 1,96 \sqrt{\frac{p \times q}{n}}$$

Em que $p = 0,33$; $q = 1 - 0,33 = 0,67$; $n = 1.000$.

Logo:

$$\begin{aligned} & 0,33 \pm 1,96 \sqrt{\frac{0,33 \times 0,67}{1000}} \\ & = 0,33 \pm 1,96 \sqrt{0,000221} \\ & = 0,33 \pm 1,96 \times 0,01487 \\ & = 0,33 \pm 0,029 \end{aligned}$$

Os limites do intervalo são 0,301 e 0,359. A margem de erro é dada pela diferença: $0,359 - 0,301 = 0,058$.

Neste exemplo, o dentista está 95% seguro de que a proporção de crianças sem cárie na população está entre 30,1 e 35,9%. A margem de erro é de 5,8%. Compare este resultado com aquele obtido no Exemplo 11.4 e verifique: a margem de erro diminui quando a amostra aumenta.

11.2 – INTERVALO DE CONFIANÇA PARA UMA MÉDIA

Imagine uma amostra casual simples de n elementos. A média dos dados dessa amostra constitui uma *estimativa* da média da população de onde essa amostra provém. Veja o Exemplo 11.2. O intervalo de confiança para a média, que veremos aqui, indica a *precisão* da estimativa. Antes, porém, de aprender como calcular o intervalo de confiança, é preciso entender o que é erro padrão da média.

11.2.1 – Erro padrão da média

Imagine uma população constituída pelos valores 4, 10 e 16. A média dessa população, que se indica por μ é:

$$\mu = \frac{4+10+16}{3} = \frac{30}{3} = 10$$

Considere, agora, todas as amostras possíveis de dois elementos que podem ser retirados dessa população, admitindo que todo elemento retirado para compor a amostra é reposto antes da retirada do segundo. Isso significa que dois elementos podem ser retirados *ad infinitum* da população. Portanto, podemos entender a população como *infinita*. Essas amostras, e as respectivas médias, estão na Tabela 11.1 e na Figura 11.1. É fácil ver, observando a Figura 11.1, que as médias das amostras distribuem-se em torno da média $\mu = 10$ da população.

TABELA 11.1
Médias das amostras de dois elementos obtidos da população constituída pelos números 4, 10 e 16.

Amostra	Média
4 e 4	4
4 e 10	7
4 e 16	10
10 e 4	7
10 e 10	10
10 e 16	13
16 e 4	10
16 e 10	13
16 e 16	16

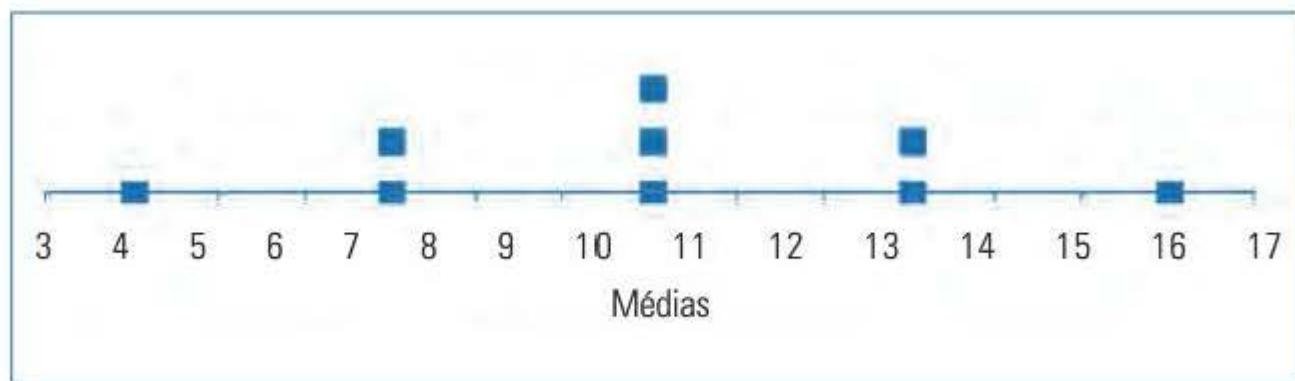


FIGURA 11.1 Distribuição das médias das amostras.

O grau de dispersão das médias das amostras em torno da média da população é dado pela *variância da média*. Essa medida, que se indica por σ_x^2 , é dada pela fórmula:

$$\sigma_x^2 = \frac{\sum_{i=1}^r (\bar{x}_i - \mu)^2}{r}$$

em que x_i é a média da i -ésima amostra e r é o número de amostras que podem ser obtidas da população.

Para as médias apresentadas na Tabela 11.1, a variância da média é:

$$\sigma_x^2 = \frac{(4-10)^2 + (7-10)^2 + (10-10)^2 + \dots + (16-10)^2}{9} = \frac{108}{9} = 12$$

Na prática, é impossível calcular a variância da média pela fórmula apresentada: o pesquisador *não* dispõe de todas as amostras possíveis — mas de uma *única* amostra, para estimar a média μ da população e obter uma medida de precisão dessa estimativa. Existe, no entanto, uma solução: já se demonstrou que uma estimativa da variância da média⁵ é dada pela fórmula:

$$s_x^2 = \frac{s^2}{n}$$

em que s^2 é a variância da amostra.

As médias, as variâncias e as variâncias das médias das amostras apresentadas na Tabela 11.1 estão na Tabela 11.2. É importante notar que a *média das médias* coincide com a média $\mu = 10$ da população e que a *média das variâncias das médias* das amostras é igual a 12, calculada anteriormente.

TABELA 11.2
Médias, variâncias e variâncias das médias das amostras apresentadas na Tabela 11.1.

Amostra	Média	Variância	Variância da média
4 e 4	4	0	0
4 e 10	7	18	9
4 e 16	10	72	36
10 e 4	7	18	9
10 e 10	10	0	0
10 e 16	13	18	9
16 e 4	10	72	36
16 e 10	13	18	9
16 e 16	16	0	0
Média	10	24	12

⁵Note que, para isto ser verdade, é preciso que as variâncias das amostras tenham sido estimadas usando os graus de liberdade como divisor.

Por definição, *erro padrão da média* é a raiz quadrada com sinal positivo da variância da média. Indica-se a estimativa do erro padrão da média por $s_{\bar{x}}$. O erro padrão da média é uma estimativa da variabilidade das médias que seriam obtidas, caso o pesquisador tivesse tomado, nas mesmas condições, um grande número de amostras. A fórmula é:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Exemplo 11.6: Erro padrão da média.

Reveja o Exemplo 11.2: A média da pressão sangüínea sistólica de 100 alunos foi 120,3mmHg com desvio padrão de 14,0mmHg. Qual é o erro padrão da média?

Aplicando a fórmula, vem:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{14,0}{\sqrt{100}} = 1,4$$

11.2.2 – Cálculo do intervalo de confiança para uma média

É pouco provável que, com base nos dados de uma amostra, o pesquisador obtenha uma estimativa (por exemplo, da média) igual ao parâmetro (no caso, da média da população). Mas — intuitivamente — você sabe que, se for examinada boa parte da população, a média da amostra terá valor próximo da média da população; se a variável variar pouco, a média terá valor próximo ao da média da população. Então uma estimativa é tanto melhor quanto *maior for a amostra* e quanto *menor for a variabilidade* dos dados.

Imagine agora que o pesquisador está estudando uma variável X com distribuição normal de média μ e variância σ^2 . Foram obtidas, com base em uma amostra casual simples de n elementos dessa população, estimativas da média, do desvio padrão e do erro padrão da média. Mas o pesquisador precisa dar indicação da *precisão* da estimativa da média. Deve, então, calcular um intervalo de confiança. Já vimos que os pesquisadores geralmente aceitam que o intervalo calculado inclua o valor populacional com probabilidade de 95%.

O intervalo de 95% de confiança para a média μ , desde que a amostra seja suficientemente grande⁴, é dado por:

$$\mu \pm 1,96 s_{\bar{x}}$$

⁴Esta fórmula serve para amostras grandes que, no caso de estimativas de médias, devem ser, pelo menos, de tamanho maior do que 30.

Exemplo 11.7: Intervalo de confiança para a média.

Reveja o Exemplo 11.2: A média da pressão sanguínea sistólica de 100 alunos foi 120,3mmHg com desvio padrão de 14,0 milímetros de mercúrio e erro padrão da média igual a 1,4mmHg. Que confiança o professor pode ter no resultado?

O intervalo de confiança é dado por:

$$\bar{x} \pm 1,96 s_{\bar{x}}$$

No exemplo, a média é 120,3 e o erro padrão da média é 1,4; $n = 100$.

Logo:

$$\begin{aligned} 120,3 &\pm 1,96 \times 1,4 \\ &= 120,3 \pm 2,74 \end{aligned}$$

Os limites do intervalo de 95% de confiança são, portanto, $120,3 - 2,74 = 117,56$ e $120,3 + 2,74 = 123,04$. Podemos então ter 95% de confiança de que a média da pressão sanguínea sistólica dos alunos que ingressam na faculdade está entre 117,56 e 123,04mmHg.

11.3 – CUIDADOS NA INTERPRETAÇÃO DOS INTERVALOS DE CONFIANÇA

A interpretação do intervalo de confiança exige cuidados. Na prática, o pesquisador dispõe de uma única amostra que fornece uma só estimativa de determinado parâmetro. O pesquisador calcula um intervalo de 95% de confiança, mas *não sabe* se o parâmetro está, ou não, contido no intervalo que calculou. Sabe apenas que 95% dos intervalos de confiança calculados da mesma forma contêm o parâmetro. A *margem de erro da estimativa* é dada pela amplitude do intervalo de confiança. Quanto maior a amostra, menor é a margem de erro — o intervalo de confiança fica menor — mas, ainda assim, não significa que contenha o parâmetro.

11.4 – PEQUENAS AMOSTRAS

Este livro não ensina como calcular o intervalo de confiança para uma proporção nos casos de pequenas amostras. No caso de variáveis contínuas, desde que a distribuição seja aproximadamente normal, é possível calcular o intervalo de confiança para a média de maneira similar à apresentada na Seção 11.2.

Você calcula o intervalo,

$$\bar{x} \pm t_{(n-1)} s_{\bar{x}}$$

em que $t_{(n-1)}$ é um valor encontrado na Tabela de distribuição de t (veja Apêndice). A variável t é obtida de uma distribuição teórica⁵ chamada distribuição t , de certa forma parecida com a distribuição normal reduzida. O gráfico da distribuição tem a forma de sino e é simétrico em torno da média zero.

Para entender como se acha o valor crítico de t , veja a Tabela 11.3, que reproduz parte da Tabela de distribuição t . É preciso especificar os *graus de liberdade*. No caso do intervalo de confiança para uma média, os graus de liberdade são os do erro padrão da média, ou seja, $(n - 1)$. Você também precisa especificar a confiança, que em geral é de 90% ou 95%. Então, para achar o valor de t que se usa na fórmula, siga os passos:

1. O tamanho da amostra é n . Digamos que $n = 15$. Ache os *graus de liberdade*, isto é $n - 1$. No caso, $15 - 1 = 14$.
2. Escolha o *nível de confiança* que você quer. Ache o valor de α subtraindo o nível de confiança de 100%. Para 95% de confiança, calcule $\alpha = 100 - 95 = 5$.
3. Procure, na Tabela de valores de t , o valor que fica no cruzamento da coluna “ $\alpha = 5\%$ ” com a linha “graus de liberdade 14”.
4. Você acha $t = 2,145$.

Então, o intervalo de 95% de confiança é:

$$\bar{x} \pm 2,145 s_{\bar{x}}$$

TABELA 11.3
Tabela (parcial) de distribuição t .

<i>Graus de liberdade</i>	<i>Nível de significância</i>		
	0,01	0,05	0,10
11	3,106	2,201	1,796
12	3,055	2,179	1,782
13	3,012	2,160	1,771
14	2,977	2,145	1,761
15	2,947	2,131	1,753
16	2,921	2,120	1,746

⁵Existe uma distribuição t para cada tamanho de amostra. Portanto, existe uma família de distribuições t .

Exemplo 11.8: Intervalo de confiança para a média, amostras pequenas.

Com base em uma amostra casual simples de $n = 25$ indivíduos, foram obtidos a média $\bar{x} = 198 \text{ mg}/100 \text{ ml}$ e o desvio padrão $s = 30 \text{ mg}/100\text{ml}$ da taxa de colesterol no plasma sanguíneo humano. Ache o intervalo de 90% de confiança.

Para um nível de 90% de confiança, $\alpha = 10\%$. Como $n = 25$ indivíduos, $n - 1 = 25 - 1 = 24$. O valor de t , na Tabela dos valores de t (veja Apêndice), para $\alpha = 10\%$ e com 24 graus de liberdade, é 1,71. A expressão do intervalo de confiança fica, então, como segue:

$$187,74 \leq \mu \leq 208,26$$

É preciso considerar aqui dois fatos importantes:

1. Na área da saúde — e em outras áreas — muitas vezes o resultado do trabalho é apresentado na forma:

$$\begin{aligned}\bar{x} &\pm s \\ \bar{x} &\pm 2s\end{aligned}$$

Como aprendemos no Capítulo 10, esses intervalos referem-se aos *dados* — porque na fórmula está o desvio padrão, que mede a variabilidade dos dados. Se a média e o desvio padrão da amostra são boas estimativas dos parâmetros μ e σ , é razoável considerar que o primeiro intervalo

$$\bar{x} \pm s$$

contenha cerca de 2/3 dos dados (68,26%) e o segundo

$$\bar{x} \pm 2s$$

contenha perto de 95% dos dados (95,44%).

2. Entretanto, é preciso deixar claro que a área da saúde — e em outras áreas — também se apresenta o resultado do trabalho na forma:

$$\bar{x} \pm s_{\bar{x}}$$

ou

$$\bar{x} \pm 2s_{\bar{x}}$$

Neste caso, o primeiro intervalo é um intervalo de 68,26% de confiança para o parâmetro μ — a *média da população*, desde que a amostra seja suficientemente grande, porque no cálculo entra o erro padrão da média. O segundo é um intervalo de 95,44% de confiança para o parâmetro μ — a *média da população*, desde que a amostra

seja suficientemente grande. Este não é, porém, verdade do caso das amostras pequenas — como as amostras de tamanho 6 ou 10.

Finalmente, um lembrete: algumas revistas internacionais não aceitam informações do tipo: $19,3 \pm 2,1$, porque não sabem exatamente o significado desse intervalo: se é um intervalo de confiança para os dados (2,1 seria o desvio padrão), ou se é um intervalo de confiança para a média (2,1 seria o erro padrão da média).

Exemplo 11.9: Intervalo de confiança para a média; amostra de tamanho 6.

Calcule o intervalo de 90% de confiança para a média de uma amostra de seis elementos.

O valor de t dado na Tabela de valores de t no final do livro é 2,02. Então o intervalo de 90% de confiança é:

$$\bar{x} \pm 2,02 s_{\bar{x}}$$

o dobro do intervalo que às vezes se apresenta, sem determinar o nível de confiança:

$$\bar{x} \pm s_{\bar{x}}$$

11.5 – EXERCÍCIOS RESOLVIDOS

11.5.1 – Dos 90 pacientes que se submeteram a uma nova técnica cirúrgica, morreram nove. Calcule o intervalo de 95% de confiança para a probabilidade de morte na cirurgia.

A proporção de mortes na amostra foi:

$$p = \frac{9}{90} = 0,10$$

e atende aos requisitos para aplicar a distribuição normal ($np = 90 \times 0,1 = 9 > 5$ e $nq = 90 \times 0,9 = 81 > 5$). Então:

$$p \pm 1,96 \sqrt{\frac{0,10 \times (1 - 0,10)}{90}} = 0,10 \pm 1,96 \times 0,0316 = 0,10 \pm 0,0620 \\ 0,0380 \leq p \leq 0,1620.$$

11.5.2 – Foi feito um ensaio⁶ com 100 pacientes para testar uma nova droga que, se presume, abaixa a pressão sanguínea. Verificou-se que a nova droga, em comparação à droga usualmente recomendada (padrão), diminui a pressão em 6%. Você pode calcular um intervalo de confiança para essa porcentagem?

Embora esta questão pareça similar à anterior, não é. Na questão anterior, havia, realmente, uma proporção. Nesta questão, a porcentagem é uma mudança em uma medida então não se pode calcular o intervalo de confiança.

11.5.3 – O extremo inferior de um intervalo de confiança para proporção pode ser negativo? Pode ser igual a zero?

É impossível o extremo inferior de um intervalo de confiança para proporção ser negativo e só é zero quando o desvio padrão é zero.

11.5.4 – A pressão sanguínea sistólica medida em 100 militares apresentou média igual a 125mmHg e desvio padrão é 9mmHg. Calcule o erro padrão da média e ache o intervalo de 95% para a média populacional.

$$s_{\bar{x}} = \frac{9}{\sqrt{100}} = 0,9$$

$$\bar{x} \pm 1,96 s_{\bar{x}} = 125 \pm 1,96 \times 0,90 = 125 \pm 1,764$$

O intervalo varia entre 123,2mmHg e 126,8mmHg.

11.5.5 – A pressão sanguínea sistólica medida em 10 militares apresentou média igual a 125mmHg e o desvio padrão é igual a 9mmHg. Calcule o erro padrão da média e ache o intervalo de 95% para a média populacional.

$$s_{\bar{x}} = \frac{9}{\sqrt{10}} = 2,846$$

$$\bar{x} \pm 1,96 s_{\bar{x}} = 125 \pm 1,96 \times 2,846 = 125 \pm 5,578$$

O intervalo varia entre 119,4mmHg e 130,6mmHg.

11.5.6 – Compare os intervalos de confiança obtidos nos exercícios 11.5.4 e 11.5.5.

O erro padrão da média diminui quando você aumenta o tamanho da amostra. Não se espera que isso aconteça com o desvio padrão, que mede a va-

⁶Este problema foi proposto em: MOTULSKY, H. *Intuitive Biostatistics*. Nova York, Oxford University Press, 1995. p.316.

riabilidade dos *dados*. É verdade que se você aumentar a amostra, os parâmetros ficam estimados com maior exatidão. O valor do desvio padrão pode, então, mudar, mas não existe tendência de o desvio padrão aumentar, ou diminuir, quando se aumenta o tamanho da amostra. No entanto, o erro padrão da média diminui porque a média da amostra tende a ter valor mais próximo da média verdadeira. E você vê isso na amplitude do intervalo de confiança.

11.6 – EXERCÍCIOS PROPOSTOS

11.6.1 – *Foi feito um estudo para levantar a proporção de adultos que sofrem de síndrome de fadiga crônica⁷. Para isso, foram selecionados ao acaso 4.000 membros saudáveis de uma organização em Seattle. Para essas pessoas, foram distribuídos questionários nos quais se perguntava se, nos seis meses anteriores, elas haviam sentido cansaço excessivo, que interferisse no trabalho ou nas responsabilidades em casa. Das 3.066 pessoas que responderam (possível tendência devido à falta de quase um quarto de não respondentes), 590 relataram fadiga crônica. Estime a proporção de pessoas que pensam ter síndrome de fadiga crônica e o intervalo de 95% de confiança.*

11.6.2 – *No estudo apresentado no problema anterior, os pesquisadores examinaram os 590 questionários de pessoas que relataram fadiga crônica e eliminaram todos aqueles cujos problemas, de natureza médica ou psiquiátrica, pudesse explicar a fadiga. Sobraram 74 questionários. Destes, apenas três tinham a síndrome (que se caracteriza por falta de concentração, falha na memória recente, dificuldade em dormir, dores musculares e nas articulações). Qual seria a proporção de adultos portadores da síndrome?*

11.6.3 – *Seja X a variável aleatória que representa a pressão sanguínea sistólica em indivíduos com idade entre 20 e 25 anos. Essa variável tem distribuição aproximadamente normal. Suponha que, com base em uma amostra de 100 indivíduos, foi obtida a média $\bar{x} = 123\text{mmHg}$ e o desvio padrão $s = 8\text{mmHg}$. Determine o intervalo de 90% de confiança para a média da população (μ).*

11.6.4 – *Seja X a variável aleatória que representa a taxa de hemoglobina em mulheres. Imagine que, com base em uma amostra aleatória de 200 mulheres, obteve-se a média $\bar{x} = 16,2\text{ g}$ de hemoglobina por 100 ml de sangue*

⁷ALIAGA, M. e GUNDERSON, B. *Interactive Statistics*. New Jersey, Prentice Hall, 2 ed. 2003. p. 539.

e o desvio padrão $s = 1,1$ g. Determine o intervalo de 95% de confiança para μ , supondo que X é uma variável com distribuição normal.

11.6.5 – Seja X a variável aleatória que representa a estatura ao nascer para o sexo masculino. Com base em 28 recém-nascidos masculinos, obtiveram-se $\bar{x} = 50$ cm e $s = 2,5$ cm. Calcule o intervalo de 90% de confiança para μ , pressupondo distribuição normal.

11.6.6 – Seja X a variável aleatória que representa a taxa de glicose no sangue humano. Determine o intervalo de 95% de confiança para μ , supondo que uma amostra de 25 pessoas forneceu média $\bar{x} = 95$ mg de glicose por 100 ml de sangue e o desvio padrão $s = 6$ mg. Suponha que X tem distribuição normal.

11.6.7 – É possível calcular⁸, com base em uma amostra, um intervalo de 100% de confiança para um parâmetro p , que indica determinada probabilidade?

11.6.8 – Num estudo sobre qualidades nutricionais⁹ de lanches rápidos, mediu-se a quantidade de gordura em 100 hambúrgueres de determinada cadeia de restaurantes. Achou-se média de 30,2 gramas e desvio padrão de 3,8 gramas. Construa um intervalo de 95% de confiança para a quantidade média de gordura nos hambúrgueres servidos nesses restaurantes.

11.6.9 – No mesmo estudo citado no Exercício 14.6.7, foi medida a quantidade de sal e se achou média 658 mg e desvio padrão 47 mg. Ache o intervalo de 98% de confiança.

11.6.10 – Uma enfermeira mediu o comprimento de 105 bebês do sexo masculino e achou o intervalo de 90% de confiança para a média, em centímetros: (45,3; 53,2). Responda brevemente às questões feitas em seguida:

- A média da população está no intervalo (45,3; 53,2)?
- A média da amostra está no intervalo (45,3; 53,2)?
- Novas amostras de 105 bebês do sexo masculino darão médias no intervalo (45,3; 53,2)?
- Um intervalo de 99% de confiança seria mais estreito?

⁸Este problema foi proposto em: MOTULSKY, H. *Intuitive Biostatistics*. Nova York, Oxford University Press, 1995. p.318.

⁹JOHNSON, R. E TSUI, K. W. *Statistical reasoning and methods*. Nova York, Wiley, 1998. p.338.

(página deixada intencionalmente em branco)

Teste de Qui-quadrado

12

(página deixada intencionalmente em branco)

As pesquisas são feitas com o objetivo de responder perguntas. Para responder perguntas, são necessárias informações que, na área de saúde, são quase sempre obtidas por meio de amostras. Mas os pesquisadores querem *generalizar seus achados* para toda a população de onde a amostra foi retirada. Isto pode ser feito, desde que a generalização seja fundamentada em um *teste de hipóteses*.

Para fazer o teste, a pergunta do pesquisador é transformada em duas *hipóteses*, ou seja, é escrita na forma de duas afirmativas que se contradizem — como nos testes de falso/verdadeiro. A idéia de construir hipóteses é complexa, mas fica bem entendida com um exemplo da área jurídica.

Exemplo 12.1: Hipóteses: inocente ou culpado.

Um réu está sendo julgado. Quais são as hipóteses possíveis?

- O réu é inocente do ato que o acusam.
- O réu é culpado do ato que o acusam.

Construídas as hipóteses, passa-se à análise dos dados para tomar *decisão* por uma das hipóteses.

Exemplo 12.2: Decisão: inocente ou culpado.

Um réu está sendo julgado. Quais são as decisões possíveis?

- Considerar o réu culpado.
- Considerar o réu inocente.

As decisões são tomadas com base em *conhecimento de parte dos fatos*. Então a decisão tomada pode estar errada.

Exemplo 12.3: Erros possíveis.

O réu está sendo julgado. Quais são os erros associados às decisões possíveis?

- Dizer que o réu é culpado, quando é inocente.
- Dizer que o réu é inocente, quando é culpado.

Vamos pensar, agora, em uma pesquisa na área da saúde.

Exemplo 12.4: Construindo as hipóteses.

Duas médicas¹ se perguntaram se a probabilidade de baixo peso ao nascer é maior quando a mãe faz uso continuado de drogas ilícitas durante a gestação.

Para responder à pergunta, é preciso comparar o peso ao nascer de filhos de dois grupos de mães:

- Que usaram drogas ilícitas durante a gestação.
- Que não usaram drogas ilícitas durante a gestação.

Quais são as hipóteses?

- A probabilidade de ter filhos com baixo peso ao nascer é a *mesma* para os dois grupos de mães.
- A probabilidade de ter filhos com baixo peso ao nascer é *maior* para mães que usaram drogas ilícitas durante a gestação.

A pergunta, escrita na forma de duas frases afirmativas que se contradizem, são as *hipóteses*. A primeira é chamada de *hipótese da nulidade* e é indicada por H_0 (lê-se: agá-zero). Na grande maioria das vezes, a hipótese da nulidade é a de que *não existe diferença* entre grupos de dados.

A segunda hipótese contradiz a primeira e é, por isso, chamada de *hipótese alternativa*. Indica-se por H_1 (lê-se: agá-um). Na grande maioria das vezes, a hipótese alternativa é o que o pesquisador gostaria de poder afirmar.

Exemplo 12.5: Coletando a amostra.

