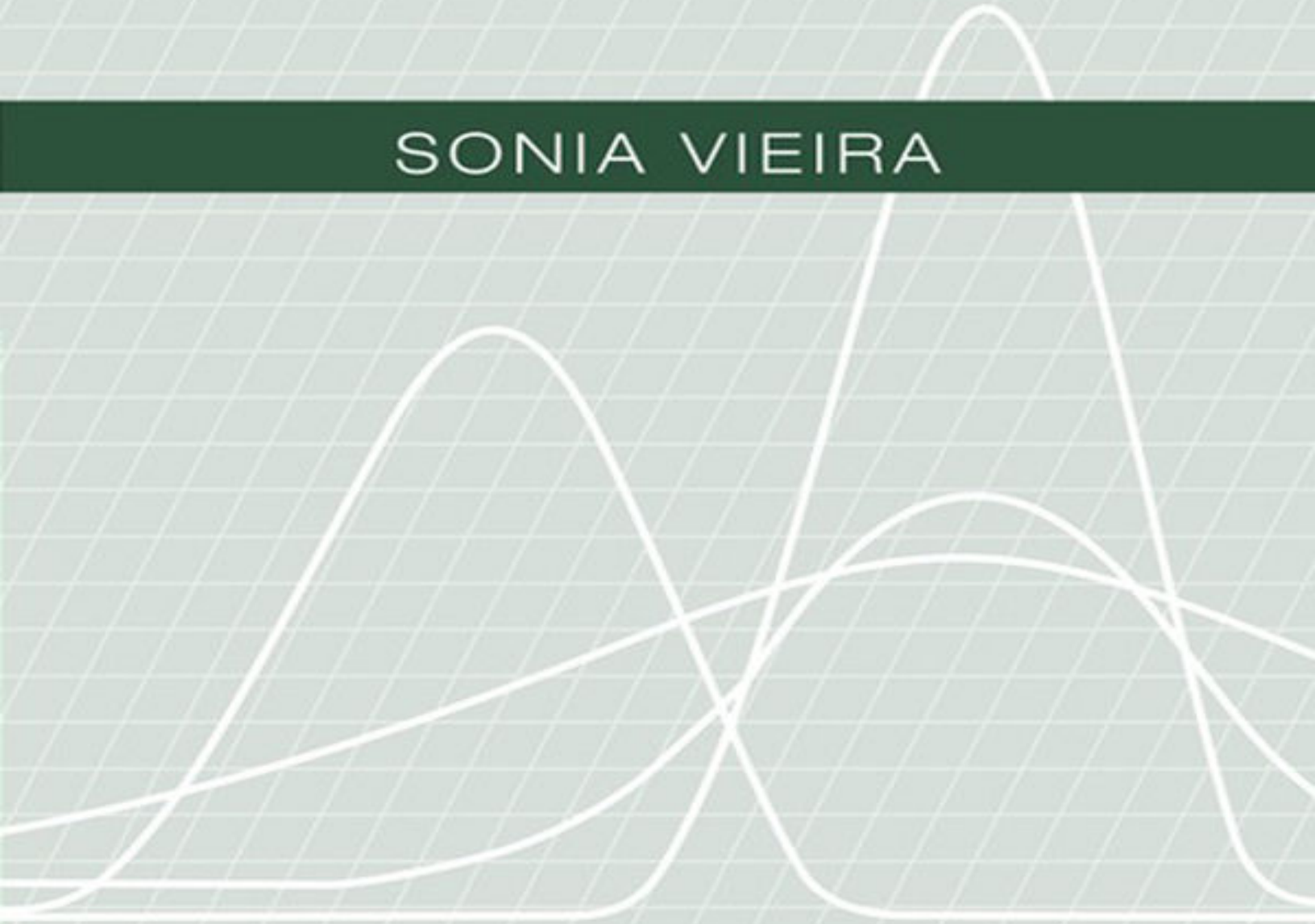
SONIA VIEIRA



Introdução à Bioestatística

5ª Edição

ELSEVIER

Introdução à Bioestatística

5ª EDIÇÃO

Sonia Vieira

Doutora em Estatística pela USP
Livre-docente em Bioestatística pela Unicamp
Professora do Curso de Mestrado e Doutorado de Bioética no Centro Universitário São Camilo, SP
Coordenadora do Comitê de Ética em Pesquisa no Centro de Pesquisas Odontológicas São Leopoldo Mandic, SP

ELSEVIER

Sumário

Capa

Folha de rosto

Copyright

Prefácio

Apresentação

Capítulo 1: Apresentação de Dados em Tabelas

- 1.1 Dados e variáveis
- 1.2 Apuração de dados
- 1.3 Normas para a construção de tabelas
- 1.4 Exercícios resolvidos
- 1.5 Exercícios propostos

Capítulo 2: Apresentação de Dados em Gráficos

- 2.1 Apresentação de dados qualitativos
- 2.2 Apresentação de dados quantitativos
- 2.3 Exercícios resolvidos
- 2.4 Exercícios propostos

Capítulo 3: Medidas de Tendência Central

- 3.1 Símbolos matemáticos
- 3.2 Média aritmética
- 3.3 Mediana
- 3.4 Moda
- 3.5 Exercícios resolvidos
- 3.6 Exercícios propostos

Capítulo 4: Medidas de Dispersão para uma Amostra

- 4.1 Mínimo, máximo e amplitude
- 4.2 Quartil
- 4.3 Desvio padrão
- 4.4 Coeficiente de variação
- 4.5 Exercícios resolvidos
- 4.6 Exercícios propostos

Capítulo 5: Noções sobre Correlação

- 5.1 Diagrama de dispersão
- 5.2 Cálculo do coeficiente de correlação
- 5.3 Cuidados na interpretação do coeficiente de correlação
- 5.4 Gráfico de linhas
- 5.5 Exercícios resolvidos
- 5.6 Exercícios propostos

Capítulo 6: Noções sobre Regressão

6.1 Regressão linear simples

6.2 Extrapolação

6.3 Escolha da variável explanatória

6.4 Coeficiente de determinação

*6.5 Regressão não linear

6.6 Exercícios resolvidos

6.7 Exercícios propostos

Capítulo 7: Noções sobre Amostragem

7.1 População e amostra

7.2 Parâmetros e estatísticas

7.3 Razões para o uso de amostras

7.4 Métodos de amostragem

7.5 Noções sobre o tamanho das amostras

7.6 A questão da representatividade

7.7 Exercícios resolvidos

7.8 Exercícios propostos

Capítulo 8: Distribuição Normal

8.1 Variável aleatória

8.2 Distribuição normal: características

8.3 Soma de variáveis aleatórias independentes

8.4 Probabilidades associadas à distribuição normal

*8.5 Distribuição normal reduzida ou padronizada

*8.6 Cálculo das probabilidades sob a distribuição normal

8.7 Usos da distribuição normal

8.8 Exercícios resolvidos

8.9 Exercícios propostos

Capítulo 9: Intervalo de Confiança

9.1 Erro padrão da média

9.2 Distribuição das médias das amostras

9.3 Cálculo do intervalo de confiança para uma média

9.4 Outras maneiras de estabelecer intervalos

9.5 Cuidados na interpretação dos intervalos de confiança

9.6 Exercícios resolvidos

9.7 Exercícios propostos

Capítulo 10: Teste t para uma Amostra

10.1 Tomada de decisão em condições de incerteza

10.2 Teste estatístico

10.3 Exercícios resolvidos

10.4 Exercícios propostos

Capítulo 11: Teste t para a Comparação de Médias

11.1 Teste t nos estudos com dados pareados

11.2 Teste t na comparação de grupos independentes

11.3 Exercícios resolvidos

11.4 Exercícios propostos

Capítulo 12: Teste $\chi 2$ para Variáveis Qualitativas

12.1 Teste χ2 para a associação de duas variáveis

12.2 Teste χ2 para comparar dois grupos em ensaios clínicos

12.3 Teste χ2 nos estudos prospectivos e retrospectivos

12.4 Risco relativo e razão de chances

12.5 Teste de uma proporção

12.6 Exercícios resolvidos

12.7 Exercícios propostos

Apêndices

Apêndice Capítulo 13: Probabilidades

Apêndice Capítulo 14: Distribuição Binomial

Anexos

Anexos Capítulo 15: Tabelas

Respostas aos Exercícios Propostos

Sugestões para leitura

Índice remissivo

Copyright

ISBN: 978-85-352-7716-6
ISBN (versão eletrônica): 978-85-352-8399-0

Capa
Olga Loureiro

Editoração Eletrônica
Thomson Digital

Elsevier Editora Ltda.

Conhecimento sem Fronteiras
Rua Sete de Setembro, n° 111 – 16° andar
20050-006 – Centro – Rio de Janeiro – RJ
Rua Quintana, n° 753 – 8° andar
04569-011 – Brooklin – São Paulo – SP
Serviço de Atendimento ao Cliente
0800 026 53 40
atendimento1@elsevier.com
Consulte nosso catálogo completo, os últimos lançamentos e os serviços exclusivos no site www.elsevier.com.br

Nota

Como as novas pesquisas e a experiência ampliam o nosso conhecimento, pode haver necessidade de alteração dos métodos de pesquisa, das práticas profissionais ou do tratamento médico. Tanto médicos quanto pesquisadores devem sempre basear-se em sua própria experiência e conhecimento para avaliar e empregar quaisquer informações, métodos, substâncias ou experimentos descritos neste texto. Ao utilizar qualquer informação ou método, devem ser criteriosos com relação a sua própria segurança ou a segurança de outras pessoas, incluindo aquelas sobre as quais tenham responsabilidade profissional.

Com relação a qualquer fármaco ou produto farmacêutico especificado, aconselha-se o leitor a cercar-se da mais atual informação fornecida (i) a respeito dos procedimentos descritos, ou (ii) pelo fabricante de cada produto a ser administrado, de modo a certificar-se sobre a dose recomendada ou a fórmula, o método e a duração da administração, e as contraindicações. É responsabilidade do médico, com base em sua experiência pessoal e no conhecimento de seus pacientes, determinar as posologias e o melhor tratamento para cada paciente individualmente, e adotar todas as precauções de segurança apropriadas.

Para todos os efeitos legais, nem a Editora, nem autores, nem editores, nem tradutores, nem revisores ou colaboradores, assumem qualquer responsabilidade por qualquer efeito danoso e/ou malefício a pessoas ou propriedades envolvendo responsabilidade, negligência etc. de produtos, ou advindos de qualquer uso ou emprego de quaisquer métodos, produtos, instruções ou ideias contidos no material aqui publicado.

O Editor

CIP-BRASIL. CATALOGAÇÃO NA PUBLICAÇÃO

SINDICATO NACIONAL DOS EDITORES DE LIVROS, RJ
V713i
5. ed.

Vieira, Sonia
Introdução à bioestatística / Sonia Vieira. - 5. ed. - Rio de Janeiro : Elsevier, 2016.
il. ; 23 cm.
Apêndice
Inclui índice remissivo
Inclui anexo
ISBN 978-85-352-7716-6
1. Bioestatística. . I. Título.
15-25725 CDD: 570.15195
CDU: 57.087.1

Prefácio

Profissionais das ciências da saúde, pesquisadores ou não, precisam saber Bioestatística. Pesquisadores, porque a Bioestatística é um dos fundamentos do trabalho científico e da pesquisa; e não pesquisadores, porque, sem ela, não conseguem avaliar, de forma crítica, o que lhes é oferecido nas publicações e nos textos.

A Bioestatística não só nos leva a aceitar ou rejeitar respostas a perguntas e dúvidas formuladas em nossa atividade investigativa e profissional, como também – e sobretudo – nos faz aprender como formular adequadamente as perguntas, sem o que não se chega à devida resposta.

Sonia Vieira, nome consagrado e respeitado na área, consegue, nesta nova edição de *Introdução à Bioestatística*, assim como nos demais livros de sua autoria, cativar o leitor já nas primeiras frases, levando-o a caminhar com satisfação na busca do conhecimento, mesmo em uma área à qual se atribui (sem razão, aliás) certa aridez.

O estilo leve, mas profundo, sóbrio e preciso, elegante e instigante da autora vai fazendo o leitor engajar-se e entusiasmar-se pela Bioestatística.