Para responder à pergunta feita, as médicas acompanharam a gravidez e anotaram o peso ao nascer dos filhos de 456 adolescentes, usuárias e não-usuárias de drogas ilícitas. Portanto, as médicas conheciam bem as adolescentes que participaram da pesquisa (amostra). Mas o que elas observaram na amostra pode ser estendido para toda a população de adolescentes de onde a amostra foi retirada?

Os pesquisadores sempre querem *generalizar seus achados* para toda a população. Querem, portanto, fazer uma *inferência*. Até que ponto os pesquisadores têm o direito de *generalizar*, para *todos* os indivíduos (a população), a informação obtida com base em *alguns* indivíduos (a amostra)? Para tomar uma decisão objetiva, os pesquisadores da área da saúde fazem *inferência estatística*.

¹QUINLIVAN, JA; EVANS, SF. The impact of continuing illegal drug use on teenage pregnancy outcomes. Australia: BJOG: An International Journal of Obstetrics & Gynaecology:109 (10):1148-53.2002.

Dizemos que uma *inferência estatística* é feita quando se estabelecem conclusões para a população com base nos dados de uma amostra e no resultado de um teste estatístico.

A inferência estatística é feita por meio de testes de hipóteses, mas, como toda inferência, está sujeita a *erro*. Os pesquisadores têm *apenas uma amostra* do imenso universo que é a população em estudo e — por puro azar — podem ter observado uma amostra *pouco representativa* da população de onde a amostra foi retirada. Quais são os tipos de erro?

- Erro tipo I: rejeitar a hipótese da nulidade quando essa hipótese é verdadeira.
- Erro tipo II: não rejeitar a hipótese da nulidade quando essa hipótese é falsa.

Exemplo 12.6: Definindo os erros.

Com base nos dados coletados e no resultado de um teste de hipóteses, as médicas devem decidir por uma das hipóteses. Quais são os erros possíveis?

Erro tipo I: rejeitar H_0 , quando H_0 é verdadeira: dizer que a probabilidade de filhos com baixo peso ao nascer é *maior* para mães usuárias de drogas ilícitas na gravidez, *se isso não for verdade*.

Erro tipo II: aceitar H_0 , quando H_0 é falsa: dizer que a probabilidade de filhos com baixo peso ao nascer é a *mesma*, para os dois grupos de mães, *se isso não for verdade*.

Os pesquisadores consideram grave o *erro de rejeitar a hipótese da nulidade quando ela é verdadeira*. Por quê? Porque isso significa mudar padrões e comportamentos sem necessidade (só porque um centro de pesquisas apontou como verdadeira uma diferença que não existe).

Exemplo 12.7: Erros tipo I.

- Dizer que uma nova droga é melhor que a tradicional, quando isso não for verdade.
- Dizer que uma dieta aumenta a longevidade, quando isso não for verdade.
- Dizer que um produto muito usado é cancerígeno, quando isso não for verdade.
- Dizer que uma vitamina faz atletas, quando isso não for verdade.

Para ter maior segurança na decisão, o pesquisador aplica um teste de hipóteses. O teste *não* elimina a probabilidade de erro, mas fornece o *p*-valor (valor de probabilidade).

O *p*-valor diz quão provável seria obter uma amostra tal qual a que foi obtida, quando a hipótese da nulidade é verdadeira.

Os pesquisadores se sentem seguros para rejeitar a hipótese da nulidade (assumir que existe a diferença procurada) quando o *p*-valor é pequeno². Isto porque seria muito pouco provável ter o resultado obtido, se a diferença não existisse. Mas quem *rejeita* H_0 não pode ter certeza absoluta (não tem 100% de confiança) de que a decisão tomada está correta — sabe, apenas, que a probabilidade de erro é pequena.

Por convenção, se o *p*-valor for menor do que 0,05 ($p < 0,05$), conclui-se que a hipótese da nulidade deve ser rejeitada. É comum dizer, nos casos em que $p < 0,05$, que os resultados são *estatisticamente significantes*.

No caso do Exemplo 12.4, as pesquisadoras não rejeitaram H_0 porque obtiveram *p*-valor maior do que 0,05 ($p > 0,05$). Concluíram³ então que *não tinham evidência suficiente* para dizer que baixo peso ao nascer *depende* de a mãe ter usado drogas ilícitas durante a gestação.

Exemplo 12.8: Interpretando o *p*-valor.

Imagine que uma enfermeira suspeita que gestantes muito jovens tenham maior probabilidade de ter filhos com baixo peso. Fez então um levantamento de dados na maternidade onde trabalha e obteve os dados. Distribuiu as mães em duas categorias: com menos de 20 anos e com 20 anos ou mais. Distribuiu, também, os recém-nascidos em duas categorias: de baixo peso e de peso normal. Obteve os dados apresentados na Tabela 12.1.

²Quando reduzimos a probabilidade de cometer um tipo de erro, aumentamos a probabilidade de cometer o outro tipo de erro. Como os pesquisadores consideram cometer erro tipo I “mais grave”, esse tipo de erro é reduzido, em geral, a 5%.

³As autoras concluíram que o uso de drogas ilícitas por gestantes parece não afetar o peso do nascituro, mas existem outros comprometimentos.

TABELA 12.1
Peso ao nascer segundo a faixa de idade da mãe.

<i>Faixa de idade materna</i>	<i>Peso ao nascer</i>		<i>Total</i>	<i>Percentual com baixo peso</i>
	<i>Menos de 2.500 g</i>	<i>2.500 g e mais</i>		
Menos de 20 anos	10	40	50	20,00%
20 anos ou mais	10	140	150	6,67%
Total	20	180	200	

A enfermeira levou, então, os dados a um estatístico para que ele fizesse a análise. O estatístico fez as hipóteses:

Hipótese da nulidade: A probabilidade de filhos com baixo peso é a mesma, para mães com menos de 20 anos e para mães com 20 anos ou mais.

Hipótese alternativa: A probabilidade de filhos com baixo peso depende da faixa etária da mãe.

Depois, fez um teste de qui-quadrado (que você aprende na seção 12.2.1) e informou à enfermeira que o *p*-valor é 0,0065.

A conclusão da enfermeira pode, então, ser escrita como segue: a probabilidade de filhos com baixo peso é显著mente maior para mães de menos de 20 anos.

Como você vê, feito o teste estatístico, a pesquisadora se sentiu segura para dizer que a diferença realmente existe.

Mas o que significa *p*-valor de 0,0175? Significa que se mães com menos de 20 anos e mães com 20 anos ou mais (as duas populações) tiverem a mesma probabilidade de ter filho com baixo peso ao nascer, somente 1,75% dos levantamentos similares aos que foram feitos mostrariam diferenças pelo menos tão grandes como a obtida, por puro acaso.

Calcular o *p*-valor é extremamente difícil e isso só é feito, hoje em dia, usando programas de computador. No entanto, não é difícil calcular a estatística do teste e comparar com valores dados em tabelas. Mas vamos ver isto na próxima seção.

12.1 – TESTE DE χ^2 DE PEARSON PARA ADERÊNCIA⁴

O teste de χ^2 proposto por Pearson tem indicação precisa: serve para testar a hipótese de que dados de freqüência se distribuem de acordo com alguma teoria ou postulado — é o teste de aderência, que veremos aqui; serve, também, para testar a hipótese de que duas variáveis nominais são independentes — é o teste de independência, que veremos na próxima seção.

Veja, então, o teste de aderência. Um pesquisador pode ter interesse em verificar se a distribuição dos elementos, numa dada amostra, está de acordo (adere) com uma dada teoria. O exemplo que será usado aqui é histórico porque se trata de um experimento feito por Gregor Mendel, o monge austríaco que, no final do século XIX, construiu as bases da Genética.

Em um célebre experimento Mendel polinizou 15 plantas de sementes lisas e alume amarelo com plantas de sementes rugosas e alume verde. As plantas resultantes desse cruzamento tinham sementes lisas e alume amarelo (amarelo-lisas). Cruzando essas plantas entre si, Mendel obteve 556 sementes, distribuídas conforme mostra a Tabela 12.2.

TABELA 12.2
Distribuição das ervilhas em um dos experimentos de Mendel.

<i>Sementes</i>	<i>Freqüência</i>
Amarelo-lisas	315
Amarelo-rugosas	101
Verde-lisas	108
Verde-rugosas	32
Total	556

Fonte: Bishop *et al.* (1975)⁵

A teoria postulada por Mendel estabelece que a segregação, neste caso, deve ocorrer na seguinte proporção:

$$\frac{9}{16} : \frac{3}{16} : \frac{3}{16} : \frac{1}{16}$$

⁴Leia-se teste de qui-quadrado para aderência. O símbolo χ é uma letra grega de nome *qui*, que equivale ao c do nosso alfabeto; lê-se qui; como está elevado à segunda potência, lê-se qui-quadrado.

⁵BISHOP, V.M.M. *et alii. Discrete multivariate analysis, theory and practice.* Cambridge, MIT Press, 1977.

Será que os resultados obtidos experimentalmente por Mendel estão de acordo com a teoria que ele postulava? Temos, então, as duas hipóteses:

- H_0 : a segregação obedece à lei de Mendel.
- H_1 : a segregação não obedece à lei de Mendel.

Para fazer o teste, os estatísticos usam um programa de computador que fornece, além do valor de χ^2 , o p -valor. Mas neste livro estamos fazendo os cálculos sem usar computador. Como é extremamente trabalhoso calcular o valor de p , vamos optar por usar as tabelas clássicas de χ^2 .

Para isso, é preciso estabelecer o nível de significância do teste. Mas o que é nível de significância?

Nível de significância do teste é a probabilidade de cometer erro tipo I, isto é, rejeitar H_0 quando H_0 é verdadeira. É usual indicar o nível de significância pela letra grega α (lê-se: alfa).

O nível de significância deve ser estabelecido antes do início do teste. Vamos, então, estabelecer $\alpha = 0,05$. Para verificar se os dados se distribuem de acordo com a teoria, vamos aplicar o teste de χ^2 . O valor de χ^2 é dado pela fórmula:

$$\chi^2 = \sum_{i=1}^r \frac{(O_i - E_i)^2}{E_i}$$

em que O_i ($i = 1, \dots, r$) representam as freqüências observadas e E_i representam as freqüências esperadas; r são as categorias da variável em análise que, no exemplo, são 4.

Foram obtidas 556 ervilhas. Então a *freqüência esperada*, pela teoria de Mendel, de amarelo-lisas é:

$$\frac{9}{16} \times 556 = 312,75$$

a *freqüência esperada* de amarelo-rugosas é:

$$\frac{3}{16} \times 556 = 104,25$$

a *freqüência esperada* de verde-lisas é:

$$\frac{3}{16} \times 556 = 104,25$$

e a freqüência esperada de verde-rugosas é:

$$\frac{1}{16} \times 556 = 34,75$$

Todos estes valores estão apresentados na Tabela 12.3.

TABELA 12.3
Distribuição dos valores esperados pela teoria de Mendel no experimento.

<i>Sementes</i>	<i>Freqüência</i>
Amarelo-lisas	312,75
Amarelo-rugosas	104,25
Verde-lisas	104,25
Verde-rugosas	34,75
Total	556,00

Compare a Tabela 12.2 com a Tabela 12.3. As diferenças entre as freqüências observadas e esperadas são, respectivamente:

$$315 - 312,75 = 2,25$$

$$101 - 104,25 = -3,25$$

$$108 - 104,25 = 3,75$$

$$32 - 34,75 = -2,75$$

Para verificar se a distribuição de freqüências observadas está de acordo com a teoria, vamos aplicar o teste de χ^2 :

$$\chi^2 = \sum_{i=1}^r \frac{(O_i - E_i)^2}{E_i}$$

Para o exemplo:

$$\chi^2 = \frac{2,25^2}{312,75} + \frac{(-3,25)^2}{104,25} + \frac{3,75^2}{104,25} + \frac{(-2,75)^2}{34,75} = 0,47$$

O valor calculado de qui-quadrado deve ser comparado com o valor da tabela de χ^2 ao nível de significância estabelecido e com $r - 1$ graus de liberdade. Então:

- Se o valor calculado da estatística for menor do que o valor crítico da tabela, não rejeite a hipótese da nulidade (H_0) ao nível estabelecido de significância.

- Se o valor calculado da estatística for *igual ou maior* do que o valor crítico da tabela, *rejeite* a hipótese da nulidade (H_0) em favor da alternativa (H_1) ao nível estabelecido de significância.

A Tabela de χ^2 é apresentada no final deste livro. Para entender como se usa essa tabela, observe a Tabela 12.4, que reproduz parte da Tabela de χ^2 do Apêndice. O valor de χ^2 com 3 graus de liberdade ao nível de significância de 5% está em negrito na Tabela 12.4.

TABELA 12.4**Tabela (parcial) de χ^2 segundo os graus de liberdade e o valor de α .**

<i>Graus de liberdade</i>	<i>Nível de significância</i>		
	<i>10%</i>	<i>5%</i>	<i>1%</i>
1	2,71	3,84	6,64
2	4,60	5,99	9,21
3	6,25	7,82	11,34
4	7,78	9,49	13,28
5	9,24	11,07	15,09

Para o exemplo que estamos desenvolvendo, o valor calculado de χ^2 foi 0,47. O valor dado na tabela de χ^2 , com $r - 1 = 4 - 1 = 3$ graus de liberdade e ao nível de 5% de significância é 7,82. Como o valor calculado ($\chi^2 = 0,47$) é menor do que o valor dado na tabela ($\chi^2 = 7,82$), *não se rejeita*, ao nível de significância de 5%, a hipótese de que a segregação ocorreu de acordo com a teoria.

12.1.1 – Resumo do procedimento

É importante saber que o teste estatístico não é uma prova — apenas indica que é muito provável que a *hipótese alternativa* seja *verdadeira*. As hipóteses são escritas de maneira que a hipótese da nulidade, colocada em teste, seja a hipótese em que o pesquisador não acredita. Para fazer o teste:

1. Defina H_0 e H_1 .
2. Escolha o valor de α .
3. Calcule o valor da estatística de teste.
4. Compare o valor calculado com o valor da tabela de valores críticos.
5. Se o valor calculado da estatística de teste for:
 - *menor* do que o valor crítico da tabela, *não rejeite* a hipótese da nulidade (H_0);

- igual ou maior do que o valor crítico da tabela, rejeite a hipótese da nulidade (H_0) em favor da alternativa (H_1).
6. Se você usou um programa de computador para fazer os cálculos, tem o p -valor. Se $p < 0,05$, rejeite a hipótese da nulidade em favor da alternativa.

Cabem aqui algumas observações sobre o nível de significância, que se indica pela letra grega α . É usual, ou tradicional, fazer testes ao nível de significância $\alpha = 5\%$ ou ao nível de significância $\alpha = 1\%$. Mas esses valores são arbitrários.

Quando se rejeita a hipótese da nulidade ao nível de significância de 5%, diz-se que o resultado “é significante”. Quando se rejeita a hipótese da nulidade ao nível de significância de 1%, diz-se que o resultado “é altamente significante”.

12.2 – TABELAS 2×2 (LÊ-SE TABELA DOIS POR DOIS)

12.2.1 – Teste de χ^2 para independência

Para estudar a efetividade de determinada droga no alívio da dor após a instrumentação endodôntica (tratamento de canal), um cirurgião-dentista fez um experimento. Antes do procedimento clínico, administrou dois comprimidos de placebo para 50 pacientes (grupo controle) e dois comprimidos da droga para 150 pacientes (grupo tratado). Os comprimidos foram acondicionados em envelopes codificados, para que o paciente não soubesse se estava recebendo a droga em teste para o alívio da dor, ou se estava recebendo placebo. Os dados estão na Tabela 12.5.

TABELA 12.5
Distribuição dos pacientes segundo o grupo e o relato sobre dor.

Grupo	Relato de dor		Total	Percentual de pacientes com dor
	Sim	Não		
Controle	10	40	50	20,0%
Tratado	15	135	150	10,0%
Total	25	175	200	

A Tabela 12.5 é uma tabela 2×2 porque apresenta duas variáveis, cada uma com duas categorias:

- Variável 1: grupo, com duas categorias: controle; tratado;
- Variável 2: relato de dor, com duas categorias: com dor; sem dor.

O pesquisador quer saber se essas variáveis são *independentes*, isto é, quer testar a hipótese da nulidade:

A probabilidade de relatar dor depois do tratamento *não depende* de o paciente ter recebido ou não a droga;

contra a hipótese alternativa:

a probabilidade de relatar dor depois do tratamento muda se o paciente tiver recebido a droga.

Vamos estabelecer o nível de significância $\alpha = 0,05$.

Para testar a hipótese de nulidade, isto é, a hipótese de que a probabilidade de relatar dor depois do tratamento não depende de o paciente ter recebido ou não a droga, aplica-se o teste de χ^2 . Mas é preciso conhecer a fórmula. Nesta seção será apresentar uma fórmula simplificada, que serve para testar a *hipótese de que duas variáveis nominais ou categorizadas são independentes*. No caso do exemplo que estamos desenvolvendo, temos duas variáveis categorizadas: grupo (tratado ou controle) e relato de dor (sim ou não).

Agora, veja a Tabela 12.6 que apresenta os valores literais no caso de uma tabela 2×2 , isto é, de uma tabela que apresenta duas variáveis categorizadas, indicadas aqui por X e Y . A variável X tem duas categorias, X_1 e X_2 ; a variável Y tem, também, duas categorias: Y_1 e Y_2 .

TABELA 12.6
Valores literais em uma tabela 2×2 .

Variável X	Variável Y		Total
	Y_1	Y_2	
X_1	a	b	$a + b$
X_2	c	d	$c + d$
Total	$a + c$	$b + d$	n

O valor de χ^2 é dado pela fórmula:

$$\chi^2 = \frac{(ad - bc)^2 n}{(a+b)(c+d)(a+c)(b+d)}$$

Nas tabelas 2×2 , como a Tabela 12.6, o valor de χ^2 está associado a 1 grau de liberdade, porque:

- você tem duas variáveis que, no caso da Tabela 12.6, são X e Y ;
- cada variável tem duas categorias;

- então você tem 1 grau de liberdade para cada variável;
- o valor de χ^2 está, então, associado a $1 \times 1 = 1$ grau de liberdade.

Para calcular o valor de χ^2 , verifique que, no exemplo que estamos desenvolvendo (veja a Tabela 12.5), temos os seguintes valores:

$$a = 10$$

$$b = 40$$

$$c = 15$$

$$d = 135$$

O valor de χ^2 é obtido como segue:

$$\begin{aligned}\chi^2 &= \frac{(10 \times 135 - 40 \times 15)^2 \times 200}{(10 + 40)(15 + 135)(10 + 15)(40 + 135)} \\ &= \frac{(1.350 - 600)^2 \times 200}{50 \times 150 \times 25 \times 175} \\ &= \frac{112.500.000}{32.812.500} = 3,429\end{aligned}$$

O valor de χ^2 é 3,429 e está associado a 1 grau de liberdade. Mas como você toma a decisão por uma das hipóteses, vendo o resultado do teste?

Toda vez que o valor calculado de χ^2 for igual ou maior do que o valor dado na Tabela de χ^2 , ao nível de significância estabelecido e com os mesmos graus de liberdade, rejeita-se H_0 .

Na Tabela de χ^2 no final do livro, para o nível de significância de 5% e com 1 grau de liberdade, encontra-se o valor 3,84. Como o valor calculado ($\chi^2 = 3,429$) é menor do que 3,84, não se rejeita a hipótese da nulidade. Portanto, a probabilidade de relatar dor depois do tratamento é a mesma para pacientes que receberam e não receberam a droga.

12.2.2 – Usos e restrições do teste de χ^2

Por questões teóricas⁶:

1. O teste de χ^2 só deve ser aplicado quando a amostra tem mais de 20 elementos.
2. Se $20 \leq n \leq 40$, o teste de χ^2 só pode ser aplicado se nenhuma frequência esperada for menor do que 1;
3. As variáveis devem ser nominais. Para variáveis ordinais, aplique o teste de χ^2 para tendências.
4. Existe uma correção — a correção de Yates — que torna o teste mais conservador⁶.

⁶Veja em: VIEIRA, S. *Bioestatística: Tópicos Avançados*. Rio de Janeiro, Campus-Elsevier, 2 ed. 5 tiragem. 2008.

12.2.3 – Medida da associação

Para medir o grau de associação de duas variáveis qualitativas, usam-se os *coeficientes de associação*. Nesta seção será explicado o coeficiente de Yule, que só se aplica às tabelas 2×2 . Para entender o que é uma associação entre variáveis, veja a Tabela 12.7.

TABELA 12.7
Participantes de uma pesquisa classificada segundo o hábito de fumar e doença periodontal.

<i>Participantes da pesquisa</i>	<i>Doença periodontal</i>		<i>Total</i>	<i>Proporção de pessoas com periodontite</i>
	<i>Não</i>	<i>Sim</i>		
Não-fumantes	18	6	24	$\frac{6}{24} = 0,250$
Fumantes	14	10	24	$\frac{10}{24} = 0,417$

A Tabela 12.7 mostra 24 fumantes e 24 não-fumantes. Também mostra a proporção de pessoas com doença periodontal (doença da gengiva, também conhecida como gengivite) em cada grupo:

- Não-fumantes : 0,250.
- Fumantes : 0,417.

A probabilidade da doença *aumenta* quando surge o *hábito de fumar*. Isto significa que existe *associação positiva* entre as variáveis (as duas aumentam juntas).

O coeficiente de Yule mede o *grau de associação* entre duas variáveis categorizadas. É indicado por Y e definido pela fórmula:

$$Y = \frac{ad - bc}{ad + bc}$$

O coeficiente de Yule varia entre -1 e $+1$, inclusive, isto é, $-1 \leq Y \leq +1$. Veja então como se interpreta o valor do coeficiente de associação:

- $Y = 1$: *associação perfeita positiva*
- $Y = -1$: *associação perfeita negativa*
- $Y = 0$: *associação nula*
- $0 < Y < 1$: *associação positiva*
- $-1 < Y < 0$: *associação negativa*

Para os dados da Tabela 12.7, o coeficiente de Yule é:

$$Y = \frac{18 \times 10 - 6 \times 14}{18 \times 10 + 6 \times 14} = \frac{96}{264} = 0,36$$

o que significa que a associação entre hábito de fumar e doença periodontal é positiva.

É importante observar que:

- O coeficiente de Associação de Yule mede o *grau de associação* entre duas variáveis nominais apresentadas numa tabela 2×2 .
- O teste de χ^2 estabelece se a associação entre duas variáveis nominais é *significante*, ou seja, se é muito provável que a *hipótese alternativa* (de associação) seja a *verdadeira*.
- Como são estatísticas diferentes — a primeira mede o grau de associação e a segunda a significância dessa associação — recomenda-se calcular as duas e, depois, discutir os resultados.

12.3 – EXERCÍCIOS RESOLVIDOS

12.3.1 – Você tem uma hipótese: determinada doença é genética e dominante. Espera-se então que metade dos filhos de pessoas com a doença, tenha também a doença. Como um teste preliminar para essa hipótese, você examina 40 filhos de pessoas doentes e encontra 14 deles com a doença. Você rejeita sua hipótese inicial?

Você espera que, em 40 filhos, 20 tenham a doença. É preciso comparar o que foi observado com o esperado, usando o teste de χ^2 para aderência. Veja a Tabela 12.8.

$$H_0: p = 0,50$$

$$H_1: p \neq 0,50$$

$$\alpha = 5\%$$

TABELA 12.8

Filhos de pais doentes, segundo o fato de terem a doença ou não.

<i>Doença</i>	<i>Número de filhos</i>		<i>O - E</i>	$(O - E)^2$
	<i>Observados (O)</i>	<i>Esperados (E)</i>		
Sim	14	20	-6	36
Não	26	20	6	36
Total	40	40	0	0

Aplicando a fórmula:

$$\chi^2 = \sum_{i=1}^r \frac{(O_i - E_i)^2}{E_i}$$

vem:

$$\chi^2 = \frac{36}{20} + \frac{36}{20} = 3,60$$

Na Tabela de χ^2 você encontra para 1 grau de liberdade e $\alpha = 5\%$, o valor 3,64. Como o valor calculado é menor do que o da tabela, não se rejeita a hipótese de que a doença é hereditária e de caráter dominante. As discrepâncias entre os valores observados e esperados são casuais.

12.3.2 – Com base nos dados apresentados na Tabela 12.9, teste a hipótese de que a proporção de recém-nascidos defeituosos é a mesma, qualquer que tenha sido a época em que a gestante foi atacada de rubéola. Faça $\alpha = 1\%$.

TABELA 12.9
Recém-nascidos segundo a época de ataque de rubéola na gestante e a condição.

<i>Época do ataque</i>	<i>Condição</i>		<i>Total</i>
	<i>Normal</i>	<i>Com defeito</i>	
Até o terceiro mês	36	14	50
Depois do terceiro mês	51	3	54
Total	87	17	104

Fonte: Hill et alii (1958)⁸

Hipótese da nulidade: A probabilidade de recém-nascidos defeituosos é a mesma, qualquer que tenha sido a época em que a gestante foi atacada de rubéola.

Hipótese alternativa: A probabilidade de recém-nascidos defeituosos depende da época em que a gestante foi atacada de rubéola.

Nível de significância: 1%.

⁸HILL, B. A. et alii. Virus diseases in pregnancy and congenital defects. Brit. J. Prev. Soc. Med., 12 (1):1958. Apud BERQUÓ, E. **Bioestatística**. São Paulo. Fac. Hig. Saúde Publ., USP, 1968.

Estatística de teste:

$$\chi^2 = \frac{(ad - bc)^2 n}{(a+b)(c+d)(a+c)(b+d)}$$

$$\chi^2 = \frac{(36 \times 3 - 14 \times 51)^2 \times 104}{(36 + 14)(51 + 3)(36 + 51)(14 + 3)}$$

$$= \frac{(108 - 714)^2 \times 104}{50 \times 54 \times 87 \times 17}$$

$$= \frac{38192544}{3993300} = 9,56$$

Na Tabela de χ^2 para $\alpha = 1\%$ e 1 grau de liberdade tem-se o valor 6,64. Como o valor calculado 9,56 é maior do que 6,64, conclui-se que a proporção de recém-nascidos com defeito é maior quando o ataque de rubéola na gestante ocorre nos três primeiros meses de gestação.