Professor ou aluno, iniciante ou veterano, pesquisador ou não, profissional da saúde e de campos afins, encontram, neste livro, condições para mais bem ensinar e para mais bem aprender.

Isso será feito com satisfação e com o sentimento de estar adquirindo mais saber e mais sabedoria.

William Saad Hossne, *Professor Emérito da Faculdade de Medicina de Botucatu (Unesp)*
Coordenador do Programa de Pós-graduação (Bioética) do Centro Universitário São Camilo
Fundador e Ex-presidente da Sociedade Brasileira de Bioética
Ex-diretor Científico da FAPESP (1964-1968 e 1975-1979)
Ex-reitor da Universidade Federal de São Carlos

Apresentação

O interesse de profissionais e alunos das áreas de saúde em Bioestatística se explica pelo uso significativo das técnicas estatísticas em pesquisa científica. Mas Bioestatística é uma ciência complexa, que não se aprende com uma simples busca de alguns poucos termos na Internet. Então, 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. Mas a leitura, a condução e a avaliação de uma pesquisa dependem, em boa parte, do conhecimento do pesquisador sobre as potencialidades e as limitações das técnicas estatísticas utilizadas. E, entre o cálculo e a interpretação do resultado, há um caminho a percorrer.

Este livro foi escrito e reescrito muitas vezes, na tentativa de facilitar a aprendizagem. Os conceitos são transmitidos mais pela intuição do que pela demonstração, sempre enfatizando as indicações e as restrições das técnicas estatísticas. Os exemplos na área da saúde, em grande quantidade, podem ser acompanhados passo a passo, com pouco trabalho de cálculo feito manualmente ou com o auxílio de calculadoras. É verdade que o uso dos computadores já se generalizou, mas quem se inicia no estudo da Estatística deve ver a fórmula para, assim, entender o conceito. Não há como ter completa segurança na discussão de uma média aritmética, por exemplo, sem nunca ter usado papel e lápis para fazer o cálculo.

A leitura do texto não demanda conhecimentos de Matemática além daqueles que são exigidos em exames vestibulares. De qualquer modo, as seções que envolvem maior gosto e aptidão para a Matemática foram assinaladas com asterisco. Tais seções podem ser evitadas sem prejuízo do entendimento das subsequentes. Assim, sem despender muito tempo com cálculos e demonstrações, o

estudante adquire, neste livro, conhecimentos suficientes para se tornar usuário competente das técnicas estatísticas mais comuns.

Uma consequência importante de se 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 a uma interpretação equivocada quando substitui o significado técnico.

A quinta edição de *Introdução à Bioestatística* 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 também é o fato de este livro ter contado com a competente e altamente especializada revisão de Martha Maria Mischan e William Saad Hossne. Ronaldo Wada fez alguns dos vários gráficos e Márcio Vieira Hoffmann fez uma leitura crítica dos originais. Também agradecemos à Editora Elsevier, pela confiança em nosso trabalho.

A autora

CAPÍTULO 1

Apresentação de Dados em Tabelas

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

O uso tão difundido de *levantamento de dados* – que, no Brasil, chamamos popularmente de "pesquisa" – faz pensar que esse trabalho é fácil. Por conta disso, ao ler um relatório de pesquisa no jornal da cidade, muita gente se considera capaz de fazer o mesmo ou 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.

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

Os estatísticos trabalham com informações. Na área de saúde, interessam informações sobre eficiência de medicamentos, causas de morte, prevalência de doenças etc. Neste capítulo, vamos aprender como essas informações são organizadas para facilitar a leitura e o entendimento. Mas, antes, é preciso saber o que são dados e o que são variáveis.

1.1 Dados e variáveis

Variável é uma condição ou característica das unidades da população.

As variáveis assumem valores diferentes em diferentes unidades. Por exemplo, se você perguntar a idade de algumas pessoas de sua família, verá valores diferentes entre si, embora todos se refiram à mesma variável: idade. Não há interesse em se levantarem constantes. Assim, não há interesse em se coletarem informações sobre analfabetismo entre universitários porque todos os estudantes universitários são alfabetizados.

Dado estatístico é toda informação coletada e registrada que se refere a uma variável.

Exemplo 1.1 Dados e variáveis

Um professor de Educação Física trabalha em uma academia de ginástica e quer saber a opinião dos clientes sobre a qualidade de seus serviços. A variável de interesse, nesse caso, é a opinião dos clientes. Os dados serão obtidos quando o professor pedir aos clientes que deem uma nota aos serviços que utilizam. Se for pedido que o cliente dê uma nota de zero a 5, os dados coletados poderão ser, por exemplo, 4, 3, 2, 3, 4, 1 etc., por serviço.

As variáveis são classificadas, conforme mostra o organograma da Figura 1.1, em dois tipos:

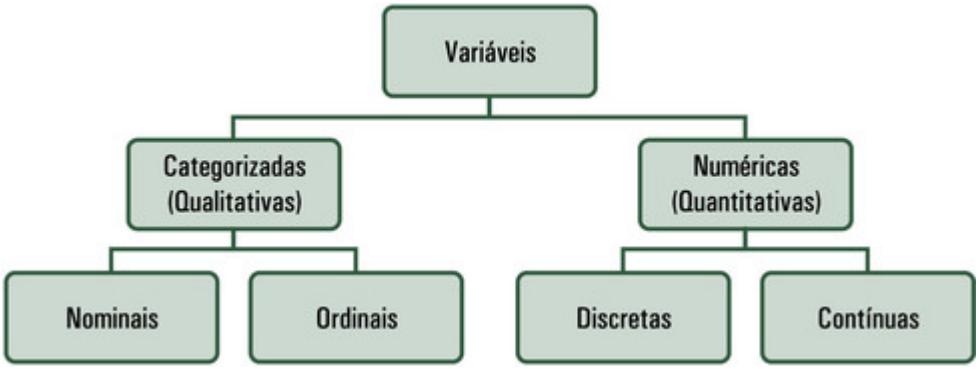


FIGURA 1.1 Tipos de variáveis

- quantitativas ou numéricas;
- qualitativas ou categorizadas.