12.3.3 – Louis Pasteur conduziu uma série de experimentos em que mostrava o papel das leveduras e das bactérias na fermentação. Esses trabalhos deram a Joseph Lister⁹, um médico britânico, a idéia de que as infecções humanas poderiam ter origem similar. Ele então usou ácido fênico como desinfetante nas salas de cirurgia. Dos 40 pacientes amputados com uso de ácido fênico, 34 sobreviveram. Dos 35 amputados sem uso de ácido fênico, 19 sobreviveram. Escreva as hipóteses que podem ser colocadas em teste. Calcule as proporções de sobreviventes, com e sem uso de ácido fênico. Faça o teste de qui-quadrado, ao nível de 1% de significância.

Hipótese da nulidade: A probabilidade de sobrevivência em cirurgias de amputação é a mesma, quer se faça ou não desinfecção na sala cirúrgica.

Hipótese alternativa: A probabilidade de sobrevivência em cirurgias de amputação está associada à desinfecção da sala cirúrgica.

Nível de significância: 1%

⁹WINSLOW, C. The Conquest of Epidemic Diseases. Princeton: Princeton University Press. 1943. p. 303. Apud ALIAGA, M. e GUNDERSON, B. *Interactive Statistics*. 2 ed. New Jersey: Prentice Hall. 2003. p. 673.

TABELA 12.10
Sobrevivência de amputados com e sem uso de ácido fênico na sala cirúrgica.

<i>Sobrevivência</i>	<i>Ácido fênico</i>		<i>Total</i>	<i>Proporção de sobreviventes</i>
	<i>Sim</i>	<i>Não</i>		
Sim	34	6	40	0,850
Não	19	16	35	0,543
Total	53	22	75	

Fonte: Winslow (1943)

Estatística de teste:

$$\chi^2 = \frac{(ad - bc)^2 n}{(a+b)(c+d)(a+c)(b+d)}$$

$$\chi^2 = \frac{(34 \times 16 - 6 \times 19)^2 \times 75}{(34+6)(19+16)(34+19)(6+16)}$$

$$= \frac{(430)^2 \times 75}{(40)(35)(53)(22)}$$

$$\frac{(430)^2 \times 75}{(40)(35)(53)(22)} = \frac{13867500}{1632400} = 8,50$$

Para $\alpha = 1\%$ e 1 grau de liberdade tem-se, na Tabela de χ^2 o valor 6,64. Como o valor 8,50 é maior do que 6,64, rejeita-se H_0 ao nível de 1% de significância.

12.3.4 – O Estudo do Coração de Helsinque (Helsinki Heart Study)¹⁰ mostrou redução na incidência de eventos cardíacos em homens de meia-idade com nível alto de colesterol, mas sem diagnóstico de doença coronariana com o uso de uma droga (genfibrozila). Dos 2.051 participantes que durante cinco anos receberam a droga para reduzir o nível de colesterol, 56 registraram evento cardíaco. Dos 2.030 participantes que receberam placebo durante cinco anos, 84 registraram evento cardíaco.

- Qual é a proporção de participantes que registraram evento cardíaco no grupo tratado?**
- Qual é a proporção de participantes que registraram evento cardíaco no grupo placebo?**

¹⁰MARSHALL, K.G. Canadian Medical Association Journal. May, 15, 1996. Apud: ALIAGA, M. e GUNDERSON, B. *Interactive Statistics*, 2 ed. New Jersey: Prentice Hall. 2003. p. 679.

- c) Existe evidência suficiente do benefício da droga?
- d) No relatório final do estudo, afirmou-se que o uso da droga reduziu a incidência de eventos cardíacos em 34%. Como isso foi calculado?
- a) e b) Veja a Tabela 12.11.

TABELA 12.11
Participantes da pesquisa segundo o tratamento e o registro ou não de evento cardíaco.

<i>Tratamento</i>	<i>Evento cardíaco</i>		<i>Total</i>	<i>Proporção com registro de evento</i>
	<i>Sim</i>	<i>Não</i>		
Droga	56	1995	2051	0,0273
Placebo	84	1946	2030	0,0414
Total	140	3941	4081	

Fonte: Marshall (1996)

- c) É preciso fazer um teste estatístico. Então:

$$\begin{aligned} H_0: p_1 &= p_2 \\ H_1: p_1 &\neq p_2 \end{aligned}$$

Nível de significância: 5%

Calcule a estatística de teste:

$$\begin{aligned} \chi^2 &= \frac{(ad - bc)^2 n}{(a+b)(c+d)(a+c)(b+d)} \\ \chi^2 &= \frac{(56 \times 1946 - 1995 \times 84)^2 \times 4081}{(56 + 1995)(84 + 1946)(56 + 84)(1995 + 1946)} \\ &= \frac{(-58604)^2 \times 4081}{(2051)(2030)(140)(3941)} = 6,10 \end{aligned}$$

Rejeita-se H_0 ao nível de 5% de significância; temos, portanto, a evidência de que a droga teve o efeito.

- d) Faça a diferença entre as duas proporções e divida pela proporção do grupo que recebeu placebo. Multiplique por 100, para ter a diferença em relação ao placebo expressa em porcentagem.

$$\frac{0,0414 - 0,0273}{0,0414} \times 100 = 34\%$$

Então usar a droga reduziu em 34% a incidência de eventos cardíacos.

12.4 – EXERCÍCIOS PROPOSTOS

12.4.1 – A proporção de recém-nascidos com defeito ou doença séria é 3%. Imagine que um médico suspeita que esta proporção tenha aumentado. Examinou então 1.000 recém-nascidos e encontrou 34 com defeito ou doença séria. Você acha que a suspeita do médico é procedente?

12.4.2 – Com base nos dados apresentados na Tabela 12.12, teste, ao nível de significância de 5%, a hipótese de que a proporção de recém-nascidos vivos portadores de anomalia é a mesma nos dois sexos.

TABELA 12.12
Recém-nascidos vivos segundo o sexo e a presença ou não de anomalia.

<i>Sexo</i>	<i>Anomalia</i>	
	<i>Sim</i>	<i>Não</i>
Masculino	28	1.485
Feminino	45	1.406

12.4.3 – Com base nos dados apresentados na Tabela 12.13, teste, ao nível de significância de 1%, a hipótese de que a ausência congênita de dentes independe do sexo.

TABELA 12.13
Escolares segundo o sexo e a ausência congênita de dentes.

<i>Sexo</i>	<i>Ausência congênita de dentes</i>	
	<i>Sim</i>	<i>Não</i>
Masculino	23	1.078
Feminino	40	859

Fonte: Vedovelo Filho (1972)¹¹

12.4.4 – Muitos pesquisadores consideram, com base em grandes amostras, que a ausência congênita de dentes está associada ao sexo da pessoa. Amostras pequenas não permitem rejeitar H_0 . Isso se deve, provavelmente, à pequena associação. Calcule o coeficiente de associação de Yule para os dados do Exercício 12.7. Você considera grande a associação? Calcule as

¹¹VEDOVELO FILHO, M. Prevalência de agenesias dentárias em escolares de Piracicaba, 1972. [Tese (mestrado) FOP-INICAMP].

proporções. As diferenças são percentualmente grandes? Veja o Exercício 12.6.3 para calcular esse percentual.

12.4.5 – Com base nos dados apresentados na Tabela 12.14 calcule o coeficiente de associação. Faça o teste de qui-quadrado.

TABELA 12.14

Resultados de casos de diagnóstico pré-natal segundo a idade da gestante e a presença ou ausência de aberração cromossômica.

<i>Idade da gestante</i>	<i>Aberração cromossômica</i>	
	<i>Sim</i>	<i>Não</i>
De 35 até 40 anos	10	447
40 anos ou mais	18	510

12.4.6 – Para determinar se existe associação entre implantes mamários e doenças do tecido conjuntivo e outras doenças¹², foram observadas, durante vários anos, 749 mulheres que haviam recebido implante e exatamente o dobro de mulheres que não haviam recebido o implante. Eles verificaram que cinco mulheres que receberam implantes e 10 das que não receberam tiveram doenças do tecido conjuntivo. Quais são as hipóteses em teste? Quais são as proporções de mulheres doentes, nos dois grupos?

12.4.7 – Com base nos dados apresentados na Tabela 12.15, você rejeita a hipótese de que a probabilidade de natimorto é a mesma para os dois sexos?

TABELA 12.15

Recém-nascidos segundo o sexo e a condição de vivo ou natimorto.

<i>Sexo</i>	<i>Condição</i>	
	<i>Vivo</i>	<i>Natimorto</i>
Masculino	1.513	37
Feminino	1.451	27

12.4.8 – Com base nos dados apresentados na Tabela 12.16, ache o coeficiente de Yule. O que significa?

¹²GABRIEL SE et alii. Risk of connective tissues diseases and other disorders after breast implantation. *New Engl J Med* 330:1697-1702, 1994. Apud: Motulsky, H. *Intuitive Biostatistics*. Nova York, Oxford University Press, 1995. p.318.

TABELA 12.16
Recém-nascidos segundo a idade materna e o tempo de gestação.

<i>Idade materna</i>	<i>Tempo de gestação</i>		<i>Total</i>
	<i>Até 36 semanas</i>	<i>De 37 a 41 semanas</i>	
De 10 a 19 anos	612	1.378	1.990
De 20 a 34 anos	13.176	34.942	48.118
Total	13.788	36.320	50.108

Fonte: Azevedo et alii (2002)¹³

12.4.9 – Com base nos dados apresentados na Tabela 12.17, você rejeita a hipótese de que a probabilidade de dormir mais de 8 horas é a mesma para as duas faixas de idade?

TABELA 12.17
Participantes da pesquisa segundo o tempo de sono, em horas, e a faixa de idade.

<i>Faixa de idade</i>	<i>Tempo de sono</i>	
	<i>Menos de 8 horas</i>	<i>8 horas ou mais</i>
De 30 a 40 anos	172	78
De 60 a 70 anos	120	130

12.4.10 – Com base nos dados apresentados na Tabela 12.18, você rejeita a hipótese de que a probabilidade de ter gripe é a mesma para pessoas vacinadas e não-vacinadas?

TABELA 12.18
Participantes da pesquisa segundo o fato de ter sido vacinada contra gripe e ter tido gripe.

<i>Vacina</i>	<i>Gripe</i>	
	<i>Sim</i>	
Sim	11	538
Não	70	464

¹³AZEVEDO, G. D. et alii. Efeito da idade materna sobre os resultados perinatais. RBGO 24 (3): 2002.

(página deixada intencionalmente em branco)

Teste *t* de
Student

13

(página deixada intencionalmente em branco)

Os pesquisadores trabalham com *amostras*, mas suas conclusões devem ser generalizadas para *as populações de onde as amostras foram retiradas*, com base na aplicação de teste estatístico. Dizemos então que foi feita uma *inferência estatística*. Os testes estatísticos testam *hipóteses* a respeito da população.

O pesquisador faz duas hipóteses: a primeira é a *hipótese da nulidade* que, na grande maioria das vezes, afirma não existir diferença entre grupos de dados. Depois, o pesquisador constrói a *hipótese alternativa* que, como diz o próprio nome, contradiz a primeira. Então, ele aplica o teste estatístico para decidir por uma das hipóteses. Como isso é feito?

Os testes estatísticos fornecem o *p*-valor (valor de probabilidade) que permite decidir, com base nos dados, se há evidência suficiente para rejeitar a hipótese da nulidade. Por convenção, se o *p*-valor é menor do que 0,05 ($p < 0,05$), a hipótese da nulidade deve ser rejeitada¹. Em outras palavras, se $p < 0,05$, os resultados são *estatisticamente significantes*.

Neste Capítulo veremos como comparar *duas médias*² da mesma variável quantitativa, obtidas de dois grupos de dados, por meio de um teste estatístico.

Exemplo 13.1: Comparando duas médias.

Para verificar se meninos e meninas aprendem a falar na mesma idade, um pesquisador obteve, para um grande número de crianças, a idade em que cada uma delas começou a falar. A primeira hipótese — da *nulidade* — é a de que a média das idades em que os meninos começam a falar (meninos da população de onde a amostra foi retirada, não apenas os da amostra) é *igual* à média das idades em que as meninas começam a falar (meninas da população de onde a amostra foi retirada, não apenas as da amostra).

- H_0 : as médias são iguais

A segunda hipótese — *alternativa* — é a de que a média das idades em que os meninos começam a falar é *diferente* da média das idades em que as meninas começam a falar.

- H_1 : as médias são diferentes

¹O *p*-valor pequeno indica ser muito improvável obter resultado igual ou menor do que o achado quando a hipótese da nulidade é verdadeira.

²Para comparar mais de duas médias, aplicam-se a análise de variância e os testes de comparações múltiplas. Veja o assunto em: VIEIRA, S. **Análise de variância (ANOVA)**. São Paulo, Atlas, 2006.

Para comparar duas médias, o teste estatístico mais usado é o teste t de Student. Vamos ver como se faz este teste em duas situações diferentes:

1. quando os dados são pareados;
2. quando os grupos são independentes.

13.1 – O TESTE t NOS ESTUDOS COM DADOS PAREADOS

Muitas vezes, as unidades — físicas ou biológicas — são medidas duas vezes, no decorrer da pesquisa. A lógica é verificar se houve ou não discrepância entre as medições. Outras vezes, as unidades são consideradas aos pares. A idéia é verificar se há ou não diferença na resposta, ou no desempenho dos pares. A análise com dados pareados é apropriada nos seguintes casos:

- Quando se mede a mesma variável nas mesmas unidades, antes e depois de uma intervenção.
- Quando os participantes da pesquisa são recrutados aos pares, ou são pareados por idade, sexo, estágio da doença. Nesses casos, um dos participantes recebe a droga em teste e o outro participante recebe o tratamento convencional.
- Quando se mede a mesma variável em gêmeos ou em um par, como mãe e filho.
- Quando se faz um experimento em laboratório com várias repetições e em cada repetição se prepara, ao mesmo tempo, um controle e um teste.

Exemplo 13.2: Ensaio com dados pareados: duas medidas no mesmo indivíduo.

Para verificar se duas drogas diferentes, usadas como antitussígenos (bloqueadores de tosse) alteram o tempo de sono, foi feito um ensaio com nove voluntários. Eles tomaram um dos antitussígenos na primeira noite, e o outro na noite seguinte. Foi registrado o tempo de sono de cada voluntário, nas duas noites. A proposta é comparar as médias de tempo de sono sob o efeito de cada antitussígeno.

Exemplo 13.3: Ensaio com dados pareados: medidas feitas em pares de unidades.

Para verificar se uma droga é eficiente na inibição do crescimento de tumores, foram injetadas células cancerosas em 14 ratos similares. Depois os tumores foram medidos e foram formados pares de ratos com tumores de mesmo tamanho. Por sorteio, um rato de cada par recebeu a droga (grupo tratado), e o outro foi mantido como controle. A idéia é comparar as médias dos tamanhos de tumores de ratos tratados e ratos controles.

Quando temos dois grupos de dados pareados, aplicamos o teste t . Mas entenda: o pareamento deve ter algum tipo de lógica; não basta que os dois grupos tenham o mesmo número de unidades. Para fazer o teste t :

1. Estabeleça as hipóteses.

2. Escolha o nível de significância.

3. Siga os passos:

a) calcule as diferenças entre todas as observações pareadas:

$$d = x_2 - x_1$$

b) calcule a média dessas diferenças:

$$\bar{d} = \frac{\Sigma d}{n}$$

c) calcule a variância dessas diferenças:

$$s^2 = \frac{\Sigma d^2 - (\Sigma d)^2}{n-1}$$

d) calcule o valor de t , que está associado a $(n - 1)$ graus de liberdade, pela fórmula:

$$t = \frac{\bar{d}}{\sqrt{\frac{s^2}{n}}}$$

e) compare o *valor absoluto* do t calculado com o valor crítico dado na Tabela de valores de t , no nível estabelecido de significância e com os mesmos graus de liberdade. Toda vez que o *valor absoluto* do t calculado for igual ou maior que o valor crítico dado na tabela, rejeite a hipótese de que as médias são iguais, no nível estabelecido de significância.

Para entender como se acha o valor crítico de t , veja a Tabela 13.1, que reproduz parte da Tabela de valores de t , incluída no final deste livro. O valor crítico de t para, por exemplo, 4 graus de liberdade e 0,05 de significância está no cruzamento da linha 4 com a coluna 0,05. É 2,78, em negrito na Tabela 13.1.

TABELA 13.1
Tabela (parcial) de valores de *t*.

<i>Graus de liberdade</i>	<i>Nível de significância</i>		
	10%	5%	1%
1	6,31	12,71	63,66
2	2,92	4,30	9,92
3	2,35	3,18	5,84
4	2,13	2,78	4,60
5	2,02	2,57	4,03

Exemplo 13.4: Aplicando o teste *t* em ensaio com dados pareados.

Lembre o Exemplo 13.2. Para verificar se duas drogas diferentes, usadas como antitussígenos (bloqueadores de tosse), alteram o tempo de sono, foi feito um ensaio com nove voluntários. Os tempos de sono dos voluntários com cada droga estão na Tabela 13.2. As hipóteses em teste são:

- H_0 : o tempo médio de sono é o mesmo para as duas drogas.
- H_1 : as drogas determinam tempos médios de sono diferentes.
- Nível de significância: 0,05.

TABELA 13.2
Tempos de sono dos voluntários, em horas, segundo a droga.

<i>Voluntário</i>	<i>Droga</i>	
	<i>A</i>	<i>B</i>
1	7	9
2	7	7
3	6	6
4	6	8
5	9	10
6	6	8
7	7	7
8	8	8
9	5	7

Para fazer o teste:

- a) calcule as diferenças entre os tempos de sono com cada droga, para cada voluntário, conforme está apresentado na Tabela 13.3;

TABELA 13.3
Tempos de sono, em horas, segundo a droga e as respectivas diferenças.

Voluntário	Droga		Diferença
	A	B	
1	7	9	2
2	7	7	0
3	6	6	0
4	6	8	2
5	9	10	1
6	6	8	2
7	7	7	0
8	8	8	0
9	5	7	2

- b) calcule a média das diferenças:

$$\bar{d} = 1$$

- c) calcule a variância das diferenças:

$$s^2 = \frac{8}{9-1} = 1$$

- d) calcule o valor de t :

$$t = \frac{1}{\sqrt{\frac{1}{9}}} = 3$$

que tem $(n - 1) = (9 - 1) = 8$ graus de liberdade.

- e) compare o valor absoluto do t calculado com o valor crítico dado em Tabela de valores de t , no nível de significância de 0,05 e com 8 graus de liberdade. Como o valor absoluto do t calculado (3,00) é maior que o valor crítico (2,31), rejeite a hipótese de que o tempo de sono para as duas drogas é, em média, o mesmo, no nível de significância de 0,05.

Em termos práticos, em média, o tempo de sono quando se administra a droga B é显著mente diferente do tempo de sono com a droga A.

Se você fizer os cálculos em computador³, para o Exemplo 13.2 você obtém o *p*-valor 0,0171. A conclusão é a mesma.

13.1.1 – Testes unilaterais e testes bilaterais

A hipótese da *nulidade* sempre afirma: “não há diferença...” ou, então, “a diferença é *nula*”. No exemplo que acabamos de ver, a hipótese alternativa afirma: “existe diferença...”, mas não informa o sinal da diferença. Pode acontecer, porém, de o pesquisador ter noção do sinal da diferença e querer testar a hipótese da nulidade contra uma *hipótese alternativa que dê o sinal da diferença*. Se a hipótese alternativa especifica o sinal da diferença, dizemos que o teste é *unilateral*. Se a hipótese alternativa não especifica o sinal da diferença, dizemos que o teste é *bilateral*.

Exemplo 13.5: Teste unilateral.

Um professor quer saber se um curso de leitura dinâmica faz aumentar a velocidade de leitura dos alunos. Mede, então, a velocidade de leitura de 22 alunos que se dispuseram a participar da pesquisa. Depois ministra um curso de leitura dinâmica e, novamente, mede a velocidade de leitura desses alunos. Quais são as hipóteses em teste?

A hipótese da nulidade é a de que, em média, a velocidade de leitura é a mesma, antes e depois do curso.

A hipótese alternativa é a de que, em média, a velocidade de leitura depois do curso é maior.

É sempre mais seguro⁴ aplicar um *teste bilateral* — aquele em que você tanto pode concluir por um aumento como uma diminuição da medida, depois da intervenção. Afinal de contas, o tratamento pode dar resultado contrário ao esperado.

Exemplo 13.6: Teste unilateral ou bilateral.

Um nutricionista quer saber se determinada dieta alimentar leva a uma diminuição de peso. Submete então 20 voluntários a essa dieta, durante um mês. Quais são as hipóteses em teste?

A hipótese da nulidade é a de que, em média, a peso das pessoas é o mesmo, antes e depois da dieta. Quanto à hipótese alternativa, é mais seguro que seja a de que os pesos antes e depois da dieta são, em média, *diferentes*. Isto porque — qual-

³É muito complicado calcular o *p*-valor, razão por que não se fornece, aqui, nenhuma fórmula de cálculo.

⁴Existem muitas razões que determinam a preferência dos estatísticos por testes bilaterais. Uma delas é o fato de eles serem mais conservadores — têm menor probabilidade de rejeitar H_0 .

quer que seja a área de conhecimentos — alguns tratamentos têm, às vezes, efeito contrário ao esperado. No caso deste exemplo, um teste bilateral estaria considerando a possibilidade de a dieta levar a aumento de peso. Mas não seria errado proceder a um teste unilateral, se houver informações de pesquisas anteriores informando que a dieta deve determinar diminuição de peso.

A questão, agora, é saber *como* se faz um teste unilateral. O procedimento é o mesmo. Muda apenas a maneira de procurar o valor crítico, na Tabela de valores de t . Para um teste unilateral ao nível de 0,05 de significância e com $n - 1$ graus de liberdade, você procura o valor crítico de t com os mesmos graus de liberdade, mas com o *dobro* do nível de significância, isto é, procure $\alpha = 0,10$.

Exemplo 13.7: Ensaio com dados pareados: teste t, unilateral.

Uma droga é tradicionalmente usada para alívio de dor nos casos de enxaqueca. Uma empresa oferece um genérico. Para verificar se o efeito do genérico não é significantemente inferior, foi feito um ensaio com sete voluntários⁵. Todos os voluntários usaram, em períodos distintos, tanto a droga tradicional como o genérico. Os tempos de alívio da dor registrados pelos voluntários com cada droga estão na Tabela 13.4.

- H_0 : o tempo médio de alívio da dor é o mesmo, para as duas drogas.
- H_1 : o tempo médio de alívio da dor é menor, quando se administra o genérico.
- Nível de significância de 5%

TABELA 13.4
Tempos de alívio da dor, em horas, segundo a droga.

<i>Voluntário</i>	<i>Droga</i>	
	<i>Tradicional</i>	<i>Genérico</i>
1	4,5	4
2	5,5	5,5
3	6	6
4	6	5
5	5,5	4,5
6	5,5	6
7	8	6,5

⁵Este tipo de teste é conhecido como de não-inferioridade. O número de voluntários deve estar em torno de 25.

Para fazer o teste, calcule as diferenças entre os tempos obtidos com a droga tradicional e o genérico, conforme está apresentado na Tabela 13.5;

TABELA 13.5
Tempos de alívio da dor, em horas, segundo a droga e as respectivas diferenças.

<i>Voluntário</i>	<i>Droga</i>		
	<i>Tradicional</i>	<i>Genérico</i>	<i>Diferença</i>
1	4,5	4	-0,5
2	5,5	5,5	0
3	6	6	0
4	6	5	-1
5	5,5	4,5	-1
6	5,5	6	0,5
7	8	6,5	-1,5

Fazendo os cálculos, você acha a média das diferenças, que é -0,5 e a variância das diferenças, que é 0,5. Aplicando a fórmula para calcular o valor de t quando os dados são pareados, você obtém:

$$t = \frac{\bar{d}}{\sqrt{\frac{s^2}{n}}}$$

$$t = -\frac{0,5}{\sqrt{\frac{0,5}{7}}} = -1,871$$

No nível de significância de 5% para um teste unilateral e com 6 graus de liberdade, o valor de t , na Tabela de valores de t , é 1,94 (leia na coluna de 10%). Como, considerando a hipótese alternativa, o valor calculado de t deve ser menor do que zero, adote o seguinte critério para decisão: se t calculado for menor do que o valor negativo do t crítico da tabela de valores de t , rejeite H_0 . Neste exemplo, o valor calculado de t (-1,871) é maior que o valor negativo do t crítico (-1,94). Então não rejeite a hipótese de que o tempo de alívio da dor é, em média, o mesmo, para a droga tradicional e o genérico.

Em termos práticos, não há evidência estatística de que o tempo de alívio da dor seja menor quando se usa o genérico. O p -valor é $0,0553 > 0,05$.

13.2 – O TESTE *t* NA COMPARAÇÃO DE DOIS GRUPOS INDEPENDENTES

Muitas vezes os pesquisadores querem comparar *dois grupos independentes*. Podem comparar, por exemplo, o novo tratamento contra o *controle* ou, então, comparar dois tratamentos conhecidos.

Exemplo 13.8

Para saber se determinado produto faz nascer cabelos em pessoas calvas um médico pode fazer um ensaio clínico: um grupo de pessoas calvas recebe o tratamento em teste — *grupo tratado* — enquanto um grupo de pessoas calvas recebe um placebo — *grupo controle*.

O teste *t* de Student é indicado para testar a igualdade de *duas* médias quando *os grupos são independentes*. Para calcular o valor de *t*, siga os passos:

- calcule a média de cada grupo;
- calcule a variância de cada grupo;
- calcule a variância ponderada, dada pela fórmula:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

- calcule o valor de *t*, que está associado a $n_1 + n_2 - 2$ graus de liberdade, pela fórmula:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)s_p^2}}$$

- compare o *valor calculado de t* (em valor absoluto) com o *valor crítico de t*, no nível estabelecido de significância e com os mesmos graus de liberdade. No caso de teste bilateral, se o valor absoluto do *t* calculado for igual ou maior do que o da tabela, rejeite a hipótese de que as médias são iguais, no nível estabelecido de significância.

Exemplo 13.9: Teste *t* para comparar dois grupos — bilateral.

Um nutricionista quer comparar o efeito de duas dietas alimentares para perda de peso. Seleciona então voluntários que querem perder peso e os divide, ao acaso, em dois grupos: um grupo é designado para a dieta A, e o outro para a dieta B. Os dados estão na Tabela 13.6. Faça o teste *t*, ao nível de 5% de significância.

TABELA 13.6
Perda de peso, em quilogramas, segundo a dieta.

<i>Dieta</i>	
<i>A</i>	<i>B</i>
12	15
8	19
15	15
13	12
10	13
12	16
14	15
11	
12	
13	

Para o exemplo apresentado neste capítulo, veja como se faz o teste *t*.

- H_0 : as perdas de peso são, em média, as mesmas, para qualquer das duas dietas.
- H_1 : as dietas determinam perdas médias de peso diferentes.
- Nível de significância: 0,05.

a) as médias de grupos são:

$$\bar{x}_1 = 12$$

$$\bar{x}_2 = 15$$

b) as variâncias de grupo são:

$$s_1^2 = 4,0$$

$$s_2^2 = 5,0$$

c) a variância ponderada é:

$$s^2 = \frac{(10-1) \times 4,0 + (7-1) \times 5,0}{10+7-2} = 4,4$$

d) o valor de *t* com $n_1 + n_2 - 2 = 10 + 7 - 2 = 15$ graus liberdade é:

$$t = \frac{15-12}{\sqrt{\left(\frac{1}{10} + \frac{1}{7}\right)4,4}} = 2,902$$

- e) como o valor calculado de t (em valor absoluto) é maior que o valor crítico de t ($2,902 > 2,13$) no nível de 5% de significância, você rejeita a hipótese de que as duas dietas determinam, em média, a mesma perda de peso.

Em termos práticos, o nutricionista pode concluir que as perdas de peso são, em média, significantemente maiores quando os voluntários são submetidos à dieta B. O p -valor, neste exemplo, é $0,0109 < 0,05$.

13.2.1 – O caso das variâncias desiguais

O teste t , tal como foi apresentado, só deve ser aplicado quando as variâncias das populações são iguais. Mas o que deve ser feito para saber se as variâncias das populações são iguais? Existe uma regra prática: compare-se as variâncias das duas amostras; se a maior variância for até quatro vezes a menor, admite-se que as duas populações têm variâncias iguais.

Exemplo 13.10: Comparação de variâncias — regra prática.

Imagine duas amostras, 1 e 2, com variâncias $s_1^2 = 15,64$ e $s_2^2 = 6,80$, respectivamente. Como:

$$\frac{s_1^2}{s_2^2} = \frac{15,64}{6,80} = 2,30 < 4,$$

é razoável admitir que as variâncias são iguais. Mas é melhor aplicar um teste estatístico.

Para testar a hipótese de que as variâncias das duas populações são iguais, aplica-se o teste F . Para fazer um teste unilateral:

1. Estabeleça as hipóteses.
 - H_0 : as variâncias na população são iguais.
 - H_1 : uma das variâncias é maior do que a outra.
2. Escolha o nível de significância.
3. Siga os passos:
 - a) Calcule a variância de cada grupo:
 - s_1^2 : variância do grupo 1
 - s_2^2 : variância do grupo 2

- b) Calcule o valor de F , dado pela razão entre a maior e a menor variância. Então, se $s_1^2 > s_2^2$, o valor

$$F = \frac{s_1^2}{s_2^2}$$

está associado a $n_1 - 1$ (numerador) e $n_2 - 1$ (denominador) graus de liberdade.

- c) Para o teste unilateral, compare o valor calculado de F com o valor dado na Tabela de valores F , com o nível de significância estabelecido e com $(n_1 - 1)$ e $(n_2 - 1)$ graus de liberdade. Para um teste bilateral, que é mais indicado, faça os cálculos da mesma maneira, mas procure, na Tabela de valores de F , o valor crítico com os mesmos graus de liberdade, mas com a metade do nível estabelecido de significância. Rejeite a hipótese de que as variâncias das duas populações são iguais toda vez que o valor calculado de F for igual ou maior do que o valor da tabela de valores F .

Para entender como se acha o valor de F na tabela, observe a Tabela 13.7 que reproduz parte dessa tabela, apresentada no final deste livro. Foi colocado em negrito o valor de F no nível de significância de 2,5% e com 7 e 8 graus de liberdade, que deve ser utilizado para um teste bilateral na forma descrita aqui, com nível de significância de 5% e com os mesmos graus de liberdade.

TABELA 13.7
Tabela (parcial) de valores de F para $\alpha = 2,5\%$.

<i>Número de graus de liberdade do denominador</i>	<i>Número de graus de liberdade do numerador</i>								
	1	2	3	4	5	6	7	8	9
1	648,0	800,0	864,0	900,0	922,0	937,0	948,0	957,0	963,0
2	38,5	39,0	39,2	39,2	39,3	39,3	39,4	39,4	39,4
3	17,4	16,0	15,4	15,1	14,9	14,7	14,6	14,5	14,5
4	12,2	10,6	9,98	9,60	9,36	9,20	9,07	8,98	8,90
5	10,0	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68
6	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52
7	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82
8	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36
9	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03

Se as variâncias são diferentes, para comparar duas médias aplica-se o teste t , na forma descrita aqui. É preciso calcular:

- a) a média de cada grupo. Indica-se:

\bar{x}_1 : média do grupo 1

\bar{x}_2 : média do grupo 2

- b) a variância de cada grupo. Indica-se:

s_1^2 : variância do grupo 1

s_2^2 : variância do grupo 2

- c) o valor de t , dado pela fórmula:

$$t = \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

onde n_1 é o número de elementos do grupo 1 e n_2 é o número de elementos do grupo 2.

- d) o número de graus de liberdade associado ao valor de t , que é a parte inteira do número g , obtido pela fórmula:

$$g = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\left(\frac{s_1^2}{n_1} \right)^2 + \left(\frac{s_2^2}{n_2} \right)^2}$$

$$n_1 - 1 \quad n_2 - 1$$

- e) feitos os cálculos, é preciso procurar o valor de t na tabela de valores de t , no nível estabelecido de significância e com g graus de liberdade. Toda vez que o valor absoluto de t calculado for igual ou maior do que o valor de t dado na tabela conclui-se que, no nível estabelecido de significância, as médias não são iguais.

Exemplo 13.11: Teste t para comparar dois grupos — variâncias diferentes.

Para verificar se determinada dieta leva à perda de peso, um médico separou, ao acaso, um conjunto de pacientes em dois grupos: um grupo foi submetido à dieta (grupo tratado), enquanto o outro manteve os mesmos hábitos alimentares (grupo controle). Decorrido determinado período de tempo, o médico obteve a perda de peso de cada paciente, em cada grupo. Os valores estão na Tabela 13.8.

TABELA 13.8**Perdas de peso em quilogramas de pacientes segundo o grupo.**

<i>Grupo</i>	
<i>Tratado</i>	<i>Controle</i>
12	1
14	0
12	0
9	1
14	0,5
14	1
9	0

Para proceder ao teste, é preciso, primeiro, estabelecer o nível de significância. Seja $\alpha = 5\%$. Depois é preciso calcular:

a) a média de cada grupo:

$$\bar{x}_1 = \frac{12 + 14 + \dots + 9}{7} = 12$$

$$\bar{x}_2 = \frac{1 + 0 + \dots + 0}{7} = 0,5$$

b) a variância de cada grupo:

$$s_1^2 = \frac{1.038 - \frac{(84)^2}{7}}{6} = 5,00$$

$$s_2^2 = \frac{3,25 - \frac{(3,5)^2}{7}}{6} = 0,25$$

c) o valor de F , porque, como as variâncias são muito diferentes, convém fazer o teste. Seja $\alpha = 5\%$.

$$F = \frac{s_1^2}{s_2^2} = \frac{5}{0,25} = 20,00$$

O valor calculado de F está associado a 6 (numerador) e 6 (denominador) graus de liberdade. A Tabela de valores F (veja no final do livro) fornece para $\alpha = 2,5\%$ com 6 e 6 graus de liberdade o valor $F = 5,82$. Então rejeita-se a hipótese de que as variâncias são iguais no nível de significância de 5%. Agora é preciso calcular:

d) o valor de *t*:

$$t = \frac{0,5 - 12}{\sqrt{\frac{5,0}{7} + \frac{0,25}{7}}}$$

$$t = \frac{-11,5}{\sqrt{\frac{5,25}{7}}} = -13,28$$

e) o número de graus de liberdade:

$$g = \frac{\left(\frac{5,0}{7} + \frac{0,25}{7}\right)^2}{\frac{\left(\frac{5,0}{7}\right)^2}{6} + \frac{\left(\frac{0,25}{7}\right)^2}{6}}$$

$$= \frac{0,5625}{0,085247} = 6,6$$

O valor calculado de *t* está associado a aproximadamente 6 graus de liberdade. Como o valor de *t* na Tabela de valores *t* (veja no final do livro), no nível de significância de 5% e com 6 graus de liberdade, é 2,45, rejeita-se a hipótese de que as médias são iguais. Em termos práticos, a perda de peso foi, em média, significativamente maior no grupo submetido à dieta.

13.3 – O TESTE *t* PARA O COEFICIENTE DE CORRELAÇÃO

O teste *t*, apresentado neste Capítulo, tem outros usos além da comparação de médias. Pode ser usado, por exemplo, para testar a hipótese de que o coeficiente de correlação entre duas variáveis é igual a zero, contra a hipótese de que é diferente de zero.

Reveja o ítem 6.2 do Capítulo 6. O coeficiente de correlação varia entre -1 e +1. Se o coeficiente de correlação entre duas variáveis for igual a zero, não existe correlação linear entre elas. E se o coeficiente calculado for $r = 0,775$? Não se pode julgar o valor desse coeficiente sem saber o tamanho da amostra. Quando a amostra é muito pequena, coeficientes de correlação com valores altos podem não ter significado estatístico.

Exemplo 13.12: Teste t para coeficiente de correlação.

O coeficiente de correlação entre duas variáveis X e Y , calculado com base em uma amostra de tamanho 14, é $r = -0,775$. Esse valor é estatisticamente significante?

Para aplicar o teste t , usa-se a fórmula:

$$t = \frac{r}{\sqrt{1-r^2}} \sqrt{n-2}$$

onde r é o valor calculado para o coeficiente de correlação e n é o tamanho da amostra. Esse valor de t está associado a $n-2$ graus de liberdade. No caso do exemplo, $r = -0,775$ e $n = 14$. Portanto:

$$t = \frac{-0,775}{\sqrt{1-0,601}} \sqrt{14-2} = \frac{-0,775}{0,632} \times 3,46 = -4,25$$

com $n-2 = 12$ graus de liberdade.

No nível de significância de 5% a Tabela de valores t (veja no final do livro) fornece, para 12 graus de liberdade, o valor $t = 2,18$. Como o valor calculado de t é, em valor absoluto, maior do que 2,18, a correlação entre as variáveis é significante no nível de 5%.

13.4 – EXERCÍCIOS RESOLVIDOS

13.4.1 – Os valores apresentados na Tabela 13.9 permitem testar a hipótese de que recém-nascidos de ambos os sexos têm, em média, a mesma estatura. Teste essa hipótese, no nível de significância de 5%.

TABELA 13.9

Tamanho da amostra, média e variância da estatura, em centímetros, de recém-nascidos, segundo o sexo.

Sexo	n	\bar{x}	s^2
Masculino	1.442	49,29	5,76
Feminino	1.361	48,54	6,30

Antes de proceder ao teste t , convém testar a igualdade das variâncias. Para isso, calcule:

$$F = \frac{6,30}{5,76} = 1,09$$

que está associado a 1.360 (numerador) e 1.441 (denominador) graus de liberdade. Para um teste bilateral no nível de significância de 5%, você deve comparar o valor calculado de F com o valor crítico de F dado na Tabela

de valores de F com $\alpha = 2,5\%$, com 1.360 e 1.441 graus de liberdade. A tabela não tem esses números de graus de liberdade, mas como os números são muito grandes, use o valor de F associado a infinitos graus de liberdade, tanto para numerador como para denominador. Esse valor é 1,00. O valor calculado de F é maior do que 1,00. Portanto, no nível de significância de 5%, as variâncias são diferentes.

O teste t — no caso de variâncias desiguais — deve ser calculado como segue:

$$t = \frac{49,29 - 48,54}{\sqrt{\frac{5,76}{1.442} + \frac{6,30}{1.361}}} = 8,076$$

que está associado aos graus de liberdade:

$$g = \frac{\left(\frac{5,76}{1.442} + \frac{6,30}{1.361} \right)^2}{\frac{\left(\frac{5,76}{1.442} \right)^2}{1.441} + \frac{\left(\frac{6,30}{1.361} \right)^2}{1.360}} = 2.772$$

O valor calculado de t é maior do que o valor dado na Tabela de valores t (veja Apêndice). Rejeite, então, no nível de significância de 5%, a hipótese de que recém-nascidos de ambos os sexos têm, em média, a mesma estatura. Em termos práticos, os meninos nascem com estatura maior do que as meninas.

13.4.2 – Com base nos dados apresentados na Tabela 13.10 teste, no nível de significância de 5%, a hipótese de que o calibre da veia esplênica é, em média, o mesmo, antes e após a oclusão da veia porta.

TABELA 13.10
Calibre da veia esplênica em seis cães antes e após a oclusão da veia porta.

<i>Número do cão</i>	<i>Oclusão da veia porta</i>	
	<i>Antes</i>	<i>Depois</i>
1	75	85
2	50	75
3	50	70
4	60	65
5	50	60
6	70	90

Note que foram tomadas duas medidas do calibre da veia esplênica em cada cão: uma antes, outra após a oclusão da veia porta. Para aplicar o teste t é preciso calcular a diferença observada em cada animal. Tais diferenças estão na Tabela 13.11.

TABELA 13.11
Diferenças de calibre da veia esplênica antes e depois a oclusão da veia porta.

<i>Número do cão</i>	<i>Oclusão da veia porta</i>		
	<i>Antes</i>	<i>Depois</i>	<i>Diferença</i>
1	75	85	10
2	50	75	25
3	50	70	20
4	60	65	5
5	50	60	10
6	70	90	20

A média das diferenças é:

$$\bar{d} = 15,0$$

e a variância é:

$$s^2 = 60,00.$$

O valor de t , associado a 5 graus de liberdade, é:

$$t = \frac{15,0}{\sqrt{\frac{60,00}{6}}} = 4,74$$

Na tabela de t , para $\alpha = 5\%$ e com 5 graus de liberdade, está o valor 2,57. Como o valor calculado de t é maior do que o da tabela no nível estabelecido de significância, a hipótese de que, em média, o calibre da veia esplênica é o mesmo, antes e depois da oclusão da veia porta, deve ser rejeitada. Em termos práticos, a oclusão da veia porta determina aumento significativo do calibre da veia esplênica.

13.4.3 – Reveja o Exemplo 5.6.11: um professor de Odontologia quer saber se alunos que começam a atender pacientes em disciplinas clínicas têm aumento na pressão sistólica. Mediú então a pressão sistólica de cinco alunos de primeiro ano (que não cursam disciplinas clínicas) e de cinco alunos do segundo ano, logo antes do primeiro atendimento de pacientes. Os

dados foram apresentados na Tabela 5.12 do Capítulo 5. Você calculou as médias e os desvios padrões. Aplique agora um teste t unilateral.

Você já calculou:

1º ano: média = 118,0; desvio padrão = 4,12.

2º ano: média = 131,0; desvio padrão = 8,66.

Faça o teste das variâncias: $F = 4,41$, não significante no nível de 5% (F crítico = 9,60) (p -valor = 0,1796). O teste t unilateral fornece $t = -3,03$, significante ao nível de 5% (t crítico = 1,86) (p -valor = 0,0082). Com base neste resultado, é razoável concluir que alunos que começam a atender pacientes em disciplinas clínicas têm aumento significante na pressão sistólica ($p < 0,05$).

13.4.4 – Um nutricionista⁶ quer saber se existe diferença entre iogurtes feitos de leite desnatado quando se adiciona (ou não) determinada bactéria. Para isso, procura amostras de leite desnatado de sete marcas comerciais diferentes. Inocula então metade da amostra de cada marca com a bactéria e a outra metade deixa sem a bactéria, para servir como controle. Depois de prontos os iogurtes, o nutricionista mede a firmeza da massa. Os dados estão apresentados na Tabela 13.12. Faça o teste.

TABELA 13.12
Firmeza da massa de iogurte, segundo a marca e a presença ou não de bactéria.

<i>Marca</i>	<i>Bactéria</i>	
	<i>Sim</i>	<i>Não</i>
A	68	61
B	75	69
C	62	64
D	86	76
E	52	52
F	46	38
G	72	68

- H_0 : a firmeza do iogurte é, em média, a mesma, com ou sem adição de bactéria.
- H_1 : a adição de bactéria muda a média da firmeza do iogurte.
- Nível de significância: 0,05.

⁶JOHNSON, R. E TSUI, K. W. *Statistical reasoning and methods*. Nova York, Wiley, 1998. p. 437.

Os resultados estão apresentados na Tabela 13.13. O valor para t é significante. Portanto, há evidência de que a bactéria modifica a firmeza do iogurte.

TABELA 13.13
Médias, desvios padrões, valor de t para firmeza da massa de iogurte.

Bactéria	Média	Desvio padrão	Teste t	p-valor
Presente	65,9	13,7		
Ausente	61,1	12,6		
Diferença	4,71	4,35	2,87	0,0285

13.5 – EXERCÍCIOS PROPOSTOS

13.5.1 – *Dez ratos machos adultos, criados em laboratório, foram separados aleatoriamente em dois grupos: um grupo foi tratado com a ração normalmente usada no laboratório, e o outro grupo foi submetido a uma nova ração (experimental). Decorrido certo período de tempo, pesaram-se os ratos. Os pesos estão apresentados na Tabela 13.14. Teste a hipótese de que o peso médio dos ratos é o mesmo, para os dois tipos de ração.*

TABELA 13.14
Pesos em gramas de ratos adultos, segundo a ração.

	Ração	
	Padrão	Experimental
	200	220
	180	200
	190	210
	190	220
	180	210

13.5.2 – *Os quocientes de inteligência (QI) de 10 crianças, medidos segundo dois testes de inteligência, A e B, estão apresentados na Tabela 13.15. Verifique, através do teste t, se os dois testes de inteligência dão, em média, o mesmo valor.*

TABELA 13.15
Valores de QI em 10 crianças, segundo o teste de inteligência aplicado.

<i>Teste</i>	
<i>A</i>	<i>B</i>
100	105
105	108
98	102
101	103
100	100
108	110
98	106
100	100
99	103
99	103

13.5.3 – A Tabela 13.16 apresenta dados de pressão sanguínea sistólica de mulheres na faixa etária de 30 a 35 anos, que usavam e que não usavam anticoncepcionais orais. Teste a hipótese de que o uso de anticoncepcionais não tem efeito sobre a pressão sanguínea sistólica.

TABELA 13.16
Pressão sanguínea sistólica de mulheres de 30 a 35 anos segundo o uso de anticoncepcionais.

<i>Uso de anticoncepcionais</i>	
<i>Sim</i>	<i>Não</i>
111	109
119	113
121	120
113	117
116	108
126	120
128	122
123	124
122	115
121	112

13.5.4 – A Tabela 13.17 apresenta o tamanho da amostra, a média e a variância dos pesos ao nascer de nascidos vivos de ambos os sexos. Teste, ao nível de significância de 1%, a hipótese de que os dois sexos têm, em média, o mesmo peso ao nascer.

TABELA 13.17

Tamanho da amostra, média e variância de pesos ao nascer de nascidos vivos, segundo o sexo.

Sexo	n	\bar{x}	s^2
Masculino	14	3,253	0,261
Feminino	13	3,130	0,265

13.5.5 – Para mais bem conhecer o efeito do frio⁷, pesquisadores fizeram um experimento com ratos de laboratório. Doze ratos foram divididos ao acaso em dois grupos. Um grupo ficou, durante 12 horas, na temperatura de 26° C e o outro grupo ficou numa temperatura de 5°C, pelo mesmo tempo. Depois os pesquisadores mediram a pressão sanguínea dos 12 ratos. Os resultados estão na Tabela 13.18. O que você conclui?

TABELA 13.18

Pressão sanguínea dos ratos segundo a temperatura a que foram submetidos.

	Temperatura	
	5°C	26°C
	152	384
	157	369
	179	354
	182	375
	176	366
	149	423

13.5.6 – Para comparar o tempo de absorção de duas drogas, A e B, nove pessoas foram designadas ao acaso para receber a droga A e sete para receber a droga B. Depois se determinou o tempo que demorou até as drogas alcançarem determinado nível no sangue. Com base nas estatísticas apresentadas na Tabela 13.19, faça o teste t.

⁷OTT, L e MENDENHALL, W. *Understanding Statistics*. Belmont, Wadsworth. 6 ed. 1994. p. 305.

TABELA 13.19

Médias e variâncias do tempo despendido para as drogas alcançarem determinado nível no sangue.

<i>Estatísticas</i>	<i>Droga</i>	
	<i>A</i>	<i>B</i>
Número de pessoas	9	7
Média	27,2	33,5
Variância	16,36	18,92

13.5.7 – Para saber se o tempo de alívio da dor no pós-operatório é significantemente maior quando se administra a droga A em lugar da droga B, mais comumente usada, observou-se o tempo do alívio da dor de 25 pessoas que receberam a droga A no pós-operatório e 20 que receberam a droga B. Com base nas estatísticas apresentadas na Tabela 13.20, faça o teste t.

TABELA 13.20

Médias e variâncias do tempo de alívio da dor, segundo a droga.

<i>Estatísticas</i>	<i>Droga</i>	
	<i>A</i>	<i>B</i>
Número de pacientes	25	20
Média	5,5	5,0
Variância	2,25	1,69

13.5.8 – Acredita-se que um novo método de armazenamento mantenha por mais tempo o ácido ascórbico do caqui do que o método usual. Foram então armazenados 20 caquis pelo novo método e 20 pelo método usual. Com base nas estatísticas apresentadas na Tabela 13.21, faça o teste t.

TABELA 13.21

Médias e variâncias do teor de ácido ascórbico em miligramas por 100 gramas da fruta, segundo o processo de armazenamento.

<i>Estatísticas</i>	<i>Armazenamento</i>	
	<i>Método usual</i>	<i>Novo método</i>
Número de caquis	20	20
Média	33,4	41,0
Variância	4,0	6,0

13.5.9 – Um nutricionista designa ao acaso 12 ciclistas para dois grupos: os dois grupos são instruídos a usar a dieta normal, mas o primeiro recebe um suplemento de vitaminas, enquanto o segundo recebe um placebo. Decorrido um mês, o nutricionista mede o tempo que cada ciclista demora em percorrer 10 km. Os dados estão na Tabela 13.22. Formule as hipóteses e faça o teste.

TABELA 13.22

Tempo, em minutos, para percorrer 10 km segundo o grupo.

<i>Grupo</i>	
<i>Suplemento de vitaminas</i>	<i>Placebo</i>
15	16
18	12
20	15
14	15
16	14
19	18

13.5.10 – Alguns estudos⁸ indicam que o açúcar torna as crianças mais ativas, outros não acham evidência de que isso aconteça. Foi feito um estudo com 25 crianças normais com idades entre 3 e 5 anos e 23 crianças que os pais diziam ficar hiperativas quando ingeriam açúcar. Os nutricionistas foram até as casas e retiraram todos os alimentos. Depois forneceram os alimentos por 4 semanas. As famílias receberam dois tipos de dieta, uma com açúcar, outra com alimentos adoçados com sacarina. Foram feitas medidas de comportamento nos dois grupos de crianças. Os dois grupos nunca foram comparados. As comparações foram feitas dentro de grupos. Esses dados constituem exemplo de dados pareados ou de grupos independentes? Que hipóteses estão em teste?

⁸ALIAGA, M. e GUNDERSON, B. *Interactive Statistics*. 2 ed. New Jersey: Prentice Hall. 2003. p. 679.