Uma variável é *qualitativa ou categorizada* quando os dados são distribuídos em categorias mutuamente exclusivas, como sexo (masculino ou feminino), tipo de sangue (O, A, B, AB), 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 como idade, estatura, número de alunos de uma escola, número de comprimidos em uma 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 nomeadas em qualquer ordem. São variáveis nominais: cor de cabelos (loiro, castanho, preto, ruivo), tipo de sangue (O, A, B, AB), não ter ou ter determinada doença.

A variável é *ordinal* quando os dados são distribuídos em categorias mutuamente exclusivas que têm ordem 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 dado intervalo. São variáveis discretas: número de filhos (nenhum, 1, 2, 3, 4, 5 ou mais), quantidade de visitas ao médico no último ano (zero, 1, 2, 3, 4 ou mais), número de pessoas na fila de espera de um serviço de saúde.

A variável *contínua* assume qualquer valor em dado intervalo. São variáveis contínuas: peso, temperatura corporal, pressão sanguínea.

1.2 Apuração de dados

Dados são registrados em fichas, cadernos, computadores, mas depois é preciso proceder à *apuração*. Se a variável for qualitativa, a apuração se resume a uma simples contagem.

Exemplo 1.2 Apuração de dados qualitativos

Para obter a porcentagem de recém-nascidos de cada sexo, em uma maternidade, um pesquisador obteve 1.000 prontuários de recém-nascidos e escreveu numa folha de papel:

Masculino
Feminino

Em seguida, examinou os prontuários e fez um traço na linha que correspondia ao sexo do recém-nascido, para cada prontuário. Cada quadrado cortado pela diagonal representa cinco recém-nascidos. O total é dado pelo número de traços em cada linha.

Masculino ⊠⊠ ... ⊠□ = 509
Feminino ⊠⊠ ... ⊠| = 491

Quando a variável é quantitativa, é preciso anotar, na apuração, cada valor observado.

Exemplo 1.3 Apuração de dados quantitativos

Para apurar peso ao nascer,[1] o pesquisador deve anotar o número do prontuário e o respectivo peso numa folha de papel. O número do prontuário, escrito ao lado do peso ao nascer, facilita a posterior verificação da apuração.

N° do prontuário	Peso ao nascer
10.525	3,250
10.526	2,010
.	.
.	.
.	.
10.624	2,208

[1] A apuração de peso ao nascer pode ser feita por sexo, se o interesse consistir em comparar peso ao nascer de meninos e de meninas.

Hoje, muitos profissionais registram dados diretamente em computador. Grandes instituições e empresas – como IBGE ou Banco do Brasil – já usam computadores na coleta de dados. São construídas as chamadas *bases de dados*, que armazenam dados de maneira a facilitar a busca de informações. O registro de dados é feito de maneira mais organizada. As bases de dados podem ser manuseadas por meio de planilhas eletrônicas, o que traz maior eficiência às pesquisas. Mas não tenha dúvida: as modernas bases de dados foram construídas a partir de ideias simples, papel e lápis, como aquelas que acabamos de apresentar. De qualquer forma, os dados coletados precisam ser organizados em tabelas.

Exemplo 1.4 Registro de dados

Em uma maternidade, é comum que os dados sobre recém-nascidos e suas mães sejam registrados em computador, não somente para a prestação de cuidados à parturiente e ao nascituro, mas também para que sejam facilmente usados pela administração e, eventualmente, em trabalhos acadêmicos. Nem tudo, porém, já está pronto. Se uma enfermeira quiser estudar o efeito do tabagismo da mãe sobre o peso ao nascer, talvez precise coletar dados sobre tempo do hábito, número de cigarros fumados por dia, se manteve o hábito durante a gestação, em associação com outros hábitos nocivos à saúde, como, por exemplo, alcoolismo.

1.3 Normas para a construção de tabelas

Os dados são apresentados em tabelas colocadas perto do ponto do texto em que são mencionadas pela primeira vez. As tabelas devem conter os seguintes elementos: título, cabeçalho, indicador de linha, células e moldura, como mostrado no Exemplo 1.5.

Exemplo 1.5 Apresentação de dados em tabela

Tabela 1.1

População residente no Brasil, segundo o sexo, de acordo com o Censo Demográfico 2010

Sexo	População residente
Homens	93.406.990
Mulheres	97.348.809
Total	190.755.799

Fonte: Censo Demográfico 2010. IBGE (2011).[2]

[2]Disponível em: <http://www.ibge.gov.br>.

O *título* explica o tipo de dado que a tabela contém. Deve-se colocá-lo acima dos dados. O *cabeçalho* especifica o conteúdo de cada coluna. O *indicador de linha* é um conjunto de termos. Cada termo descreve o conteúdo de uma linha.

Exemplo 1.6 Componentes da tabela

Observe a Tabela 1.1. O título explica a *natureza* (população residente) e a *abrangência* dos dados (Brasil, 2010). O cabeçalho

está destacado em seguida. Na primeira coluna, coloque a denominação da variável, que é sexo, enquanto na segunda coluna está o número (ou frequência) de pessoas de cada sexo residentes no Brasil.

Sexo	População residente

O indicador de linha é mostrado em seguida: a primeira linha apresenta dados sobre homens; a segunda linha, dados sobre mulheres e a terceira linha, o total.

Homens
Mulheres
Total

A *célula* resulta do cruzamento de uma linha com uma coluna e deve conter um dado numérico. Nenhuma célula da tabela deve ficar em branco. Toda célula deve apresentar um número ou, se o dado não existir, coloca-se um traço na célula (–) em que o dado deveria estar escrito.

As tabelas devem ter moldura. Entende-se por *moldura* o conjunto de traços que dão estrutura aos dados numéricos e aos termos necessários à sua compreensão. Então:

- as tabelas devem ser delimitadas, no alto e embaixo, por traços horizontais. Esses traços podem ser mais fortes do que os traços feitos no interior da tabela; as tabelas não devem ser delimitadas, à direita e à esquerda, por traços verticais;
- o cabeçalho deve ser delimitado por traços horizontais;
- é possível fazer traços verticais no interior da tabela, separando as colunas;
- são comuns os traços verticais no interior do cabeçalho, para separar as especificações.

As tabelas ainda podem conter fonte e notas.

A *fonte* identifica o responsável (pessoa física ou jurídica) pelos dados. Deve ser colocada na primeira linha do rodapé da tabela e precedida pela palavra Fonte. Não se indica a fonte nos casos em que os dados foram obtidos pelo pesquisador, ou pelo grupo de pesquisadores, ou pela instituição que apresenta a tabela. Veja o Exemplo.

Exemplo 1.7 Fonte dos dados

Observe a Tabela 1.1. Os dados apresentados nessa tabela são de responsabilidade do Instituto Brasileiro de Geografia e Estatística (IBGE), conforme explica a fonte.

As *notas* são informações de natureza geral que servem para esclarecer o conteúdo das tabelas ou para explicar o método utilizado no levantamento dos dados. São colocadas no rodapé da tabela, logo após a fonte, se houver, e devem ser precedidas pela palavra Nota. Veja o Exemplo 1.8.

Exemplo 1.8 Tabela com fonte e nota

Tabela 1.2

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

Nota: Suprimidos os casos com idade ou local de residência ignorados.

Fonte: Ministério da Saúde/SE/Datasus – Sistema de Informações Hospitalares do SUS (SIH/SUS).

1.3.1 Tabelas de distribuição de frequências para dados qualitativos

Quando observamos *dados qualitativos*, classificamos cada observação em determinada categoria. Depois, contamos o número de observações em cada categoria. A ideia seguinte é resumir as informações na forma de uma tabela que mostre essas contagens (frequências) por categoria. Temos, então, uma *tabela de distribuição de frequências*.

Exemplo 1.9 Tabela de distribuição de frequências para dados ordinais

Pesquisa realizada pelo Datafolha, entre os dias 15 e 16 de julho de 2014, em 233 municípios brasileiros, para saber a opinião das

pessoas (a margem de erro é de 2% para mais ou para menos) sobre o trabalho do técnico Luiz Felipe Scolari na Seleção Brasileira de Futebol em 2014, mostrou que, dos 5.377 entrevistados, 1.075 consideravam o técnico ótimo ou bom, 1.506 julgavam o técnico regular e 2.635 o consideravam péssimo; 161 não tinham opinião ou não quiseram responder. A Tabela 1.3 apresenta as respostas dadas pelos entrevistados (primeira coluna) e as respectivas frequências dessas respostas (segunda coluna).

Tabela 1.3

Opinião dos brasileiros sobre o técnico de futebol

Resposta	Frequência
Ótimo ou bom	1.075
Regular	1.506
Péssimo	2.635
Não sabe/ não respondeu	161
Total	5.377

Fonte: dimassantos.com.br/pesquisa-aponta-tite-para-futuro-tecnico-da-selecao. Acesso em: Setembro de 2014.

As tabelas de distribuição de frequências podem apresentar, além das frequências, a *proporção* (frequência relativa) de unidades que recaem em cada categoria. Para obter a *proporção* (frequência relativa) de unidades de determinada categoria, calcule:

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

As frequências relativas são, muitas vezes, expressas em percentuais, porque as pessoas entendem mais facilmente proporções dadas em porcentagens. Para obter o percentual de determinada categoria, multiplique a frequência relativa por 100.

Convém exibir sempre o total (tamanho da amostra), que é o indicador da credibilidade da informação.[3]

Exemplo 1.10 Tabela de distribuição de frequências, com frequências relativas

A Tabela 1.4 apresenta, na terceira coluna, as frequências relativas para os dados contidos na Tabela 1.3.

Tabela 1.4

Opinião dos brasileiros sobre o técnico de futebol

Resposta	Frequência	Frequência relativa
Ótimo ou bom	1.075	
Regular	1.506	
Ruim	2.635	
Não sabe	161	
Total	5.377	1,00

Fonte: dimassantos.com.br/pesquisa-aponta-tite-para-futuro-tecnico-da-selecao. Acesso em setembro de 2014.

1.3.2 Tabelas de contingência

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

Exemplo 1.11 Tabela de contingência

Diabetes mellitus durante a gravidez aumenta o risco de complicações perinatais. Para comparar a redução obtida sob novo tratamento com a redução obtida sob tratamento de rotina em diferentes raças ou grupos étnicos, foi conduzido um ensaio

clínico randomizado.[4] Os dados sobre raça e etnia das voluntárias, bem como o grupo de estudo ao qual foram designadas, estão na Tabela 1.5.

Tabela 1.5

Raça ou etnia das voluntárias segundo o grupo

Raça ou grupo étnico	Grupo	
	Intervenção	Rotina
Branca	356	396
Asiática	92	72
Outro	42	42

Fonte: Crowther, CA *et alii*. Effect of Treatment of Gestational Diabetes Mellitus on Pregnancy Outcomes. N Engl J Med 2005; 352:2.477-2.486, June 16, 2005.

[4]Ver Vieira, S. e Hossne, WS. Metodologia científica para a área de saúde. 2 ed. Rio de Janeiro: Elsevier, 2015.

As tabelas de contingência devem apresentar os totais porque não é possível confiar nos resultados obtidos de amostras muito pequenas. Também podem expor percentuais.[5]

Exemplo 1.12 Tabela de contingência com totais

A Tabela 1.6 reapresenta a Tabela 1.5, agora com os totais. Fica fácil ver que havia mais brancas no estudo e mais voluntárias no tratamento de rotina.

Tabela 1.6

Raça ou etnia das voluntárias segundo o grupo

Raça ou grupo étnico	Grupo		Total
	Intervenção	Rotina	
Branca	356	396	752
Asiática	92	72	164
Outro	42	42	84
Total	490	510	1.000

Fonte: Crowther, CA *et alii*. Effect of Treatment of Gestational Diabetes Mellitus on Pregnancy Outcomes. N Engl J Med 2005; 352: 2.477-2.486, June 16, 2005.

1.3.3 Apresentação de dados quantitativos

Os dados quantitativos são apresentados na ordem em que foram coletados. Os pesquisadores podem identificar a unidade que forneceu o dado por um número. No caso de pesquisas em seres humanos, alguns pesquisadores identificam os participantes pelas iniciais de seus nomes e apresentam os dados obedecendo à ordem alfabética das iniciais.