Respostas aos Exercícios Propostos

(página deixada intencionalmente em branco)

CAPÍTULO 1

- 1.9.1 –** Podem ser obtidas seis amostras diferentes: 1. Antônio e Luís; 2. Antônio e Pedro; 3. Antônio e Carlos; 4. Luís e Pedro; 5. Luís e Carlos; 6. Pedro e Carlos.
- 1.9.2 –** Podem ser selecionados: a) os elementos de ordem par; b) os elementos de ordem ímpar; c) os quatro primeiros elementos.
- 1.9.3 –** Numeram-se os alunos e sorteiam-se seis.
- 1.9.4 –** Divida 10 por cinco e obterá dois. Sorteie um dos dois primeiros números, isto é, 1 ou 2. Se sair 1, chame, para a amostra, o primeiro, o terceiro, o quinto, o sétimo e o nono nomes; se sair 2, chame o segundo, o quarto, o sexto, o oitavo e o décimo nomes.
- 1.9.5 –** O tipo de serviço odontológico que uma família demanda depende da sua renda. A amostragem com base na lista telefônica é incorreta porque seleciona apenas aqueles que têm telefone fixo, o que está associado com renda.
- 1.9.6 –** a) qualquer conjunto de 10 unidades como, por exemplo: 3; 5; 8; 13; 19; 22; 26; 27; 30; 40. b) no caso da amostra sugerida na resposta anterior: 0,3 ou 30%; c) 0,5 ou 50%; d) Boa (nota: não são boas as estimativas 0; 0,1; 0,9; 1).
- 1.9.7 –** Questão fechada: Você costuma escovar os dentes todos os dias?
Sim Não
Questão aberta: Como você limpa seus dentes?
- 1.9.8 –** A média da população (parâmetro) é 5. As médias das amostras (estatísticas) são: João e José: 8; João e Paulo: 7; João e Pedro: 5; José e Paulo: 5; José e Pedro: 3; Paulo e Pedro: 2. A média das médias das amostras é 5, igual à média da população.
- 1.9.9 –** O costume é escolher uma cidade “representativa” de todo o Estado.
- 1.9.10 –** a) alunos da universidade; b) percentual de alunos que têm trabalho remunerado; c) não, porque talvez no restaurante fiquem mais alunos que têm trabalho; d) não, porque excluiria os que têm condução própria.

1.9.11 – Leitores de livros técnicos.

1.9.12 – 143 policiais militares.

CAPÍTULO 2

2.8.1 – a) peso de pessoas: numérica contínua; b) marcas comerciais de um mesmo analgésico: nominal; c) temperatura de pessoas: numérica contínua; d) quantidade anual de chuva na cidade de São Paulo: numérica contínua; e) religião: nominal; f) número de dentes permanentes irrompidos em uma criança: numérica discreta; g) número de bebês nascidos por dia em uma maternidade: numérica discreta; h) comprimento de cães: numérica contínua.

2.8.2 –

Distribuição das pessoas segundo a opinião.

<i>Opinião</i>	<i>Freqüência</i>	<i>Percentual</i>
Favorável	425	49,9%
Contrária	368	43,2%
Não tem/não sabe	59	6,9%
Total	852	100,0%

2.8.3 –

Distribuição das notas de 200 alunos.

<i>Nota do aluno</i>	<i>Freqüência</i>	<i>Freqüência relativa</i>
De 9 a 10	16	0,08
De 8 a 8,9	36	0,18
De 6,5 a 7,9	90	0,45
De 5 a 6,4	30	0,15
Abaixo de 5	28	0,14
Total	200	1

2.8.4 –**Distribuição dos pacientes segundo o estágio da doença.**

<i>Estágio da doença</i>	<i>Freqüência</i>	<i>Freqüência relativa</i>
Leve	8	0,40
Moderado	9	0,45
Severo	3	0,15
Total	20	1,00

2.8.5 –

Não está definido se os valores iguais aos extremos de classe estão ou não incluídos na classe. Os intervalos se sobrepõem (por exemplo, de 20 a 30 e de 30 a 40; o valor 30 aparece nos dois intervalos) e falta uma classe: de 50 a 60.

2.8.6 –**Distribuição dos doadores de sangue segundo o tipo de sangue.**

<i>Tipo de sangue</i>	<i>Freqüência</i>	<i>Freqüência relativa</i>
0	15	0,375
A	16	0,4
B	6	0,15
AB	3	0,075
Total	40	1

2.8.7 – 20 alunos.**2.8.8 –****Distribuição das crianças segundo o hábito de sucção.**

<i>Hábito de sucção</i>	<i>Freqüência</i>	<i>Percentual</i>
Sucção do polegar	190	9,4%
Chupeta	588	29,2%
Mamadeira	618	30,7%
Não têm o hábito	615	30,6%
Total	2.011	100,0%

2.8.9 -

<i>Classe</i>
70 – 75
75 – 80
80 – 85
85 – 90
90 – 95
95 – 100
100 – 105
105 – 110
110 – 115
115 – 120

2.8.10 - O intervalo de classes é 5 (enfermeiros em serviço). O intervalo de toda a distribuição é 30.

2.8.11 -

Distribuição de pacientes acidentados no trabalho segundo o tempo de internação, em dias.

<i>Classe</i>	<i>Freqüência</i>
0 – 3	5
3 – 6	8
6 – 9	11
9 – 12	4
12 – 15	6
15 – 17	2
Total	36

Distribuição de pacientes acidentados no trabalho segundo o tempo de internação, em dias.

<i>Classe</i>	<i>Freqüência</i>
1 dia	2
De 2 a 3 dias	6
De 4 a 7 dias	12
De 8 a 14 dias	14
Mais de 14 dias	2
Total	36

- 2.8.12 –** Conjunto A: para achar o número de classes: $\sqrt{50} = 7,01 \approx 7$; amplitude dos dados: $70 - 24 = 46$. Dividindo a amplitude total pelo número de classes, acha-se o intervalo de classe: $46 \div 7 = 6,6 \approx 7$.

24 |– 31
31 |– 38
38 |– 45
45 |– 52
52 |– 59
59 |– 66
66 |– 73

Conjunto B: para achar o número de classes: $\sqrt{100} \approx 10$; amplitude dos dados: $821 - 187 = 634$. Dividindo a amplitude total pelo número de classes, acha-se o intervalo de classe: $634 \div 10 = 63,4 \approx 65$. Para facilitar os cálculos, faça o extremo inferior da primeira classe igual a 185.

185 |– 250
250 |– 315
315 |– 380
380 |– 445
445 |– 510
510 |– 575
575 |– 640
640 |– 705
705 |– 770
770 |– 835

2.8.13 –**Taxa de abandono do tratamento contra tuberculose pulmonar segundo a zona de moradia.**

Zona	Abandono do tratamento			Taxa de abandono
	Sim	Não	Total	
Urbana	15	80	95	15,8%
Rural	70	35	105	66,7%
Total	85	115	200	42,5%

2.8.14 –**Distribuição dos dentistas segundo a adoção de métodos de prevenção de cáries e doenças gengivais no consultório.**

Prevenção	Freqüência	Percentual
Sim	78	78,0%
Não	22	22,0%
Total	100	100,0%

A prática da prevenção deveria ser adotada por 100% dos dentistas.

2.8.15 –**Número de óbitos por grupos de causas. Brasil, 2004.**

Grupos de causas	Masculino		Feminino	
	Nº	%	Nº	%
Doenças infecciosas e parasitárias	27.437	5,2%	18.615	5,0%
Neoplasias	76.065	14,5%	64.724	17,3%
Doenças do aparelho circulatório	150.383	28,8%	135.119	36,2%
Doenças do aparelho respiratório	55.785	10,7%	46.369	12,4%
Afecções originadas no período perinatal	17.530	3,4%	13.165	3,5%
Causas externas	107.032	20,5%	20.368	5,4%
Demais causas definidas	88.563	16,9%	75.399	20,2%
Total	522.795	100,0%	373.759	100,0%

Foram 896.554 óbitos com causa definida, 58,3% homens e 41,7% mulheres. Doenças do aparelho circulatório respondem pela maior proporção de mortes. Chama atenção a grande proporção de óbitos de homens por causas externas (acidentes e homicídios).

2.8.16 –**Pacientes portadores de carcinoma epidermóide de base de língua, segundo a faixa etária, em anos.**

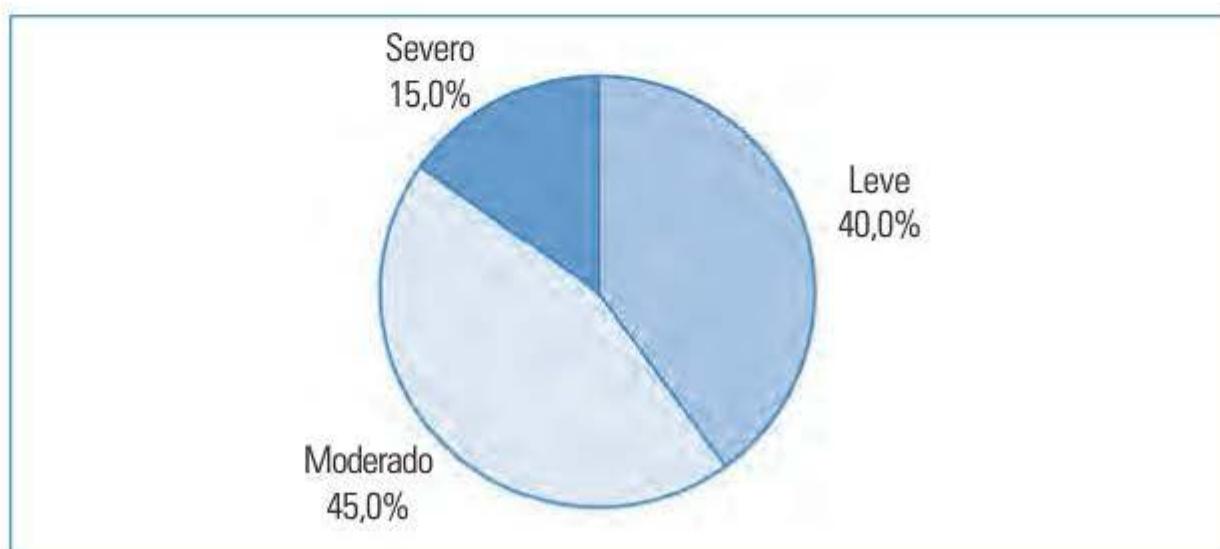
<i>Faixa etária</i>	<i>Número</i>	<i>Freqüência relativa</i>
30 – 40	10	3,4%
40 – 50	66	22,8%
50 – 60	119	41,0%
60 – 70	66	22,8%
70 – 80	24	8,3%
80 e mais	5	1,7%
Total	290	100,0%

A faixa etária de maior risco: dos 50 aos 60 anos.

2.8.17 –**Número de órgãos obtidos de doadores cadáveres.**

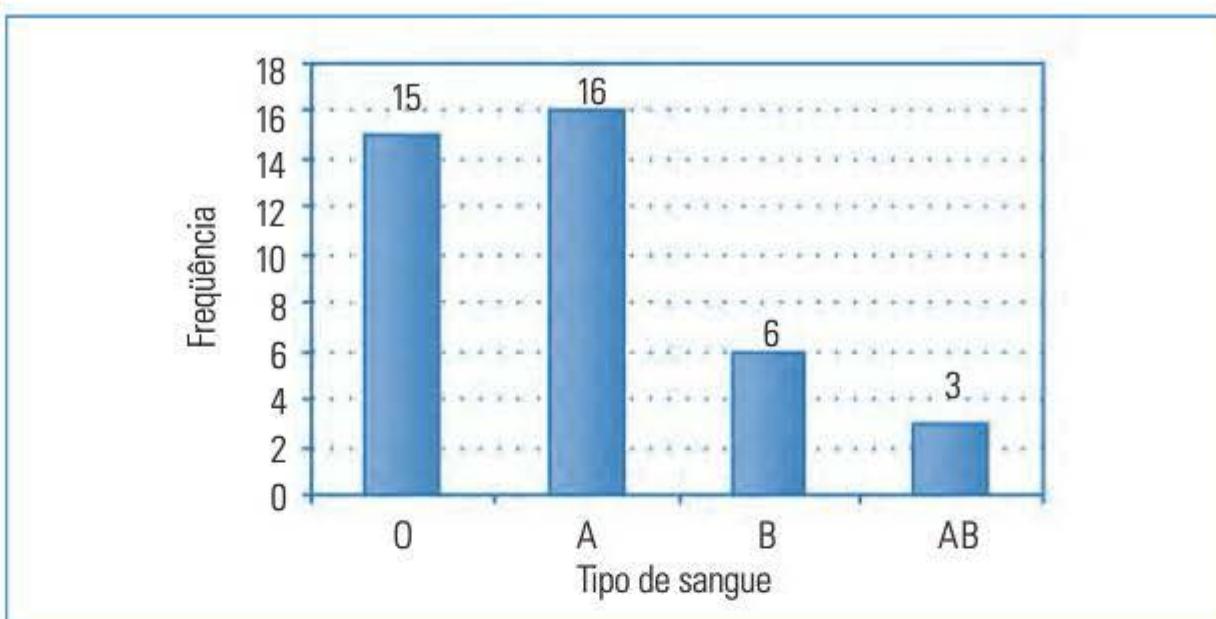
<i>Órgão</i>	<i>Número de doadores</i>	<i>Número de órgãos aproveitados</i>	<i>Taxa de aproveitamento</i>
Rim	105	210	100,0%
Coração	105	45	42,9%
Fígado	105	20	19,0%
Pulmões	105	17	8,1%

Nota: Cada cadáver é potencialmente doador de dois rins, um coração, um fígado e dois pulmões. A taxa de aproveitamento é sobre número de órgãos — não de cadáveres.

CAPÍTULO 3**3.5.1 –**

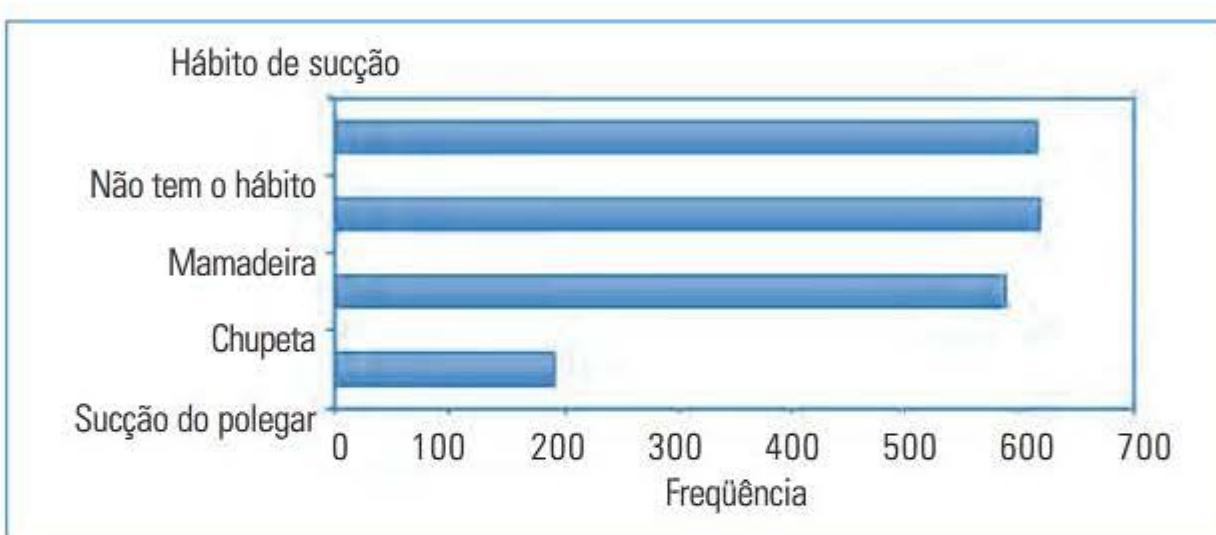
Distribuição dos pacientes segundo o estágio da doença.

3.5.2 -



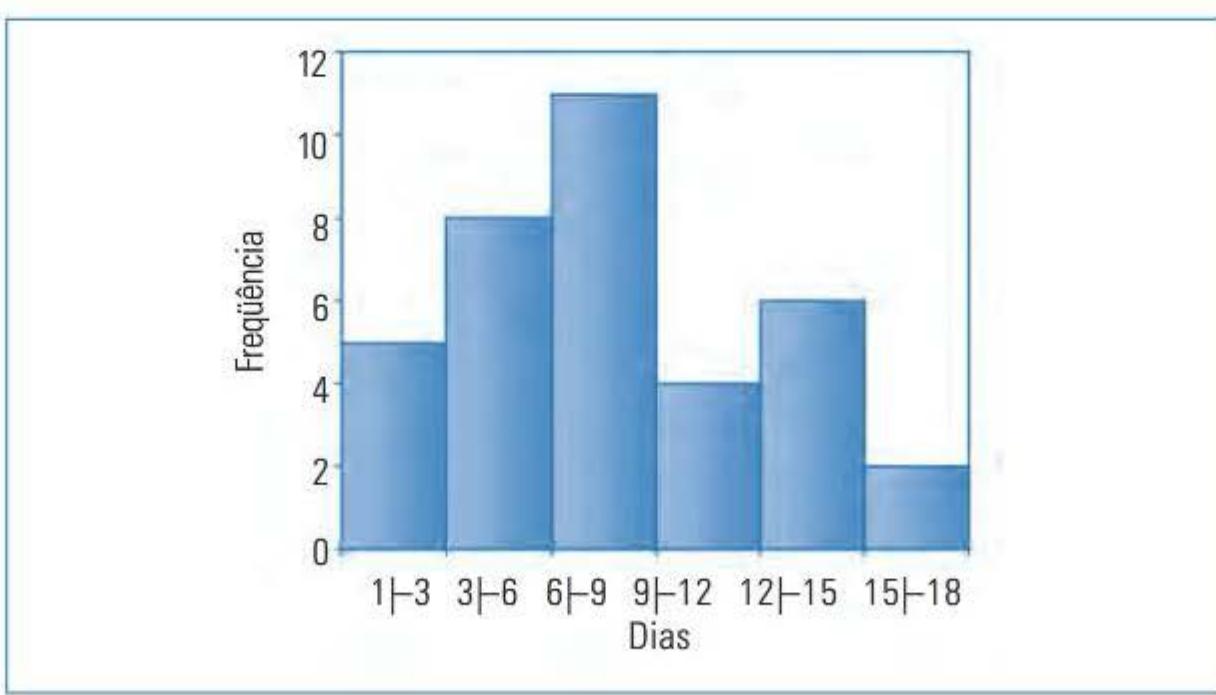
Distribuição dos doadores de sangue segundo o tipo de sangue.

3.5.3 -

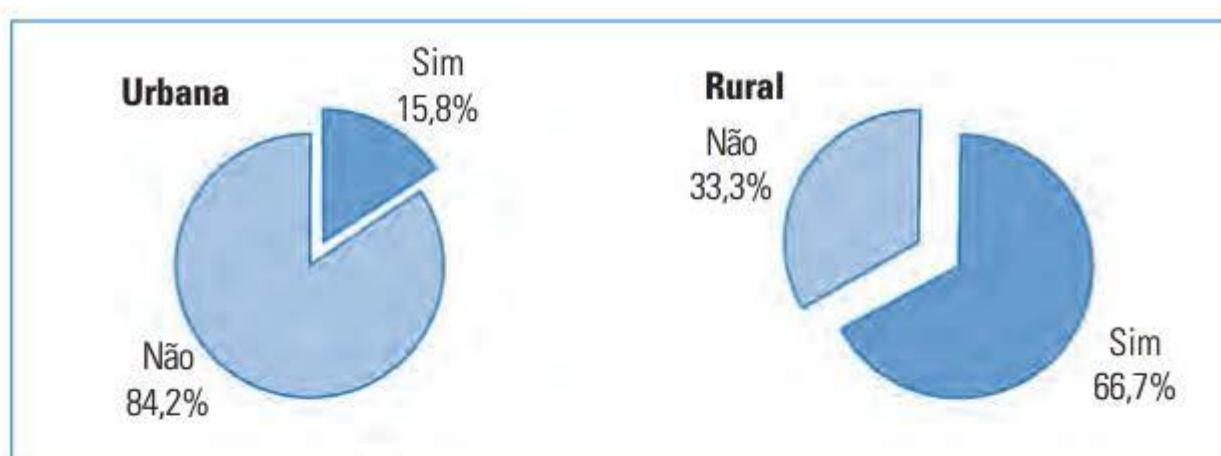


Distribuição das crianças segundo o hábito de sucção.

3.5.4 -



Distribuição de pacientes acidentados no trabalho segundo o tempo de internação, em dias.

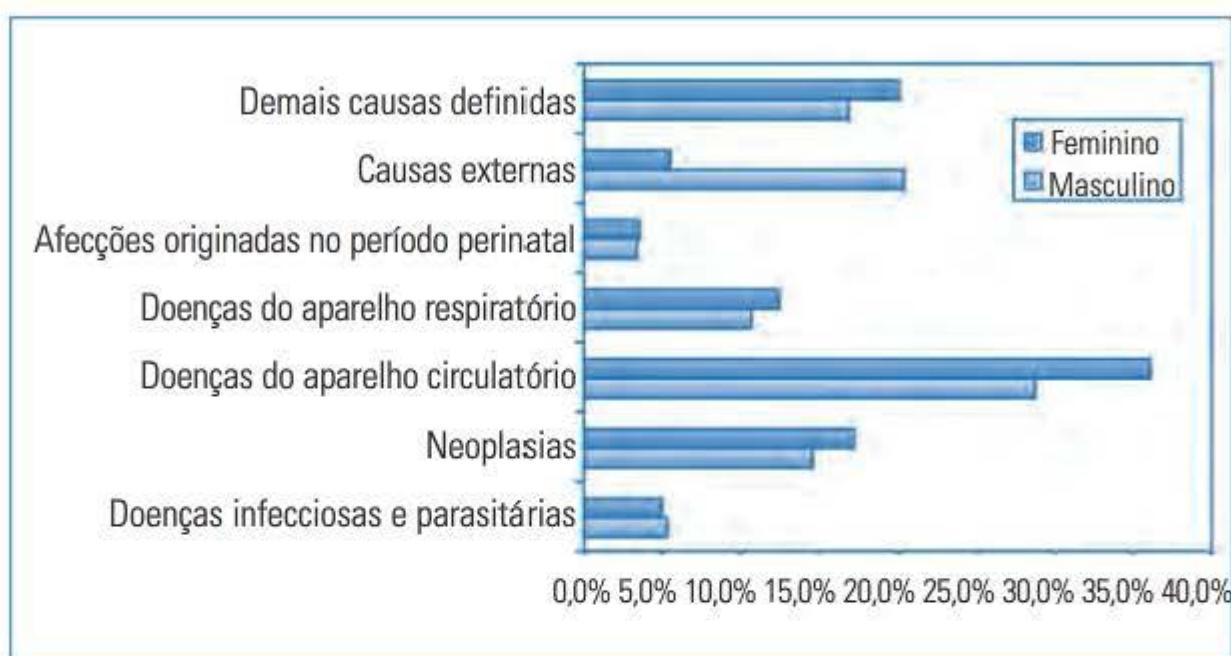
3.5.5 –

Taxa de abandono do tratamento contra tuberculose pulmonar segundo a zona de moradia.

3.5.6 –

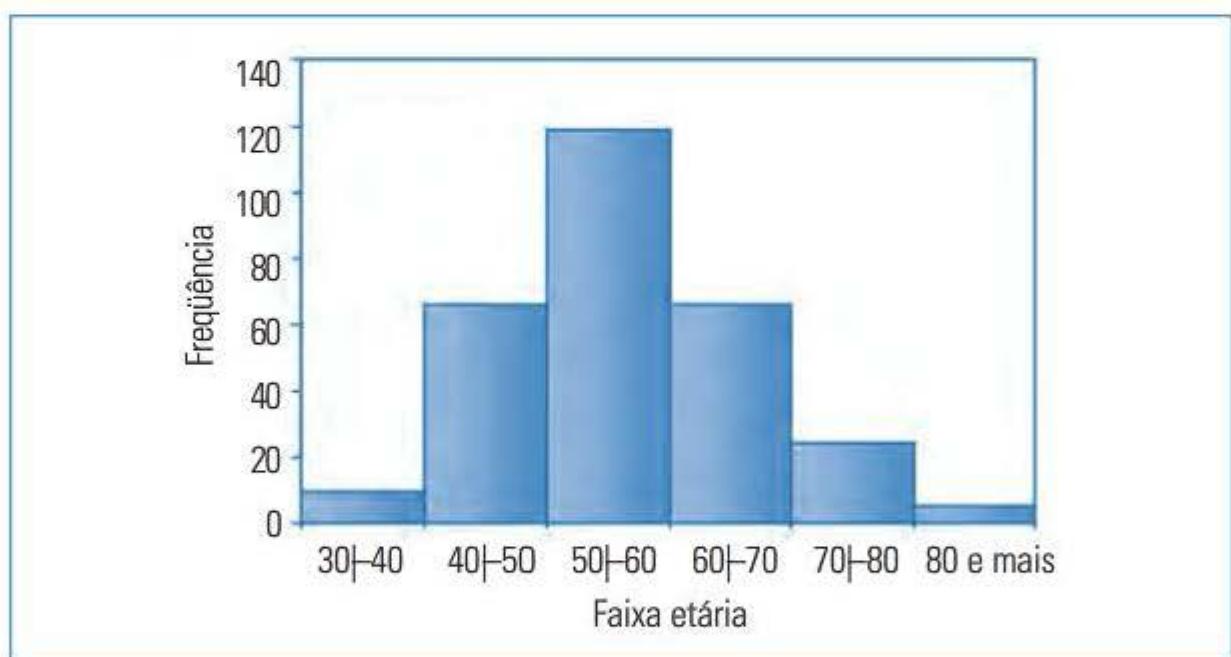
Proporção de óbitos por grupos de causas. Brasil, 2004.

Nesses gráficos, as grandes causas foram colocadas em ordem decrescente, considerando as porcentagens. Mas os dois gráficos podem ser reunidos em um só, como na figura que se segue.