Exemplo 1.13 Apresentação de dados quantitativos

Foram coletados dados de 48 pacientes que participaram de uma pesquisa. A Tabela 1.7 apresenta os dados de seis deles.

Tabela 1.7

Idade, peso, altura, pressão arterial sistólica, pressão arterial diastólica em seis pacientes

Paciente	Idade (anos)	Peso (kg)	Altura (metros)	PAS mmHg	PAD mmHg
AG	56,3	85,5	1,55	191	79
AAS	50,5	72,25	1,58	152	92
ABS	64,1	68	1,65	204	113
ACPS	38,7	96	1,69	169	86
ACS	58,6	64,1	1,46	145	84
CVB	44,1	80,1	1,7	170	95

Nota: Não são apresentados todos os dados porque isso tornaria a tabela muito extensa, e a finalidade, aqui, é mostrar como se faz uma tabela.

Fonte: Sousa, MG. Determinantes das propriedades funcionais e estruturais das grandes artérias e as relações com lesão de órgãos-alvo em hipertensos estágio 3. Tese (doutorado). Faculdade de Medicina da USP. 2012.

1.3.4 Tabelas de distribuição de frequências para dados quantitativos

Dados quantitativos podem ser apresentados em *tabelas de distribuição de frequências,* como mostrado no Exemplo 1.14. Se os dados são *discretos,* para organizar a tabela de distribuição de frequências:

- escreva os dados em ordem crescente;
- conte quantas vezes cada valor se repete;
- organize a tabela apresentando os valores numéricos em ordem natural.

Exemplo 1.14 Tabela de distribuição de frequências para dados discretos

É mais fácil entender os dados da Tabela 1.8 se forem apresentados como mostra a Tabela 1.9.

Tabela 1.8

Número de faltas de trinta funcionários ao trabalho. Clínica ABC, segundo semestre de 2014

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

Tabela 1.9

Número de faltas de trinta funcionários ao trabalho. Clínica ABC, segundo semestre de 2014

N°. de faltas	Frequência	Porcentagem
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 contínuos* não dão ao leitor visão rápida e global do fenômeno. É difícil dizer como os valores se distribuem. Por essa razão, dados contínuos – desde que em grande número – são apresentados em *tabelas de distribuição de frequências*. Mas veja os dados apresentados no Exemplo 1.15.

Exemplo 1.15 Apresentação de dados contínuos

Os dados apresentados na Tabela 1.10 não dão visão rápida sobre peso ao nascer.

Tabela 1.10

Peso ao nascer, em quilogramas, de nascidos vivos

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 frequências 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 1.16, as faixas são de peso). Essas faixas recebem o nome de *classes*;
- o resultado da divisão é o *intervalo de classe*. Sempre é melhor arredondar o valor obtido para o intervalo de classes para um valor mais alto, o que facilita o trabalho;
- organize as classes, de maneira que a primeira contenha o menor valor observado.

Exemplo 1.16 Construção de tabela de distribuição de frequências (dados contínuos)

Observe os dados apresentados na Tabela 1.10. O menor valor é 1,570 kg, e o maior valor, 4,600 kg. A amplitude dos dados é:

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

Para organizar *sete* classes, calcule:

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

Arredonde o valor calculado para intervalo de classe, que resultou em 0,433, para 0,500 e construa a primeira classe, que será de 1,5 kg a 2,0 kg (essa classe contém o menor valor); em seguida, construa a segunda classe, que será de 2,0 kg a 2,5 kg, e assim por diante, como mostra o esquema a seguir:

$$1{,}5 \vdash 2{,}0$$

$$2{,}0 \vdash 2{,}5$$

$$2{,}5 \vdash 3{,}0$$

$$3{,}0 \vdash 3{,}5$$

$$3{,}5 \vdash 4{,}0$$

$$4{,}0 \vdash 4{,}5$$

$$4{,}5 \vdash 5{,}0$$

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 frequências, se os valores iguais aos extremos estão ou não incluídos na classe. Veja a notação usada no Exemplo 1.16. A primeira classe é

$$1{,}5 \vdash 2{,}0$$

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

Exemplo 1.17 Tabela de distribuição de frequências para dados contínuos

Os dados de peso ao nascer de nascidos vivos foram organizados em uma tabela de distribuição de frequências. Veja a Tabela 1.11.

Tabela 1.11

Distribuição de frequências para peso ao nascer de nascidos vivos, em quilogramas

Classe	Frequência
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 neste momento que, para indicar se extremos de classe estão ou não incluídos em determinada classe, é possível adotar outros métodos. 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 aqueles que estão na véspera de completar 5 anos.

O número de classes deve ser escolhido pelo pesquisador, em função do que pretende mostrar. Em geral, convém estabelecer de 5 a 20 classes. Se o número de classes for demasiadamente 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 1.18 Cálculo do número de classes

Reveja a Tabela 1.10. Com $n = 100$, aplicando a primeira fórmula, 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 na Tabela 1.11, aplicou-se a segunda fórmula e, por isso, foram construídas *sete* classes.

Às vezes, as classes de uma distribuição de frequências já estão definidas por tabelas que informam, por exemplo, os intervalos de normalidade. Essa situação é comum nas ciências biológicas. Nesses casos, a distribuição de frequências deve obedecer às definições dos especialistas.

Exemplo 1.19 Tabela de distribuição de frequências para dados contínuos com classes de tamanhos definidos por especialistas

É difícil dizer, observando os dados apresentados na Tabela 1.12, o número de obesos, por exemplo. Fica mais fácil observar os dados mostrados na Tabela 1.13.

Tabela 1.12

IMC de hipertensos estágio 3, com idade média de 53,6 anos

35,6	25,2	43,3	30,1	33,4	24,8	29,1	41,3
28,9	36,6	26,2	31,3	30,5	28,7	29,3	28,7
25,0	33,9	30,1	27,6	25,7	34,7	32,7	24,4
33,6	32,7	38,4	30,6	29,3	30,4	18,9	35,3
30,1	30,5	26,1	29,4	28,4	29,8	21,8	39,5
27,7	25,2	35,6	23,5	36,8	28,7	28,7	26,6

Fonte: Sousa, MG. Determinantes das propriedades funcionais e estruturais das grandes artérias e as relações com lesão de órgãos-alvo em hipertensos estágio 3. Tese (doutorado). Faculdade de Medicina da USP. 2012.

Tabela 1.13

Distribuição dos pacientes hipertensos classificados segundo o IMC

IMC	Frequência	Porcentagem
Abaixo do peso	1	2,1
Normal	4	8,3
Acima do peso	20	41,7
Obesidade I	14	29,2
Obesidade II	7	14,6
Obesidade III	2	4,2
Total	48	100,0

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

Exemplo 1.20 Tabela de distribuição de frequências para dados contínuos com classes de tamanhos diferentes e extremo superior da última classe não definido

Para dar uma ideia geral sobre pressão sanguínea sistólica de mulheres com 30 anos, um pesquisador apresentou não os valores observados, mas o número de mulheres por faixas de pressão. Veja a Tabela 1.14, que também é um exemplo no qual o extremo superior da última classe não está definido.

Tabela 1.14

Distribuição de frequências para pressão sanguínea sistólica, em milímetros de mercúrio, de mulheres com 30 anos

Classe	Frequência
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 frequências mostram a distribuição da variável, *mas perdem em exatidão*. Por exemplo, a Tabela 1.14 revela que seis mulheres apresentaram pressão sanguínea sistólica entre 90 e 100, mas não dá o valor exato para cada uma delas.

1.4 Exercícios resolvidos

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

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

1.4.2. 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.

1.4.3. Para estudar a distribuição dos erros cometidos por alunos nas radiografias intrabucais, foram obtidos os dados que estão na Tabela 1.15. As frequências relativas e o total estão apresentados na Tabela 1.16.

Tabela 1.15

Erros técnicos em radiografias intrabucais

Erros	Frequência
Ângulo horizontal	459
Exposição insuficiente	355
Resultado amarelado	158
Excesso de exposição	141
Corte do dente	130
Resultado manchado	63
Corte cônico	44
Outros erros	46

Fonte: Carvalho, PL *et al*. Erros técnicos nas radiografias intrabucais realizadas por alunos de graduação. RGO, Porto Alegre, v. 57, n.2, p. 151-155, abr./jun. 2009.

Tabela 1.16

Erros técnicos em radiografias intrabucais

Erros	Frequência	Porcentagem
Ângulo horizontal	459	32,9
Exposição insuficiente	355	25,4
Resultado amarelado	158	11,3
Excesso de exposição	141	10,1
Corte do dente	130	9,3
Resultado amarelado	63	4,5
Corte cônico	44	3,2
Outros erros	46	3,3
Total	1.396	100,0

1.4.4. De acordo com o Sistema Nacional de Informações Tóxico-Farmacológicas (Sinitox), em 2005 foram registrados, no Brasil, 23.647 casos de intoxicação humana 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. Esses dados estão apresentados na Tabela 1.17.

Tabela 1.17

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

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).[6]

1.4.5. Construa uma tabela de distribuição de frequências para apresentar os dados da Tabela 1.18.

Tabela 1.18

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

O número k de classes para apresentar $n = 49$ dados pode ser obtido pela segutinte fórmula:

$$k = \sqrt{n} = \sqrt{49} = 7$$

Podem ser constituídas 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. Veja a Tabela 1.19.

Tabela 1.19

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

1.4.6. Imagine[7] que você quer comparar as distribuições de frequências da mesma variável, para homens e mulheres, separadamente, mas o número de mulheres é consideravelmente maior. Você compararia as frequências ou as frequências relativas? Por quê? Dê um exemplo.

Devem-se comparar, em cada categoria, as proporções obtidas para homens e para mulheres. As frequências não são comparáveis, uma vez que as amostras são de tamanhos diferentes. Para entender essa informação, imagine que são, no total, 200 mulheres e 50 homens e que, para uma dada categoria, a frequência seja de 4, em ambas as distribuições. Isso significa 2% das mulheres (4 /200 = 0,02) e 8% dos homens (4/50 = 0,08), uma diferença muito grande.

---

[6] http://www.saude.rj.gov.br/animaispeconhentos/estatisticas.html. Disponível em 30 de maio de 2008.

[7] Minium, E. W., Clarke, R. C., Coladarci, T. *Elements of Statistical Reasoning*. 2 ed. New York, Wiley, 1999, p. 33.

1.5 Exercícios propostos

1.5.1. Especifique o tipo (qualitativa, quantitativa, nominal etc.) 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.

1.5.2. 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 se mostravam contrárias.

1.5.3. Complete a Tabela 1.20.

Tabela 1.20

Distribuição das notas de 200 alunos

Nota do aluno	Frequência	Frequência relativa
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

1.5.4. Uma doença pode ser classificada em três estágios (leve; moderada; severa). Foram examinados vinte pacientes, obtendo-se os seguintes 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 frequência de cada categoria; b) calcule a frequência relativa de cada categoria.

1.5.5. Qual é o erro na distribuição de frequências dada em seguida?

Classe
20 – 30
30 – 40
40 – 50
60 – 70
70 e mais

1.5.6. São dados os tipos de sangue de quarenta 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; A; O; O; O; A; O; A; O; O; A; O; AB; O; O; A; AB; B; B. Apresente os dados em uma tabela de distribuição de frequências.

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

1.5.8. Foram avaliadas, por cirurgiões dentistas com especialização em Ortodontia, crianças no estágio de dentadura decídua, entre 3 e 6 anos de idade. Dessas crianças, 615 não tinham hábitos de sucção, 190 tinham o hábito de sucção do polegar, 588 usavam chupeta e 618 usavam mamadeira. Apresente os dados em tabela. Calcule o total e as frequências relativas.

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

1.5.10. O número de enfermeiros em serviço varia muito em um hospital. Foi feita uma distribuição de frequências com as seguintes classes: $20 \vdash 35$; $35 \vdash 40$; $40 \vdash 45$; $45 \vdash 50$; $50 \vdash 55$. Qual é

o intervalo de classes e qual é o intervalo de toda a distribuição de frequências?