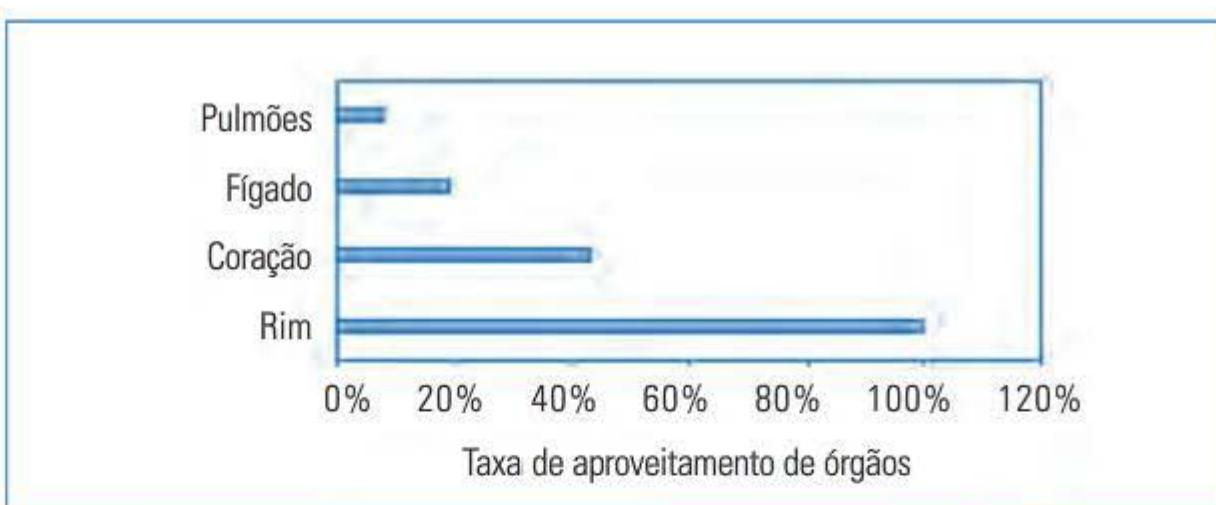


Proporção de óbitos por grupos de causas. Brasil, 2004.

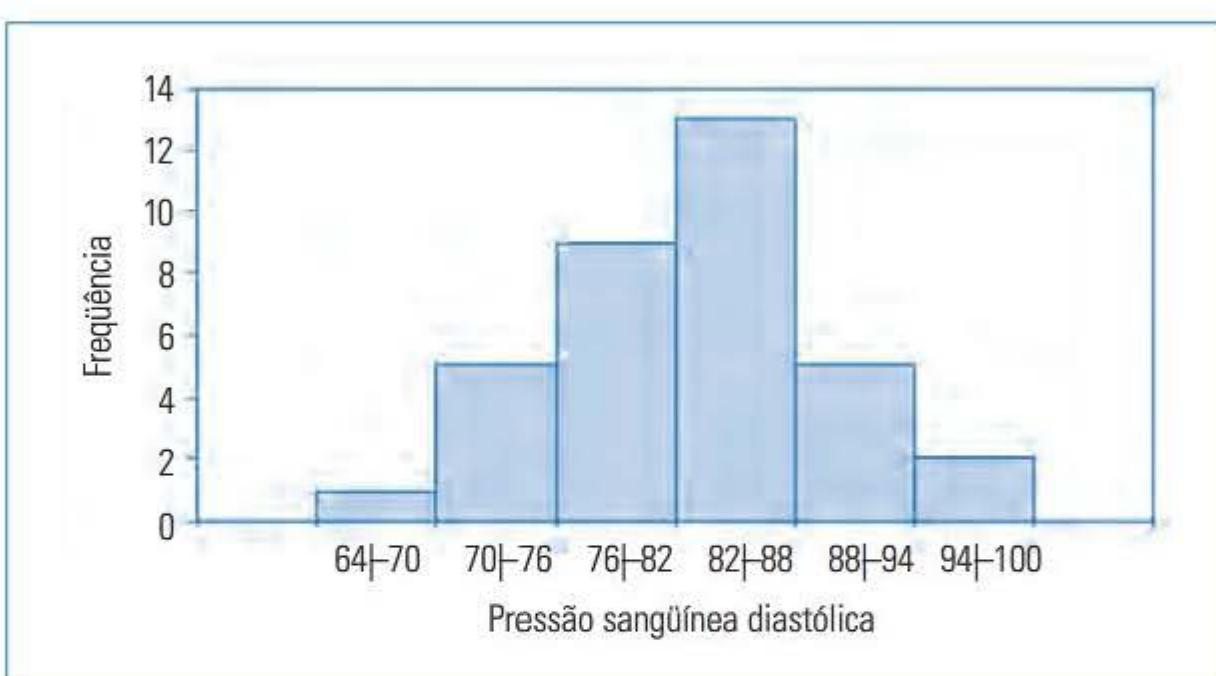
3.5.7 -



Pacientes portadores de carcinoma epidermóide de base de língua, segundo a faixa etária, em anos.

3.5.8 –

Taxa de aproveitamento de órgãos obtidos de doadores cadáveres.

3.5.9 –

Pressão sanguínea diastólica de 35 enfermeiros que trabalham em um hospital.

3.5.10 –

Pressão sanguínea diastólica de 35 enfermeiros que trabalham em um hospital

CAPÍTULO 4

- 4.6.1 –** a) Média = 5; mediana = 6; moda = 8
b) Média = 8; mediana = 8; moda = 8.
c) Média = 11; mediana = 10; moda = 10.
d) Média = 1; mediana = 0; não tem moda.
e) Média = 2 mediana = 1; duas modas: 1 e 2.
- 4.6.2 –** Mediana.
- 4.6.3 –** Moda.
- 4.6.4 –** 24 anos.
- 4.6.5 –** A média é 100 miligramas por 100 ml de sangue, e a mediana é 99,5 miligramas por 100 ml de sangue.
- 4.6.6 –** Estatura: média = 1,70 m; mediana = 1,68 m.
Peso: média = 72,5 kg; mediana = 70 kg.
Pressão arterial: média = 165,5mmHg; mediana 160mmHg.
- 4.6.7 –** Masculino: média = 0,88 dente cariado; feminino: média = 1 dente cariado.
- 4.6.8 –** 1,06 minuto. O rato que não dormiu não entra na média, porque tempo de latência é o tempo para a droga fazer efeito — no caso, dormir.
- 4.6.9 –** Masculino: média = 7,00 gramas por dia; mediana = 6,5 gramas por dia.
Feminino: média = 7,00 gramas por dia; mediana = 7,0 gramas por dia.
- 4.6.10 –** Masculino: média = 0,90 litro por dia; mediana = 0,85 litro por dia.
Feminino: média = 0,80 litro por dia; mediana = 0,75 litro por dia.
- 4.6.11 –** Metade das pacientes retornou às atividades menos de 27,5 dias depois de submetidas à histerectomia; não houve moda, ou seja, nenhum número de dias foi mais freqüente.
- 4.6.12 –** 3,62 miligramas de ácido ascórbico em 100 ml.
- 4.6.13 –** Sim: 1; 2; 3; 3; 4; 5; a média, a mediana e a moda são iguais a 3.
- 4.6.14 –** A média, porque a última classe não tem o extremo superior definido.

CAPÍTULO 5

5.6.1 – a) 1; b) 5; c) 4.

5.6.2 – a) $\sum x = 35$; b) $\sum(x - \bar{x})^2 = 20$

5.6.3 – A média é 4, e o desvio padrão é 3.

5.6.4 – O tamanho da amostra é 6.

5.6.5 – Média, 24 e variância 80.

5.6.6 – Antônio: média = 5; desvio padrão = 0.

João: média = 5; desvio padrão = 1.

Pedro: média = 5; desvio padrão = 5.

As notas de Antônio não variaram; as notas de Pedro variaram muito mais do que as de João.

5.6.7 – a) O desvio padrão pode ser maior do que o valor da média; exemplo: -2; 0; 2 b) O valor do desvio padrão pode ser igual ao valor da média; exemplo: 10; 10; 5; 0; 0; c) O valor do desvio padrão *não* pode ser negativo, por definição. d) O desvio padrão é igual a zero quando todos os dados do conjunto são iguais entre si.

5.6.8 – A variância é 16, o desvio padrão é 4 e o coeficiente de variação é 4%.

5.6.9 – A média é 5 e a variância é 0,8.

5.6.10 – a) Desvantagem de usar a amplitude: os dois conjuntos podem ter amplitudes iguais e variabilidades diferentes. b) Não. c) Sim, quando menor do que 1.

5.6.11 – 1º ano: média = 118,0; desvio padrão = 4,12.

2º ano: média = 131,0; desvio padrão = 8,66.

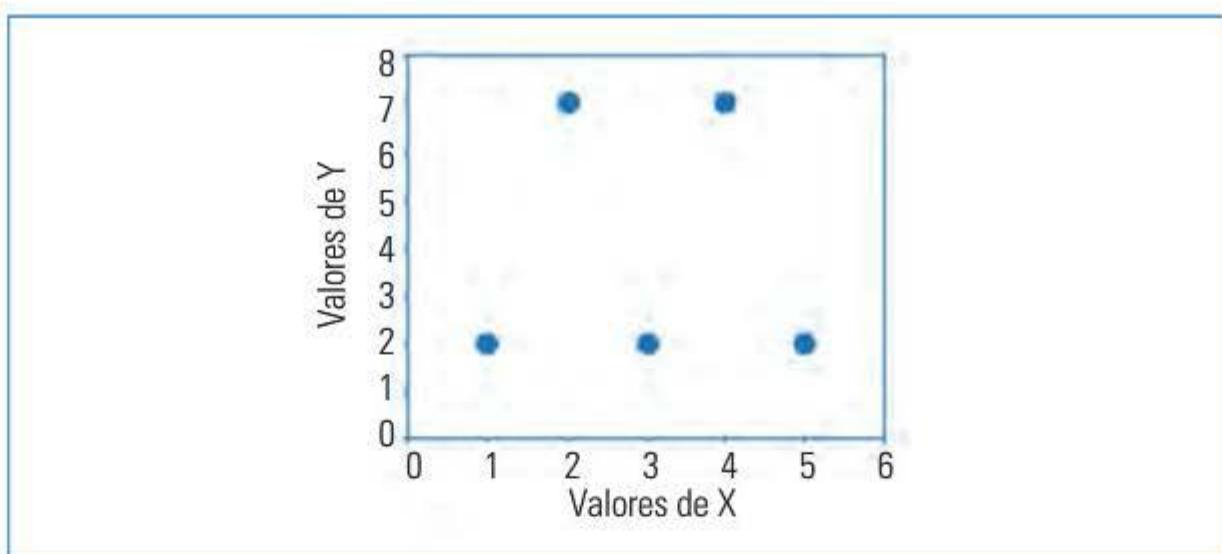
A média do 2º ano é 11% maior do que a do 1º ano e a variabilidade é praticamente o dobro.

5.6.12 – A diferença de médias não é muito grande, mas a diferença de variabilidades é tão grande que justificaria preferir a primeira dieta para perda de peso. Como as respostas são mais homogêneas, a expectativa do resultado é mais previsível.

- 5.6.13 –** Diurno: média = 47,5; desvio padrão = 9,3.
Noturno: média = 45,4; desvio padrão = 9,4.
A média é um pouco maior no diurno, mas as variabilidades são praticamente as mesmas.

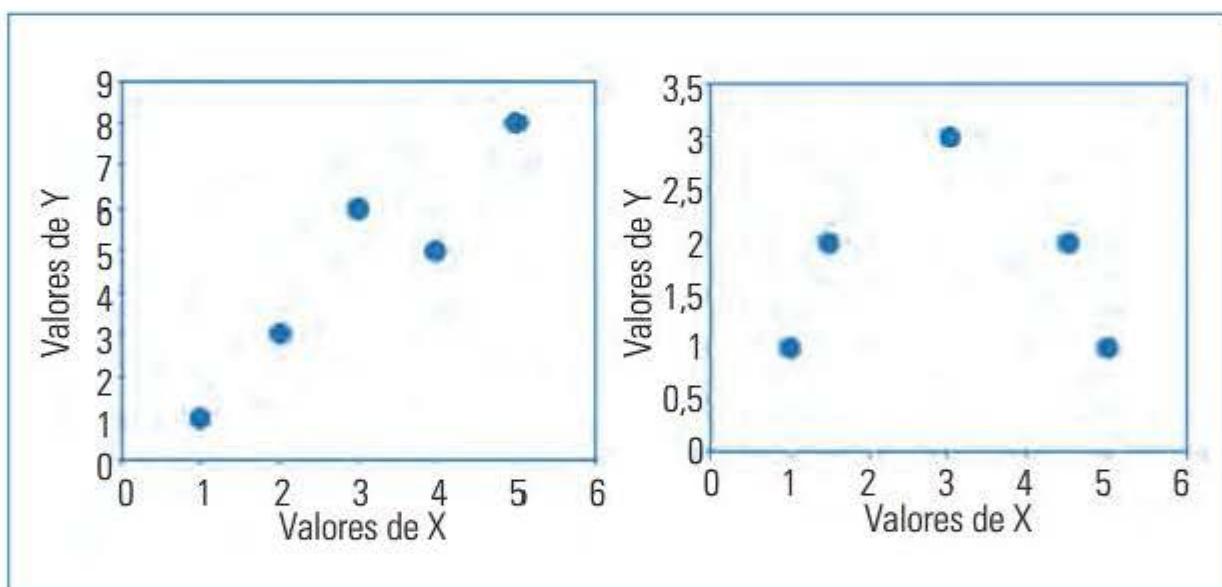
CAPÍTULO 6

- 6.6.1 –** a) $r = 1$: correlação perfeita positiva.
b) $r = -1$: correlação perfeita negativa.
c) $r = 0$: correlação nula.
d) $r = 0,90$: correlação positiva alta.
e) $r = -0,90$: correlação negativa alta
- 6.6.2 –** a) correlação negativa
b) correlação positiva
c) correlação nula.
- 6.6.3 –** O sobrepeso pode ser um fator de risco para a morte por doenças do coração
- 6.6.4 –** Não.
- 6.6.5 –** a) Correlação perfeita negativa
b) Forte correlação positiva
c) Correlação nula ou próxima de zero
- 6.6.6 –** 1; 1 ou -1; positiva ou negativa; zero; maior.
- 6.6.7 –** Negativa
- 6.6.8 –** Se as variáveis estão ou não correlacionadas
- 6.6.9 –** Não existe correlação entre as variáveis: $r = 0$. O diagrama de dispersão mostra isso.



Dados relativos as duas variáveis X e Y .

- 6.6.10-** Para o Conjunto A, $r = 0,936$, portanto alta correlação positiva. Para o Conjunto B, $r = 0$, o que, no caso, não significa correlação nula, mas, como mostra o gráfico, correlação não-linear.



Dois conjuntos de pares de valores de duas variáveis.

- 6.6.11 -** Não é possível¹ calcular o valor de r , mas, obviamente, não existe correlação entre as variáveis: X cresce e Y permanece constante.
- 6.6.12 -** $\sum x = 255$, $\sum x^2 = 9.443$, $\sum y = 17,25$, $\sum y^2 = 50,4375$, $\sum xy = 660,25$. Logo, $r = 0,913$.

¹Divisão por zero, uma vez que a variância de Y , que aparece no denominador, é zero.

6.6.13 – Para o Conjunto A, $r = 1$, portanto correlação perfeita positiva. Para o Conjunto B, $r = 0$; o valor altamente discrepante anula a correlação. Mas, atenção: retire o valor discrepante apenas no caso de ter havido erro na leitura ou no registro do dado. Outras situações demandam discussão. Note ainda: o valor discrepante mudou totalmente o valor de r pelo fato de a amostra ser pequena.

6.6.14 – O valor de r é 0,774 (correlação positiva alta).

6.6.15 –



Duração do exercício, em minutos e $\text{VO}_{\text{2máx}}$ em mililitros por quilograma por minuto para 12 homens saudáveis.

Olhando o diagrama, é razoável afirmar que $\text{VO}_{\text{2máx}}$ diminui quando aumenta a atividade.

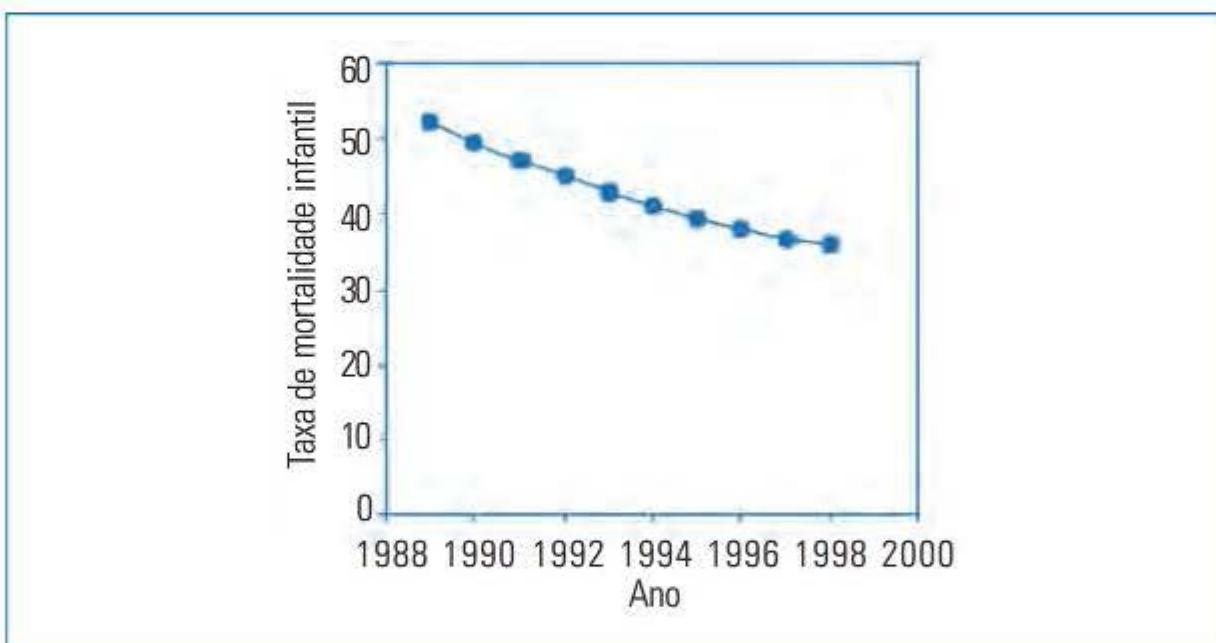
CAPÍTULO 7

7.8.1 – A razão de sexos, que se inicia acima de 100 (o que significa que nascem mais homens do que mulheres) começa a diminuir dos 15 aos 30 anos, tende a estabilizar a queda até os 55 anos, depois cai cada vez mais rapidamente.



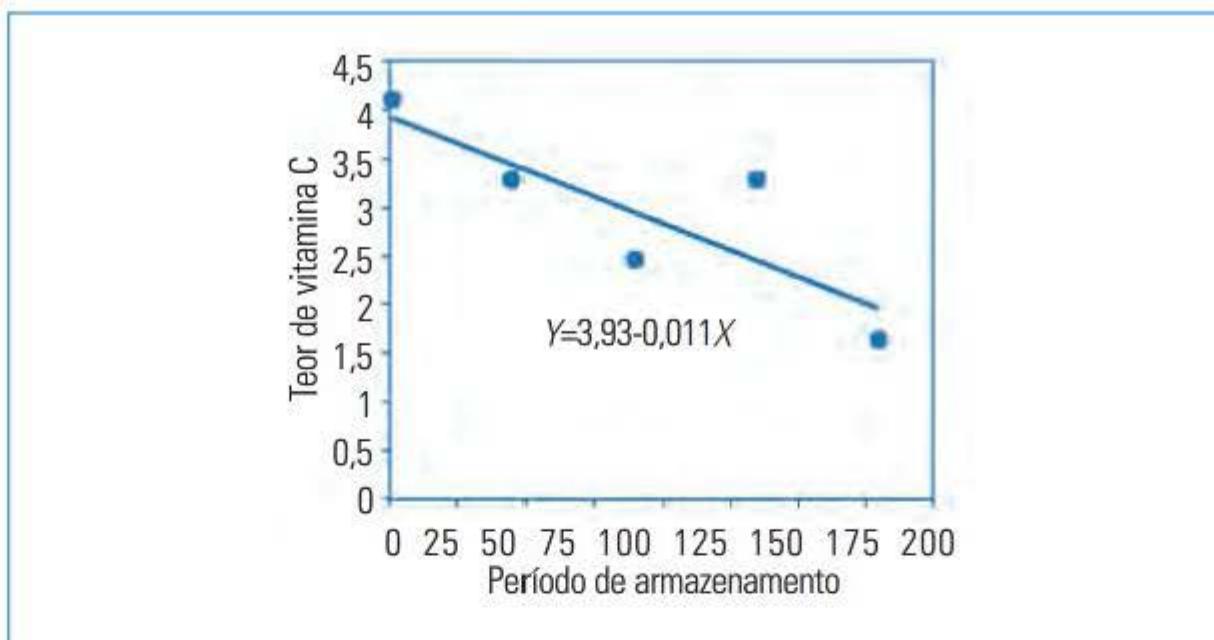
Razão de sexos no Brasil, em 2005.

- 7.8.2 –** A taxa de mortalidade infantil diminuiu no período, mas ainda não é baixa.



Taxa de mortalidade infantil no Brasil, de 1889 a 1998.

- 7.8.3 –** Tanto o gráfico como a reta ajustada, indicam que o teor de vitamina C no suco de maçã diminui à medida que aumenta o tempo de armazenamento.



Teor de vitamina C (mg de ácido ascórbico/100 ml de suco de maçã)
em função do período de armazenamento em dias.

- 7.8.4 –** O coeficiente de correlação não muda, mas a reta de regressão será outra. As duas retas se cruzarão no ponto de coordenadas iguais às médias de X e Y .
- 7.8.5 –** Não.
- 7.8.6 –** $\hat{Y} = 5 + X$
- 7.8.7 –** Não seria possível achar o valor de b pela fórmula, uma vez que o denominador seria zero. Mas a idéia é de uma reta paralela ao eixo das ordenadas.
- 7.8.8 –** Os dados são poucos para discutir assunto tão complexo, mas, em geral, pode-se afirmar que escolaridade está associada ao nível de renda, que significa maiores gastos com produtos de higiene e maior busca de profissionais de saúde, além da facilidade de ter e buscar novos conhecimentos. De qualquer forma, ensinar métodos preventivos dá bons resultados. O que não se pode é usar estatísticas de má qualidade — traçou-se a reta pelos pontos médios de X e pelas médias de Y , o que determinou maior R^2 — mesmo que seja para “provar” assuntos comprovados, ou para demonstrar boas intenções.
- 7.8.9 –** Os gastos com propaganda aumentaram as vendas. O valor de $R^2 = 0,984$ indica que a proporção da variação do volume de vendas Y explicada pela variação do gasto em propaganda é muito alta.
Mas cuidado: não se pode extrapolar.

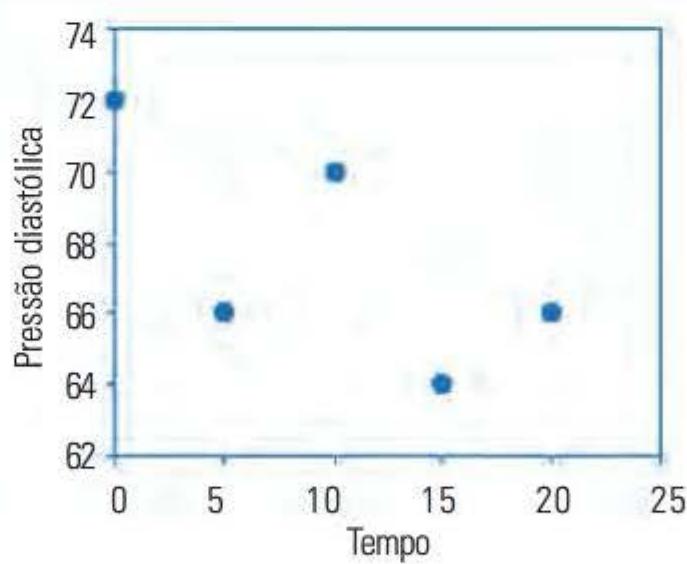


Gastos com propaganda, em reais, na semana, e valores recebidos, em reais, nas vendas.

7.8.10 – $\hat{Y} = 11,24 + 1,309 X$

7.8.11 – O $V_0_{\text{máx}}$ inalado diminui linearmente quando aumenta a atividade, no intervalo estudado.

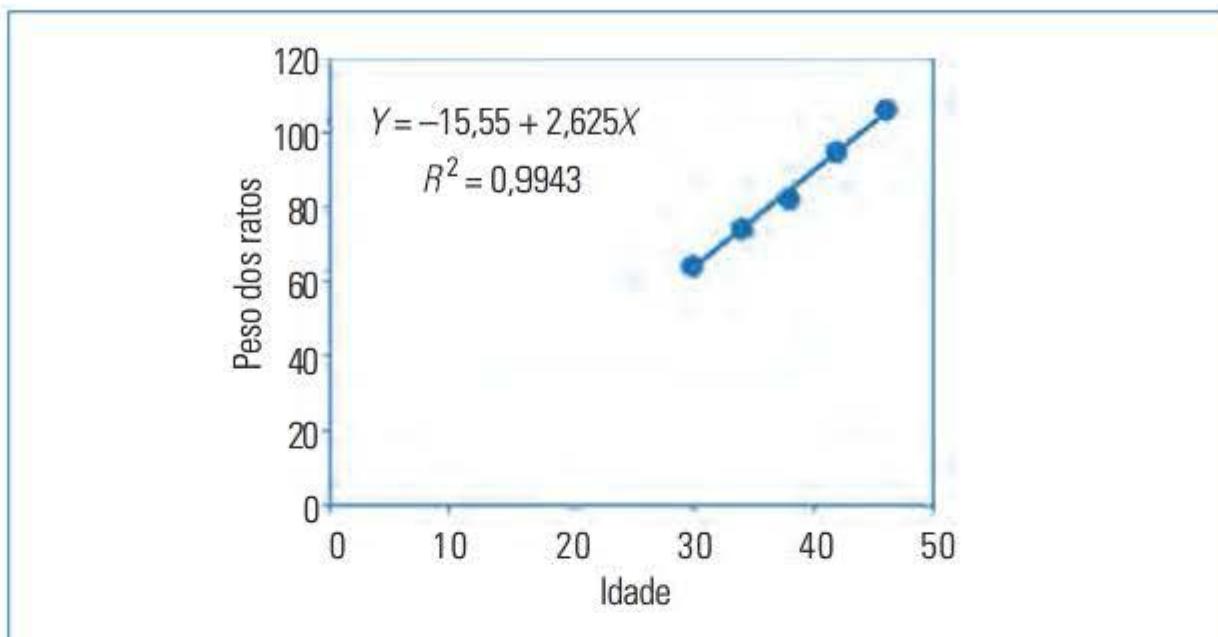
$$\hat{Y} = 162,57 - 8,841X$$



Tempo em minutos desde o início do repouso e pressão sangüínea diastólica, em milímetros de mercúrio.

- 7.8.12 –** Para se ajustar uma reta de regressão aos dados, é preciso que as observações sejam *independentes*. Observações feitas ao longo do tempo não são independentes.

7.8.13 –



Idade, em dias, e peso médio, em gramas, de 10 ratos machos da raça Wistar.

Peso aos 32 dias = 68,45 gramas.

- 7.8.14 –** A regressão *exponencial* traz a variável explanatória no expoente. Escreve-se:

$$\hat{Y} = ae^{bx}$$

Para ajustá-la, é preciso calcular o logaritmo neperiano de X. Ajusta-se:

$$\hat{Y} = A + b\ln X.$$

Cálculos auxiliares:

X	Y	In Y	XIn Y	X ²
28	1,25	0,22314	6,24802	784
32	1,25	0,22314	7,14059	1.024
35	1,75	0,55962	19,58655	1.225
38	2,25	0,81093	30,81535	1.444
39	3,25	1,17865	45,96754	1.521
41	3,25	1,17865	48,32485	1.681
42	4,25	1,44692	60,77060	1.764
$\Sigma X = 255$		$\Sigma Y = 17,25$	$\Sigma \ln Y = 5,62106$	$\Sigma X \ln Y = 218,85351$
				$\Sigma X^2 = 9.443$

Aplicando as fórmulas, obtém-se:

$$\hat{Y} = -2,535 + 0,09164 \ln X$$

$$\hat{Y} = 0,0792e^{0,0916x}$$

CAPÍTULO 8

8.8.1 - a) $\frac{4}{52} = \frac{1}{13}$

b) $\frac{13}{52} = \frac{1}{4}$

c) $\frac{1}{52}$

8.8.2 - a) $\frac{8}{10}$

b) $\frac{7}{10}$

c) $\frac{2}{10}$

8.8.3 - a) $\frac{7}{15}$

b) $\frac{8}{15}$

c) zero

8.8.4 - É mais fácil resolver o problema construindo o espaço amostral.

1	2	3	4	5	6	7	8	9	10
ABC	ABD	ABE	ACD	ACE	ADE	BCD	BCE	BDE	CDE

a) $\frac{6}{10}$

b) $\frac{3}{10}$

8.8.5 - a) $\frac{1}{6}$

b) $\frac{1}{6}$

- 8.8.6 –** Os eventos “ser reprovado em Matemática” e “ser reprovado em Português” *não* são independentes porque a condição de independência, dada em seguida, não é satisfeita.

$$P(A \cap B) = P(A) + P(B)$$

Temos:

$$P(\text{Reprovado em Português}) = 0,10$$

$$P(\text{Reprovado em Matemática}) = 0,20$$

$$P(\text{Reprovado em Português} \cap \text{Reprovado em Matemática}) = 0,05$$

$$0,05 \neq 0,10 \times 0,20$$

- 8.8.7 –** a) 50% b) 50%

- 8.8.8 –** 0,1%

- 8.8.9 –** 50%

- 8.8.10 –** a) 36% b) 1%

CAPÍTULO 9

- 9.6.1 –**

Eventos e respectivos resultados no jogo.

<i>Eventos</i>	<i>Resultados possíveis</i>
12	Ganha
13	Perde
21	Perde
23	Perde
31	Perde
32	Ganha

O jogador perde mais vezes do que ganha, porque só 2 é par e 1 e 3 são ímpares. O jogo é injusto.

9.6.2 –**Distribuição do número de meninos em uma família de cinco crianças.**

X	P(X)
0	1/32
1	5/32
2	10/32
3	10/32
4	5/32
5	1/32

9.6.3 – $\mu = 5$, $\sigma^2 = 2,5$ **9.6.4 –** $\mu = 2$, $\sigma^2 = 1,6$ **9.6.5 –** 2,7%**9.6.6 –** 27/64 ou 42,2%**9.6.7 –** 0,001%

9.6.8 – a) As respostas têm distribuição binomial. b) Depende da taxa de respostas, que deve ser igual ou superior a 70%, isto é, pelo menos 70% dos questionários devem ter sido respondidos. Um cuidado importante, aqui, é saber se a pergunta feita não induz um tipo de resposta (por exemplo, dizer “não” pode ser prejudicial para a enfermeira, ou pode ofender colegas). Nesse caso, as respostas poderiam, eventualmente, ser tendenciosas, e a taxa de respostas pequena.

9.6.9 – 35,4%.

9.6.10 – Se considerarmos cada dia como um ensaio, em cada dia podem ocorrer mais de dois eventos (ocorreu acidente ou não). Interessa o número de acidentes por dia e depois o estudo da distribuição de freqüências: em quantos dias houve um acidente, 2, 3 etc. e o estudo das causas. Portanto, a variável não é binomial.

CAPÍTULO 10

10.6.1 – 49,01%

10.6.2 – a) $\pm 0,67$; b) $\pm 1,64$; c) $\pm 1,96$.

10.6.3 – a) 78,88% b) 10,56%

10.6.4 – a) 4,75% b) 45,25%

10.6.5 – a) 97,72% b) 2,28%

10.6.6 – a) 21,19% b) 21,19%

10.6.7 – Usando apenas os conhecimentos adquiridos com a distribuição normal, é razoável dizer que a média, mais um desvio padrão, é ponto de alerta (no caso, 142,5 5 mEq/L de plasma); média mais dois desvios padrões (no caso 145,55 mEq/L de plasma) seria ponto de corte para dizer que está alta a concentração de sódio no plasma de uma pessoa.

10.6.8 – a) 0,1587 ou 15,87%; b) 0,0228 ou 2,28% c) 0,5 ou 50% d) 0,1003 ou, aproximadamente, 10%.

10.6.9 – Sim, metade dos escores é positiva e metade é negativa porque a distribuição normal reduzida é simétrica em torno da média.

10.6.10 – 0,0475 ou 4,75%.

CAPÍTULO 11

11.6.1 – A proporção de adultos que pensam que sofrem da síndrome é $590 \div 3.066 = 0,1924$. O intervalo de 95% de confiança vai de 0,178 a 0,206.

11.6.2 – A resposta mais razoável talvez seja $3 \div 3.066 = 0,000978$ ou 0,0978%.

11.6.3 – O intervalo de 90% de confiança vai de 121,7 a 124,3mmHg.

11.6.4 – O intervalo de 99% de confiança vai de 15,50 a 16,90 g de hemoglobina por 100 ml de sangue.

11.6.5 – O intervalo de 90% de confiança vai de 49,20 a 50,80 cm.

11.6.6 – O intervalo de 95% de confiança vai de 92,5 a 97,5 mg de glicose por 100 ml de sangue.

11.6.7 – O intervalo teria de ser $0 \leq p \leq 1$. Mas esse intervalo não tem qualquer utilidade.

11.6.8 – O intervalo de 95% de confiança vai de 29,46 a 30,94 g.

11.6.9 – O intervalo de 98% de confiança vai de 647,05 a 668,95 mg.

- 11.6.10 –**
- a) Não necessariamente.
 - b) Sim.
 - c) Não necessariamente.
 - d) Não.

CAPÍTULO 12

12.4.1 – Um teste de qui-quadrado ao nível de 5% de significância não rejeita a hipótese de que é de 3% a proporção de recém-nascidos com defeito ou doença séria.

12.4.2 – $\alpha^2 = 4,82$. A proporção de recém-nascidos portadores de anomalia congênita é maior no sexo feminino.

12.4.3 – $\alpha^2 = 9,04$. A ausência congênita de dentes ocorre mais em meninas.

12.4.4 – O coeficiente de Yule é -0,372. A anodontia está associada ao sexo, na ordem de 37%.

12.4.5 – $\alpha^2 = 1,32$. A associação é -0,22, relativamente pequena. O teste não rejeita a hipótese de que presença de aberração cromossômica no feto não depende da faixa de idade da gestante ser de 35 até 40 anos, ou de 40 anos ou mais.

12.4.6 – Hipótese da nulidade: existe associação entre implantes mamários e doenças do tecido conjuntivo e outras doenças. Hipótese alternativa: doenças do tecido conjuntivo e outras não estão associadas aos implantes mamários. A proporção é 0,00668 nos dois grupos.

12.4.7 – Hipótese da nulidade: a probabilidade de natimorto é a mesma para os dois sexos. Hipótese alternativa: a probabilidade de natimorto é maior para um dos sexos. $\alpha = 5\%$. Calculado: $\chi^2 = 1,15$, menor que o da Tabela de χ^2 , com 1 grau de liberdade. Não se rejeita H_0 .

12.4.8 – O coeficiente de Yule é 0,0816. Associação positiva, mas muito pequena (da ordem de 8%).

12.4.9 – Hipótese da nulidade: a probabilidade de dormir mais de 8 horas é a mesma para as duas faixas de idade. Hipótese alternativa: a probabilidade de dormir mais de 8 horas é diferente para as duas faixas de idade. $\alpha = 1\%$. $\chi^2 = 22,26$, portanto rejeite H_0 ao nível de 1% de significância.

12.4.10 – $\chi^2 = 48,24$; rejeita-se H_0 ao nível de 1%.

CAPÍTULO 13

13.5.1 – A tabela dada em seguida apresenta as médias e os desvios padrões de pesos de ratos.

Médias e desvios padrões de pesos de ratos.

<i>Estatísticas</i>	<i>Ração</i>	
	<i>Padrão</i>	<i>Experimental</i>
Média	188,0	212,0
Desvio padrão	3,7	3,7

O valor de t é 4,536, significante a 5%. Os ratos submetidos à ração experimental ganharam mais peso.

13.5.2 – Observações pareadas; $t = 4,226$, significante ao nível de 5%. O teste B dá, em média, resultados显著mente maiores de QI do que o teste A.

13.5.3 – $t = 1,642$, não-significante a 5%. Os dados não mostram que o uso de anticoncepcionais orais aumenta a pressão sanguínea sistólica.

13.5.4 – $t = 0,623$, não-significante a 5%. Os dados não mostram diferença de peso ao nascer entre sexos.

13.5.5 – A tabela dada em seguida apresenta as médias e as variâncias da pressão sanguínea dos ratos.

Médias e variâncias da pressão sanguínea dos ratos segundo a temperatura a que foram submetidos.

<i>Estatísticas</i>	<i>Temperatura</i>	
	<i>5° C</i>	<i>26° C</i>
Média	165,8	378,5
Variância	218,17	573,90

Não se rejeita a hipótese de variâncias iguais ($p > 0,05$). Rejeita-se a hipótese de médias iguais ($p < 0,05$).

- 13.5.6 –** Rejeita-se a hipótese de médias iguais ($p = 0,0097$).
- 13.5.7 –** Estatísticas para comparar do tempo de alívio da dor obtido com a nova droga, em relação à antiga.

<i>Estatística</i>	<i>Resultado</i>
Valor de F	1,33
p -valor	0,2652
Variância ponderada	2,003
Valor de t	-1,18
p -valor (unilateral)	0,1227

Não se rejeita a hipótese de variâncias iguais ($p > 0,05$). Também não há evidência de que a droga nova seja melhor do que a antiga ($p > 0,05$).

- 13.5.8 –** Estatísticas para comparar os dois métodos de processamento

<i>Estatística</i>	<i>Resultado</i>
Valor de F	1,50
p -valor	0,1924
Variância ponderada	5,000
Valor de t	10,75
p -valor (unilateral)	0,0000

Não se rejeita a hipótese de variâncias iguais ($p > 0,05$). Rejeita-se a hipótese de médias iguais ($p = 0,0000 < 0,05$).

13.5.9 – Estatísticas para comparar as duas dietas

<i>Estatística</i>	<i>Resultado</i>
Valor de <i>F</i>	1,18
<i>p</i> -valor	0,4290
Variância ponderada	2,183
Valor de <i>t</i>	-2,34
<i>p</i> -valor (unilateral)	0,0205

Não se rejeita a hipótese de variâncias iguais ($p > 0,05$). Rejeita-se a hipótese de médias iguais ($p = 0,0205 < 0,05$).

13.5.10 – Teste *t* pareado, porque a mesma criança foi observada duas vezes: a) quando recebeu alimentos adoçados com açúcar e b) quando recebeu alimentos adoçados com sacarina. Os dois grupos (de crianças mais velhas, hiperativas e de crianças mais novas, “normais”), não são comparáveis porque diferem quanto a dois fatores: idade e hiperatividade.

Tabelas

(página deixada intencionalmente em branco)

TABELA 1
Distribuição normal reduzida $P(0 < Z < z)$

	<i>Último dígito</i>									
	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2703	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4658	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990

TABELA 2
Valores de χ^2 , segundo os graus de liberdade e o valor de α

<i>Graus de liberdade</i>	α		
	10%	5%	1%
1	2,71	3,84	6,64
2	4,60	5,99	9,21
3	6,25	7,82	11,34
4	7,78	9,49	13,28
5	9,24	11,07	15,09
6	10,64	12,59	16,81
7	12,02	14,07	18,48
8	13,36	15,51	20,09
9	14,68	16,92	21,67
10	15,99	18,31	23,21
11	17,28	19,68	24,72
12	18,55	21,03	26,22
13	19,81	22,36	27,69
14	21,06	23,68	29,14
15	22,31	25,00	30,58
16	23,54	26,30	32,00
17	24,77	27,59	33,41
18	25,99	28,87	34,80
19	27,20	30,14	36,19
20	28,41	31,41	37,57
21	29,62	32,67	38,93
22	30,81	33,92	40,29
23	32,01	35,17	41,64
24	33,20	36,42	42,98
25	34,38	37,65	44,31
26	35,56	38,88	45,64
27	36,74	40,11	46,96
28	37,92	41,34	48,28
29	39,09	42,56	49,59
30	40,26	43,77	50,89

TABELA 3

Valores de F para $\alpha = 2,5\%$, segundo o número de graus de liberdade do numerador e do denominador

<i>Nº de g. l. do denominador</i>	<i>Número de graus de liberdade do numerador</i>								
	1	2	3	4	5	6	7	8	9
1	648	800	864	900	922	937	948	957	963
2	38,5	39,0	39,2	39,2	39,3	39,3	39,4	39,4	39,4
3	17,4	16,0	15,4	15,1	14,9	14,7	14,6	14,5	14,5
4	12,2	10,6	9,98	9,60	9,36	9,20	9,07	8,98	8,90
5	10,0	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68
6	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52
7	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82
8	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36
9	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03
10	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78
11	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,59
12	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44
13	6,41	4,97	4,35	4,00	3,77	3,60	3,48	3,39	3,31
14	6,30	4,86	4,24	3,89	3,66	3,50	3,38	3,29	3,21
15	6,20	4,77	4,15	3,80	3,58	3,41	3,29	3,20	3,12
16	6,12	4,69	4,08	3,73	3,50	3,34	3,22	3,12	3,05
17	6,04	4,62	4,01	3,66	3,44	3,28	3,16	3,06	2,98
18	5,98	4,56	3,95	3,61	3,38	3,22	3,10	3,01	2,93
19	5,92	4,51	3,90	3,56	3,33	3,17	3,05	2,96	2,88
20	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84
21	5,83	4,42	3,82	3,48	3,25	3,09	2,97	2,87	2,80
22	5,79	4,38	3,78	3,44	3,22	3,05	2,93	2,84	2,76
23	5,75	4,35	3,75	3,41	3,18	3,02	2,90	2,81	2,73
24	5,72	4,32	3,72	3,38	3,15	2,99	2,87	2,78	2,70
25	5,69	4,29	3,69	3,35	3,13	2,97	2,85	2,75	2,68
26	5,66	4,27	3,67	3,33	3,10	2,94	2,82	2,73	2,65
27	5,63	4,24	3,65	3,31	3,08	2,92	2,80	2,71	2,63
28	5,61	4,22	3,63	3,29	3,06	2,90	2,78	2,69	2,61
29	5,59	4,20	3,61	3,27	3,04	2,88	2,76	2,67	2,59
30	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57
40	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,45
60	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33
120	5,15	3,80	3,23	2,89	2,67	2,52	2,39	2,30	2,22
∞	5,02	3,69	3,12	2,79	2,57	2,41	2,29	2,19	2,11

continua

Continuação da Tabela 3

Nº de g. 1. do denominador	Número de graus de liberdade do numerador									
	10	12	15	20	24	30	40	60	120	∞
1	969	977	985	993	997	1.000	1.010	1.010	1.010	1.020
2	39,4	39,4	39,4	39,4	39,5	39,5	39,5	39,5	39,5	39,5
3	14,4	14,3	14,3	14,2	14,1	14,1	14,0	14,0	13,9	13,9
4	8,84	8,75	8,66	8,56	8,51	8,46	8,41	8,36	8,31	8,26
5	6,62	6,52	6,43	6,33	6,28	6,23	6,18	6,12	6,07	6,02
6	5,46	5,37	5,27	5,17	5,12	5,07	5,01	4,96	4,90	4,85
7	4,76	4,67	4,57	4,47	4,42	4,36	4,31	4,25	4,20	4,14
8	4,30	4,20	4,10	4,00	3,95	3,89	3,84	3,78	3,73	3,67
9	3,96	3,87	3,77	3,67	3,61	3,56	3,51	3,45	3,39	3,33
10	3,72	3,62	3,52	3,42	3,37	3,31	3,26	3,20	3,14	3,08
11	3,53	3,43	3,33	3,23	3,17	3,12	3,06	3,00	2,94	2,88
12	3,37	3,28	3,18	3,07	3,02	2,96	2,91	2,85	2,79	2,72
13	3,25	3,15	3,05	2,95	2,89	2,84	2,78	2,72	2,66	2,60
14	3,15	3,05	2,95	2,84	2,79	2,73	2,67	2,61	2,55	2,49
15	3,06	2,96	2,86	2,76	2,70	2,64	2,59	2,52	2,46	2,40
16	2,99	2,89	2,79	2,68	2,63	2,57	2,51	2,45	2,38	2,32
17	2,92	2,82	2,72	2,62	2,56	2,50	2,44	2,38	2,32	2,25
18	2,87	2,77	2,67	2,56	2,50	2,44	2,38	2,32	2,26	2,19
19	2,82	2,72	2,62	2,51	2,45	2,39	2,33	2,27	2,20	2,13
20	2,77	2,68	2,57	2,46	2,41	2,35	2,29	2,22	2,16	2,09
21	2,73	2,64	2,53	2,42	2,37	2,31	2,25	2,18	2,11	2,04
22	2,70	2,60	2,50	2,39	2,33	2,27	2,21	2,14	2,08	2,00
23	2,67	2,57	2,47	2,36	2,30	2,24	2,18	2,11	2,04	1,97
24	2,64	2,54	2,44	2,33	2,27	2,21	2,15	2,08	2,01	1,94
25	2,61	2,51	2,41	2,30	2,24	2,18	2,12	2,05	1,98	1,91
26	2,59	2,49	2,39	2,28	2,22	2,16	2,09	2,03	1,95	1,88
27	2,57	2,47	2,36	2,25	2,19	2,13	2,07	2,00	1,93	1,85
28	2,55	2,45	2,34	2,23	2,17	2,11	2,05	1,98	1,91	1,83
29	2,53	2,43	2,32	2,21	2,15	2,09	2,03	1,96	1,89	1,81
30	2,51	2,41	2,31	2,20	2,14	2,07	2,01	1,94	1,87	1,79
40	2,39	2,29	2,18	2,07	2,01	1,94	1,88	1,80	1,72	1,64
60	2,27	2,17	2,06	1,94	1,88	1,82	1,74	1,67	1,58	1,48
120	2,16	2,05	1,94	1,82	1,76	1,69	1,61	1,53	1,43	1,31
∞	2,05	1,94	1,83	1,71	1,64	1,57	1,48	1,39	1,27	1,00

Fonte: SCHEFFÉ (1959)

TABELA 4

Valores de F para $\alpha = 5\%$, segundo o número de graus de liberdade do numerador e do denominador

Nº de g. 1. do denominador	Número de graus de liberdade do numerador								
	1	2	3	4	5	6	7	8	9
1	161	200	216	225	230	234	237	239	241
2	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4	19,4
3	10,1	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88

continua

Continuação da Tabela 4

Nº de g. 1. do denominador	Número de graus de liberdade do numerador									
	10	12	15	20	24	30	40	60	120	∞
1	242	244	246	248	249	250	251	252	253	254
2	19,4	19,4	19,4	19,4	19,5	19,5	19,5	19,5	19,5	19,5
3	8,79	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	4,74	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36
6	4,06	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	3,64	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	3,35	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	3,14	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	2,98	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	2,85	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	2,67	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	2,60	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	2,45	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	2,41	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	2,38	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	2,32	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	2,30	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	2,27	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	2,25	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	2,24	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	2,22	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	2,20	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	2,19	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
29	2,18	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	2,08	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	1,99	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	1,91	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
∞	1,83	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00

Fonte: SCHEFFÉ (1959)

TABELA 5

Valores de F para $\alpha = 10\%$, segundo o número de graus de liberdade do numerador e do denominador

<i>Nº de g. l. do denominador</i>	<i>Número de graus de liberdade do numerador</i>								
	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	<i>7</i>	<i>8</i>	<i>9</i>
1	39,9	49,5	53,6	55,8	57,2	58,2	58,9	59,4	59,9
2	8,53	9,00	9,16	9,24	9,29	9,33	9,35	9,37	9,38
3	5,54	5,46	5,39	5,34	5,31	5,28	5,27	5,25	5,24
4	4,54	4,32	4,19	4,11	4,05	4,01	3,98	3,95	3,94
5	4,06	3,78	3,62	3,52	3,45	3,40	3,37	3,34	3,32
6	3,78	3,46	3,29	3,18	3,11	3,05	3,01	2,98	2,96
7	3,59	3,26	3,07	2,96	2,88	2,83	2,78	2,75	2,72
8	3,46	3,11	2,92	2,81	2,73	2,67	2,62	2,59	2,56
9	3,36	3,01	2,81	2,69	2,61	2,55	2,51	2,47	2,44
10	3,29	2,92	2,73	2,61	2,52	2,46	2,41	2,38	2,35
11	3,23	2,86	2,66	2,54	2,45	2,39	2,34	2,30	2,27
12	3,18	2,81	2,61	2,48	2,39	2,33	2,28	2,24	2,21
13	3,14	2,76	2,56	2,43	2,35	2,28	2,23	2,20	2,16
14	3,10	2,73	2,52	2,39	2,31	2,24	2,19	2,15	2,12
15	3,07	2,70	2,49	2,36	2,27	2,21	2,16	2,12	2,09
16	3,05	2,67	2,46	2,33	2,24	2,18	2,13	2,09	2,06
17	3,03	2,64	2,44	2,31	2,22	2,15	2,10	2,06	2,03
18	3,01	2,62	2,42	2,29	2,20	2,13	2,08	2,04	2,00
19	2,99	2,61	2,40	2,27	2,18	2,11	2,06	2,02	1,98
20	2,97	2,59	2,38	2,25	2,16	2,09	2,04	2,00	1,96
21	2,96	2,57	2,36	2,23	2,14	2,08	2,02	1,98	1,95
22	2,95	2,56	2,35	2,22	2,13	2,06	2,01	1,97	1,93
23	2,94	2,55	2,34	2,21	2,11	2,05	1,99	1,95	1,92
24	2,93	2,54	2,33	2,19	2,10	2,04	1,98	1,94	1,91
25	2,92	2,53	2,32	2,18	2,09	2,02	1,97	1,93	1,89
26	2,91	2,52	2,31	2,17	2,08	2,01	1,96	1,92	1,88
27	2,90	2,51	2,30	2,17	2,07	2,00	1,95	1,91	1,87
28	2,89	2,50	2,29	2,16	2,06	2,00	1,94	1,90	1,87
29	2,89	2,50	2,28	2,15	2,06	1,99	1,93	1,89	1,86
30	2,88	2,49	2,28	2,14	2,05	1,98	1,93	1,88	1,85
40	2,84	2,44	2,23	2,09	2,00	1,93	1,87	1,83	1,79
60	2,79	2,39	2,18	2,04	1,95	1,87	1,82	1,77	1,74
120	2,75	2,35	2,13	1,99	1,90	1,82	1,77	1,72	1,68
∞	2,71	2,30	2,08	1,94	1,85	1,77	1,72	1,67	1,63

continua

Continuação da Tabela 5

Nº de g. 1. do denominador	Número de graus de liberdade do numerador									
	10	12	15	20	24	30	40	60	120	∞
1	60,2	60,7	61,2	61,7	62,0	62,3	62,5	62,8	63,1	63,3
2	9,39	9,41	9,42	9,44	9,45	9,46	9,47	9,47	9,48	9,49
3	5,23	5,22	5,20	5,18	5,18	5,17	5,16	5,15	5,14	5,13
4	3,92	3,90	3,87	3,84	3,83	3,82	3,80	3,79	3,78	3,76
5	3,30	3,27	3,24	3,21	3,19	3,17	3,16	3,14	3,12	3,10
6	2,94	2,90	2,87	2,84	2,82	2,80	2,78	2,76	2,74	2,72
7	2,70	2,67	2,63	2,59	2,58	2,56	2,54	2,51	2,49	2,47
8	2,54	2,50	2,46	2,42	2,40	2,38	2,36	2,34	2,32	2,29
9	2,42	2,38	2,34	2,30	2,28	2,25	2,23	2,21	2,18	2,16
10	2,32	2,28	2,24	2,20	2,18	2,16	2,13	2,11	2,08	2,06
11	2,25	2,21	2,17	2,12	2,10	2,08	2,05	2,03	2,00	1,97
12	2,19	2,15	2,10	2,06	2,04	2,01	1,99	1,96	1,93	1,90
13	2,14	2,10	2,05	2,01	1,98	1,96	1,93	1,90	1,88	1,85
14	2,10	2,05	2,01	1,96	1,94	1,91	1,89	1,86	1,83	1,80
15	2,06	2,02	1,97	1,92	1,90	1,87	1,85	1,82	1,79	1,76
16	2,03	1,99	1,94	1,89	1,87	1,84	1,81	1,78	1,75	1,72
17	2,00	1,96	1,91	1,86	1,84	1,81	1,78	1,75	1,72	1,69
18	1,98	1,93	1,89	1,84	1,81	1,78	1,75	1,72	1,69	1,66
19	1,96	1,91	1,86	1,81	1,79	1,76	1,73	1,70	1,67	1,63
20	1,94	1,89	1,84	1,79	1,77	1,74	1,71	1,68	1,64	1,61
21	1,92	1,88	1,83	1,78	1,75	1,72	1,69	1,66	1,62	1,59
22	1,90	1,86	1,81	1,76	1,73	1,70	1,67	1,64	1,60	1,57
23	1,89	1,84	1,80	1,74	1,72	1,69	1,66	1,62	1,59	1,55
24	1,88	1,83	1,78	1,73	1,70	1,67	1,64	1,61	1,57	1,53
25	1,87	1,82	1,77	1,72	1,69	1,66	1,63	1,59	1,56	1,52
26	1,86	1,81	1,76	1,71	1,68	1,65	1,61	1,58	1,54	1,50
27	1,85	1,80	1,75	1,70	1,67	1,64	1,60	1,57	1,53	1,49
28	1,84	1,79	1,74	1,69	1,66	1,63	1,59	1,56	1,52	1,48
29	1,83	1,78	1,73	1,68	1,65	1,62	1,58	1,55	1,51	1,47
30	1,82	1,77	1,72	1,67	1,64	1,61	1,57	1,54	1,50	1,46
40	1,76	1,71	1,66	1,61	1,57	1,54	1,51	1,47	1,42	1,38
60	1,71	1,66	1,60	1,54	1,51	1,48	1,44	1,40	1,35	1,29
120	1,65	1,60	1,55	1,48	1,45	1,41	1,37	1,32	1,26	1,19
∞	1,60	1,55	1,49	1,42	1,38	1,34	1,30	1,24	1,17	1,00

Fonte: SCHEFFÉ (1959)

TABELA 6
Valores de t , segundo os graus de liberdade e o valor de α

<i>Graus de liberdade</i>	α		
	10%	5%	1%
1	6,31	12,71	63,66
2	2,92	4,30	9,92
3	2,35	3,18	5,84
4	2,13	2,78	4,60
5	2,02	2,57	4,03
6	1,94	2,45	3,71
7	1,90	2,36	3,50
8	1,86	2,31	3,36
9	1,83	2,26	3,25
10	1,81	2,23	3,17
11	1,80	2,20	3,11
12	1,78	2,18	3,06
13	1,77	2,16	3,01
14	1,76	2,14	2,98
15	1,75	2,13	2,95
16	1,75	2,12	2,92
17	1,74	2,11	2,90
18	1,73	2,10	2,88
19	1,73	2,09	2,86
20	1,73	2,09	2,84
21	1,72	2,08	2,83
22	1,72	2,07	2,82
23	1,71	2,07	2,81
24	1,71	2,06	2,80
25	1,71	2,06	2,79
26	1,71	2,06	2,78
27	1,70	2,05	2,77
28	1,70	2,05	2,76
29	1,70	2,04	2,76
30	1,70	2,04	2,75
40	1,68	2,02	2,70
60	1,67	2,00	2,66
120	1,66	1,98	2,62
∞	1,64	1,96	2,58

TABELA 7

Valores da amplitude total estudentizada (q) para $\alpha = 5\%$, segundo o número de tratamento (k) os graus de liberdade do resíduo

Nº de graus de lib. do resíduo	Número de tratamentos (k)																		
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	8,0	27,0	32,8	37,1	40,4	43,1	45,4	47,4	49,1	50,6	52,0	53,2	54,3	55,4	56,3	57,2	58,0	58,8	59,6
2	6,08	8,33	9,80	10,9	11,7	12,4	13,0	13,5	14,0	14,4	14,7	15,1	15,4	15,7	15,9	16,1	16,4	16,6	16,8
3	4,50	5,91	6,82	7,50	8,04	8,48	8,85	9,18	9,46	9,72	9,95	10,2	10,3	10,5	10,7	10,8	11,0	11,1	11,2
4	3,93	5,04	5,76	6,29	6,71	7,05	7,35	7,60	7,83	8,03	8,21	8,37	8,52	8,66	8,79	8,91	9,03	9,13	9,23
5	3,64	4,60	5,22	5,67	6,03	6,33	6,58	6,80	6,99	7,17	7,32	7,47	7,60	7,72	7,83	7,93	8,03	8,12	8,21
6	3,46	4,34	4,90	5,30	5,63	5,90	6,12	6,32	6,49	6,65	6,79	6,92	7,03	7,14	7,24	7,34	7,43	7,51	7,59
7	3,34	4,16	4,68	5,06	5,36	5,61	5,82	6,00	6,16	6,30	6,43	6,55	6,66	6,76	6,85	6,94	7,02	7,10	7,17
8	3,26	4,04	4,53	4,89	5,17	5,40	5,60	5,77	5,92	6,05	6,18	6,29	6,39	6,48	6,57	6,65	6,73	6,80	6,87
9	3,20	3,95	4,41	4,76	5,02	5,24	5,43	5,59	5,74	5,87	5,98	6,09	6,19	6,28	6,36	6,44	6,51	6,58	6,64
10	3,15	3,88	4,33	4,65	4,91	5,12	5,30	5,46	5,60	5,72	5,83	5,93	6,03	6,11	6,19	6,27	6,34	6,40	6,47
11	3,11	3,82	4,26	4,57	4,82	5,03	5,20	5,35	5,49	5,61	5,71	5,81	5,90	5,98	6,06	6,13	6,20	6,27	6,33
12	3,08	3,77	4,20	4,51	4,75	4,95	5,12	5,27	5,39	5,51	5,61	5,71	5,80	5,88	5,95	6,02	6,09	6,15	6,21
13	3,06	3,73	4,15	4,45	4,69	4,88	5,05	5,19	5,32	5,43	5,53	5,63	5,71	5,79	5,86	5,93	5,99	6,05	6,11
14	3,03	3,70	4,11	4,41	4,64	4,83	4,99	5,13	5,25	5,36	5,46	5,55	5,64	5,71	5,79	5,85	5,91	5,97	6,03

TABELA 7 (cont.)**Valores da amplitude total estudentizada (q) para $\alpha = 5\%$, segundo o número de tratamento (k) os graus de liberdade do resíduo**

Nº de graus de lib. do resíduo	Número de tratamentos (k)																		
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
15	3,01	3,67	4,08	4,37	4,59	4,78	4,94	5,08	5,20	5,31	5,40	5,49	5,57	5,65	5,72	5,78	5,85	5,90	5,96
16	3,00	3,65	4,05	4,33	4,56	4,74	4,90	5,03	5,15	5,26	5,35	5,44	5,52	5,59	5,66	5,73	5,79	5,84	5,90
17	2,98	3,63	4,02	4,30	4,52	4,70	4,86	4,99	5,11	5,21	5,31	5,39	5,47	5,54	5,61	5,67	5,73	5,79	5,84
18	2,97	3,61	4,00	4,28	4,49	4,67	4,82	4,96	5,07	5,17	5,27	5,35	5,43	5,50	5,57	5,63	5,69	5,74	5,79
19	2,96	3,59	3,98	4,25	4,47	4,65	4,79	4,92	5,04	5,14	5,23	5,31	5,39	5,46	5,53	5,59	5,65	5,70	5,75
20	2,95	3,58	3,96	4,23	4,45	4,62	4,77	4,90	5,01	5,11	5,20	5,28	5,36	5,43	5,49	5,55	5,61	5,66	5,71
24	2,92	3,53	3,90	4,17	4,37	4,54	4,68	4,81	4,92	5,01	5,10	5,18	5,25	5,32	5,38	5,44	5,49	5,55	5,59
30	2,89	3,49	3,85	4,10	4,30	4,46	4,60	4,72	4,82	4,92	5,00	5,08	5,15	5,21	5,27	5,33	5,38	5,43	5,47
40	2,86	3,44	3,79	4,04	4,23	4,39	4,52	4,63	4,73	4,82	4,90	4,98	5,04	5,11	5,16	5,22	5,27	5,31	5,36
60	2,83	3,40	3,74	3,98	4,16	4,31	4,44	4,55	4,65	4,73	4,81	4,88	4,94	5,00	5,06	5,11	5,15	5,20	5,24
120	2,80	3,36	3,68	3,92	4,10	4,24	4,36	4,47	4,56	4,64	4,71	4,78	4,84	4,90	4,95	5,00	5,04	5,09	5,13
∞	2,77	3,31	3,63	3,86	4,03	4,17	4,29	4,39	4,47	4,55	4,62	4,68	4,74	4,80	4,85	4,89	4,93	4,97	5,01

Fonte: SCHEFFÉ, (1959)

TABELA 8

Valores da amplitude total estudentizada (q) para $\alpha = 10\%$, segundo o número de tratamento (k) e os graus de liberdade do resíduo

Nº de graus de lib. do resíduo	Número de tratamentos (k)																		
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	8,93	13,4	16,4	18,5	20,2	21,5	22,6	23,6	24,5	25,2	25,9	26,5	27,1	27,6	28,1	28,5	29,0	29,3	29,7
2	4,13	5,73	6,77	7,54	8,14	8,63	9,05	9,41	9,72	10,0	10,3	10,5	10,7	10,9	11,1	11,2	11,4	11,5	11,7
3	3,33	4,47	5,20	5,74	6,16	6,51	6,81	7,06	7,29	7,49	7,67	7,83	7,98	8,12	8,25	8,37	8,48	8,58	8,68
4	3,01	3,98	4,59	5,03	5,39	5,68	5,93	6,14	6,33	6,49	6,65	6,78	6,91	7,02	7,13	7,23	7,33	7,41	7,50
5	2,85	3,72	4,26	4,66	4,98	5,24	5,46	5,65	5,82	5,97	6,10	6,22	6,34	6,44	6,54	6,63	6,71	6,79	6,86
6	2,75	3,56	4,07	4,44	4,73	4,97	5,17	5,34	5,50	5,64	5,76	5,87	5,98	6,07	6,16	6,25	6,32	6,40	6,47
7	2,68	3,45	3,93	4,28	4,55	4,78	4,97	5,14	5,28	5,41	5,53	5,64	5,74	5,83	5,91	5,99	6,06	6,13	6,19
8	2,63	3,37	3,83	4,17	4,43	4,65	4,83	4,99	5,13	5,25	5,36	5,46	5,56	5,64	5,72	5,80	5,87	5,93	6,00
9	2,59	3,32	3,76	4,08	4,34	4,54	4,72	4,87	5,01	5,13	5,23	5,33	5,42	5,51	5,58	5,66	5,72	5,79	5,85
10	2,56	3,27	3,70	4,02	4,26	4,47	4,64	4,78	4,91	5,03	5,13	5,23	5,32	5,40	5,47	5,54	5,61	5,67	5,73
11	2,54	3,23	3,66	3,96	4,20	4,40	4,57	4,71	4,84	4,95	5,05	5,15	5,23	5,31	5,38	5,45	5,51	5,57	5,63
12	2,52	3,20	3,62	3,92	4,16	4,35	4,51	4,65	4,78	4,89	4,99	5,08	5,16	5,24	5,31	5,37	5,44	5,49	5,55
13	2,50	3,18	3,59	3,88	4,12	4,30	4,46	4,60	4,72	4,83	4,93	5,02	5,10	5,18	5,25	5,31	5,37	5,43	5,48
14	2,49	3,16	3,56	3,85	4,08	4,27	4,42	4,56	4,68	4,79	4,88	4,97	5,05	5,12	5,19	5,26	5,32	5,37	5,43

TABELA 8 (cont.)Valores da amplitude total estudentizada (q) para $\alpha = 10\%$, segundo o número de tratamento (k) e os graus de liberdade do resíduo

Nº de graus de lib. do resíduo	Número de tratamentos (k)																		
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
15	2,48	3,14	3,54	3,83	4,05	4,23	4,39	4,52	4,64	4,75	4,84	4,93	5,01	5,08	5,15	5,21	5,27	5,32	5,38
16	2,47	3,12	3,52	3,80	4,03	4,21	4,36	4,49	4,61	4,71	4,81	4,89	4,97	5,04	5,11	5,17	5,23	5,28	5,33
17	2,46	3,11	3,50	3,78	4,00	4,18	4,33	4,46	4,58	4,68	4,77	4,86	4,93	5,01	5,07	5,13	5,19	5,24	5,30
18	2,45	3,10	3,49	3,77	3,98	4,16	4,31	4,44	4,55	4,65	4,75	4,83	4,90	4,98	5,04	5,10	5,16	5,21	5,26
19	2,45	3,09	3,47	3,75	3,97	4,14	4,29	4,42	4,53	4,63	4,72	4,80	4,88	4,95	5,01	5,07	5,13	5,18	5,23
20	2,44	3,08	3,46	3,74	3,95	4,12	4,27	4,40	4,51	4,61	4,70	4,78	4,85	4,92	4,99	5,05	5,10	5,16	5,20
24	2,42	3,05	3,42	3,69	3,90	4,07	4,21	4,34	4,44	4,54	4,63	4,71	4,78	4,85	4,91	4,97	5,02	5,07	5,12
30	2,40	3,02	3,39	3,65	3,85	4,02	4,16	4,28	4,38	4,47	4,56	4,64	4,71	4,77	4,83	4,89	4,94	4,99	5,03
40	2,38	2,99	3,35	3,60	3,80	3,96	4,10	4,21	4,32	4,41	4,49	4,56	4,63	4,69	4,75	4,81	4,86	4,90	4,95
60	2,36	2,96	3,31	3,56	3,75	3,91	4,04	4,16	4,25	4,34	4,42	4,49	4,56	4,62	4,67	4,73	4,78	4,82	4,86
120	2,34	2,93	3,28	3,52	3,71	3,86	3,99	4,10	4,19	4,28	4,35	4,42	4,48	4,54	4,60	4,65	4,69	4,74	4,78
∞	2,33	2,90	3,24	3,48	3,66	3,81	3,93	4,04	4,13	4,21	4,28	4,35	4,41	4,47	4,52	4,57	4,61	4,65	4,69

Fonte: SCHEFFÉ, (1959)

(página deixada intencionalmente em branco)

Sugestões para leitura

- ALIAGA, M. e GUNDERSON, B. **Interactive Statistics.** New Jersey, Prentice Hall, 2 ed. 2003.
- ARMITAGE, P. **Statistical methods in medical research.** Oxford, Blackwell Scientific Publications, 1971.
- BLAND, M. **An introduction to medical statistics.** Oxford, Oxford Medical Publications, 1987.
- BROWN, B.W. e HOLLANDER, M. **Statistics: a biomedical introduction.** New York, Wiley, 1977.
- BISHOP, V.M.M. et alii. **Discrete multivariate analysis, theory and practice.** Cambridge, MIT Press, 1977.
- BUSSAB, W.e MORETTIN, P. A. **Estatística Básica.** São Paulo: Saraiva. 2002.
- COCHRAN, W. **Sampling techniques.** New York, Wiley, 1977.
- CHOW, S. C. e LIU, J.L. **Design and analysis of clinical trials.** New York, Wiley, 2004.
- DANIEL, C. **Applications of Statistics.** New York, Wiley. 1976.
- DANIEL, W.W. **Biostatistics: a foundation for analysis in the health sciences.** New York, Wiley, 1987.
- DAWSON, B., TRAPP, R.G. **Bioestatística básica e clínica.** Rio de Janeiro, McGraw, 3 ed. 1994.
- DEAN, A., VOSS, D. **Design and analysis of experiments.** New York, Springer, 1999.
- ELSTON, R.C. e JOHNSON, W.D. **Essentials of biostatistics.** Philadelphia, F.A. Davis Company, 1987.
- FREUND, J. E. E SMITH, R. M. **Statistics: a first course.** Englewood Cliffs, Prentice Hall, 4 ed. 1986.
- GLANTZ, S.A. **Primer of biostatistics.** New York, McGraw, 1987.
- JOHNSON, R. E TSUI, K. W. **Statistical reasoning and methods.** Nova York, Wiley, 1998.
- LOHR, S. L. **Sampling: Design and analysis.** Pacific Grove, Brooks, 1999.
- MATTHEWS, D.E. e FAREWELL, V. **Using and understanding medical statistics.** New York. Karger, 1985.
- MINIUM, E. W., CLARKE, R. C., COLADARCI, T. **Elements of Statistical Reasoning.** New York, Wiley, 2 ed. 1999.
- MOTULSKY, H. **Intuitive Biostatistics.** New York, Oxford Press, 1995.
- OTT, L e Mendenhall, W. **Understanding Statistics.** Belmont, Wadsworth,, 6 ed. 1994.

- SCHORK, M. A. e REMINGTON, R. D. **Statistics with applications to the biological and health sciences.** New Jersey, Prentice Hall, 3 ed. 2000.
- VIEIRA, S. **Elementos de Estatística,** São Paulo, Atlas, 5 ed. 2003.
- VIEIRA, S. **Bioestatística: Tópicos Avançados.** Rio de Janeiro, Campus-Elsevier, 2 ed., 5^a tiragem.2008.
- VIEIRA, S. E HOSSNE, W. S. **Metodologia científica para a área de saúde.** São Paulo, Rio de Janeiro, Campus-Elsevier, .
- VIEIRA, S. **Análise de variância.** São Paulo, Atlas.2006.
- VIEIRA, S. e HOSSNE, W. S. **Experimentação com seres humanos.** São Paulo, Moderna, 3 ed, 1988.
- ZAR, J. H. **Biostatistical analysis.** New Jersey, Prentice Hall, 4.ed. 1999.

Índice

A

- Amostra, 4
 - aleatória estratificada, 6
 - aleatória ou probabilística, 5, 9
 - não-probabilística ou de conveniência, 9
 - não-representativa, 13
 - por conglomerados, 7, 9
 - por quotas, 8, 9
 - razões de trabalhar com, 4-5
 - representativa, 13
 - semiprobabilística, 6
 - sistemática, 7, 9
 - tendenciosa, 13
- Amplitude, 34, 87
- Apresentação de dados numéricos, 31, 56
- Apresentação de dados qualitativos, 28, 49
- Apuração de dados, 24
- Áreas sob a curva normal, 210
- Associação positiva, 259

C

- Cabeçalho, tabela, 26
- Cálculo de probabilidade, 164
- Cálculo de probabilidade condicional, 171
- Cálculo de probabilidades na distribuição binomial, 194
- Cálculo do intervalo de confiança para uma média, 236
- Cálculo do intervalo de confiança para uma proporção, 230, 231
- Cálculo do número de classes, 37
- Cálculo dos coeficientes de regressão, 139
- Caracterização da distribuição binomial, 192
- Caudas da distribuição, 213
- Classe modal, 76
- Coeficiente angular da reta, 137
- Coeficiente de correlação, 115
- Coeficiente de correlação de Pearson, 115
- Coeficiente de determinação, 143, 144, 145
- Coeficiente de variação, 98
- Coeficiente de Yule, 259
- Coeficientes de associação, 259
- Coeficientes de regressão, cálculo dos, 139
- Coluna indicadora, tabela, 26
- Colunas, tabela, 26
- Comparação de variâncias, 281
- Componentes das tabelas, 26
- Condição de independência, 167
- Corpo, tabela, 26

Correlação, 185

- de Pearson, coeficiente de, 115
- forte, 111, 112
- fraca, 111, 112
- negativa, 109
- nula, 113
- perfeita, 111, 112
- positiva, 109

Cuidados na interpretação dos intervalos de confiança, 237

D

- Dados, 23
- Dados contínuos, 33
- Dados discrepantes, 74
- Dados discretos, 32
- Dados numéricos, apresentação de, 31, 56
- Dados qualitativos, apresentação de, 28
- Dados, apuração de, 24
- Desvio padrão, 95, 209, 211, 212
- Desvio padrão da amostra, 93
- Diagrama de caixa (Box plot), 91
- Diagrama de dispersão, 109, 149
- Diagrama de linhas, 56
- Dispersão, 87
- Dispersão relativa, 98
- Distância interquartílica, 90
- Distribuição binomial, 189, 193, 194
 - cálculo de probabilidades na, 194
 - caracterização da, 192
 - média na, 194
 - variância na, 194
- Distribuição de freqüências, 189
- Distribuição de Gauss, 208
- Distribuição de probabilidades, 187, 188
- Distribuição normal, 208, 209
 - padronizada, 213
 - reduzida, 213
 - usos da, 219
- Distribuição teórica, 208
- Distribuições empíricas, 207

E

- Ensaio com dados pareados, 272
- Equação da reta, 137
- Erro, 249
- Erro padrão da média, 233, 236, 239
- Erro tipo I, 249
- Erros, definindo os, 249

Escolha da variável explanatória, 142
 Espaço amostral, 163
 Estatística, definição, 3, 10
 Estimativas de probabilidade, 209
 freqüência relativa como, 164, 165
 Evento, 163
 Evento certo, 164
 Evento impossível, 164
 Eventos dependentes, 175
 Eventos independentes, 166, 168, 174, 175
 diferença de eventos mutuamente exclusivos, 170
 Eventos mutuamente exclusivos, 166
 diferença de eventos independentes, 170
 Extrapolação, 140
 Extremos de classe, 35

F

Fonte e notas, tabela, 27
 Freqüência esperada, 253
 Freqüência relativa, 29, 30
 como estimativa de probabilidade, 164, 165

G

Gráfico de barras, 49
 com 3 D, 52
 com grades, 51
 com percentuais nas barras, 51
 horizontais, 52
 Gráfico de linhas, 133
 Gráfico de pontos, 57
 Gráfico de setores, 54
 em 3D, 55
 Grau de associação, 259
 Grau de correlação linear, 115
 Graus de liberdade, 95, 238

H

Hipótese alternativa, 248
 Hipótese da nulidade, 248
 Hipóteses, 247
 Histograma, 57, 58

I

Inferência, 248
 Inferência estatística, 249
 Interpretando o *p*-valor, 250
 Intervalo de classe, 34
 Intervalo de confiança
 cuidados na interpretação dos, 237
 para uma média, 233
 para uma proporção, 230

L

Levantamento de dados, 3

Limites dos intervalos de classe, 35
 Linhas, tabela, 26

M

Margem de erro, 12, 229, 232
 Máximo, 87
 Média, 185, 209, 211, 212, 229
 Média aritmética, 68
 Média da amostra, 68
 Média da população, 239
 Média dos quadrados dos desvios, 95
 Média na distribuição binomial, 194
 Mediana da amostra, 74
 Medida da associação, 259
 Medida de variabilidade, 87
 Medidas de tendência central, 67
 Mínimo, 87
 Moda da amostra, 75

N

Nível de confiança, 12
 Nível de significância, 256
 Nível de significância do teste, 253
 Notação de somatório, 68
 Número de classes, 37

P

Parâmetros, definição, 10
 Polígono de freqüências, 58
 População, 4
 População infinita, 209
 Precisão, 236
 Probabilidade
 associada à distribuição normal, 213
 cálculo de, 164
 condicional, 170, 171, 172
 definição clássica de, 163
 distribuição de, 187, 188
 na distribuição normal reduzida, 215, 216
 na distribuição normal, 216, 217
 Proporção (freqüência relativa), 29
p-valor, 250

Q

Qualidade de uma estimativa, 11
 Quartil, 89

R

Regra do “e”, 167, 174
 Regra do “ou”, 166, 173
 Regressão
 linear simples, 151
 múltipla, 151
 não-linear, 147
 Relação não-linear entre duas variáveis, 114

Relações determinísticas, 144
 Relações entre variáveis, 109
 Relações probabilísticas, 144
 Representatividade, 13
 Reta de regressão, 135
 traçado da, 140

S

Soma de quadrados dos desvios, 94
 Somatório, notação de, 68

T

Tabela de distribuição de freqüências, 28
 Tabela de distribuição de t , 238
 Tabela de distribuição normal reduzida, 214
 Tabelas 2 X 2, 256
 Tabelas de contingência, 30
 Tabelas de distribuição de freqüências, 32, 33
 Tabelas, componentes das, 26
 Tamanho da amostra, 11
 Tendência, 13
 Tendência central, medidas de, 67
 Teorema da soma, 173
 Teorema do produto, 174
 Teoria das probabilidades, 164
 Teste de χ^2 para independência, 256, 258
 Teste de aderência, 252
 Teste de hipóteses, 247
 Teste F , 281, 283
 Teste t
 na comparação de dois grupos
 independentes, 279

para dados pareados, 272
 para o coeficiente de correlação, 285
 Testes bilaterais, 276
 Testes unilaterais, 276
 Título, tabela, 26
 Traçado da reta de regressão, 140
 Traços horizontais, tabela, 27
 Traços verticais, tabela, 27
 Transformação logarítmica, 150

V

Valor máximo, 34
 Valor mínimo, 34
 Variabilidade, 87, 185, 236
 Variância, 93, 94
 da média, 234
 na distribuição binomial, 194
 desiguais, 281
 Variável
 categorizada, 23
 contínua, 24
 dependente, 133
 discreta, 24
 explanatória, 133
 nominal, 24
 numérica, 23
 ordinal, 24
 qualitativa, 23
 quantitativa, 23
 Variável aleatória, 185