1.5.11. Construa uma tabela de distribuição de frequências para apresentar os dados da Tabela 1.21, usando intervalos de classes iguais. Em seguida, 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 1.21

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

1.5.12. Imagine dois conjuntos de dados, A e B; no primeiro conjunto, n = 50, e, no segundo, n = 100. No conjunto A, o valor mínimo é 24 e o valor máximo, 70; no conjunto B, o valor mínimo é 187 e o valor máximo, 821. Construa intervalos de classe para cada conjunto.

1.5.13. Com base nos dados apresentados na Tabela 1.22, calcule o percentual de pacientes que abandonaram o tratamento contra tuberculose pulmonar (taxa de abandono), segundo a zona de moradia.

Tabela 1.22

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

Zona	Abandono do tratamento	
	Sim	Não
Urbana	15	80
Rural	70	35

1.5.14. Perguntou-se, a cem 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 esses dados em uma tabela de distribuição de frequências e discuta os resultados. Os dados mostram que os dentistas adotam a prática de prevenção?

1.5.15. Calcule as frequências relativas para os dados apresentados na Tabela 1.23 e comente.

Tabela 1.23

Número de óbitos por grupos de causa. Brasil, 2004

Grupos de causa	Número	
	Masculino	Feminino
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

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.

Fonte: Ministério da Saúde/SVS – Sistema de Informações sobre Mortalidade (SIM)[8]

1.5.16. Calcule as frequências relativas para os dados apresentados na Tabela 1.24 e aponte a faixa etária de maior risco.

Tabela 1.24

Pacientes portadores de carcinoma epidermoide 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

1.5.17. Com base nos dados apresentados na Tabela 1.25, calcule o percentual de órgãos aproveitados (taxa de aproveitamento para cada órgão).

Tabela 1.25
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

[8]Disponível em http://tabnet.datasus.gov.br/CGI/tabcgi.exe?idb2006/c04.def. Acesso em: 4 mai. 2008.

[3]Não tem sentido fornecer resultados em porcentagens quando a amostra é muito pequena. Por exemplo, não teria sentido fornecer porcentagens se a amostra fosse constituída por cinco ou seis pessoas.

[5]Ver o Capítulo 12 deste livro.

CAPÍTULO 2

Apresentação de Dados em Gráficos

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.

2.1 Apresentação de dados qualitativos

2.1.1 Gráfico de barras

O *gráfico de barras*[2] é usado para apresentar variáveis *qualitativas*, sejam elas nominais ou ordinais. Para construir um *gráfico de barras*:

- desenhe o sistema de eixos cartesianos;
- anote as categorias da variável estudada no eixo das abscissas (eixo horizontal);
- escreva as frequências ou as frequências 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 frequência ou pela frequência relativa (em geral, em porcentagem) da categoria;
- coloque legendas nos dois eixos e título na figura.

Exemplo 2.1 Gráfico de barras

Foram entrevistadas cem pessoas que haviam sido submetidas a uma cirurgia estética reparadora. Indagadas se consideravam que a cirurgia havia melhorado a aparência delas, responderam como segue: 66 afirmaram que sim, 20 disseram que em parte, 8 disseram que não e 6 não quiseram responder. Os dados estão na Tabela 2.1 e o gráfico de barras está apresentado na Figura 2.1.

Tabela 2.1

Você acha que a cirurgia melhorou sua aparência?

Resposta	Frequência	Porcentagem
Sim	66	66
Em parte	20	20
Não	8	8
Sem resposta	6	6
Total	100	100

FIGURA 2.1 Você acha que a cirurgia melhorou sua aparência?

Para facilitar a leitura dos percentuais de cada categoria, é possível fazer linhas auxiliares (linhas de grade).

Exemplo 2.2 Gráfico de barras com grades

Com os dados da Tabela 2.1, foi desenhado um gráfico de barras com linhas auxiliares, apresentado na Figura 2.2.

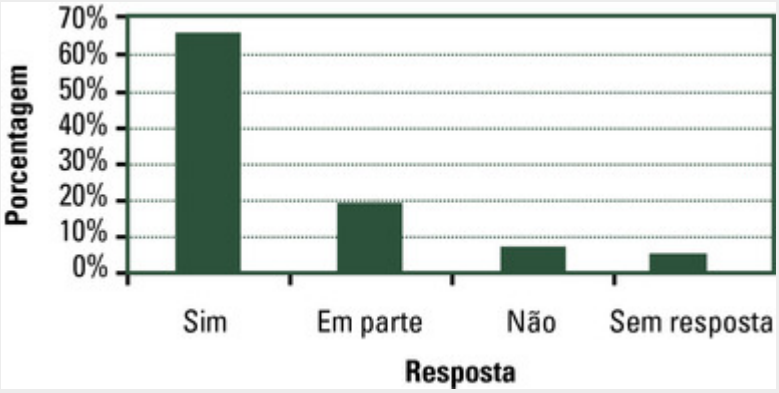


FIGURA 2.2 Você acha que a cirurgia melhorou sua aparência?

Os percentuais podem ser apresentados nas barras (rótulos dos dados), em diversas posições.

Exemplo 2.3 Gráfico de barras com percentuais nas barras

Com os dados da Tabela 2.1, foi desenhado o gráfico de barras da Figura 2.3, com percentuais escritos acima das barras.

FIGURA 2.3 Você acha que a cirurgia melhorou sua aparência?

Os gráficos de barras podem ser feitos em três dimensões. São, então, conhecidos como gráficos em 3D. São agradáveis de ver, mas de difícil compreensão quando apresentam muitas categorias.

Exemplo 2.4 Gráfico de barras com 3 D

Com os dados da Tabela 2.1, foi feito o gráfico de barras em três dimensões apresentado na Figura 2.4.

FIGURA 2.4 Você acha que a cirurgia melhorou sua aparência?

Quando o *gráfico de barras* é usado para apresentar variáveis ordinais, deve-se obedecer à ordem das categorias da variável, mas devem ser colocadas, no final, as categorias "não sabe", "não respondeu" etc.

Exemplo 2.5 Gráfico de barras para dados ordinais

Veja os dados apresentados na Tabela 1.3 do Capítulo 1. A ordem das categorias foi respeitada e é mostrado o número de respondentes em cada categoria.

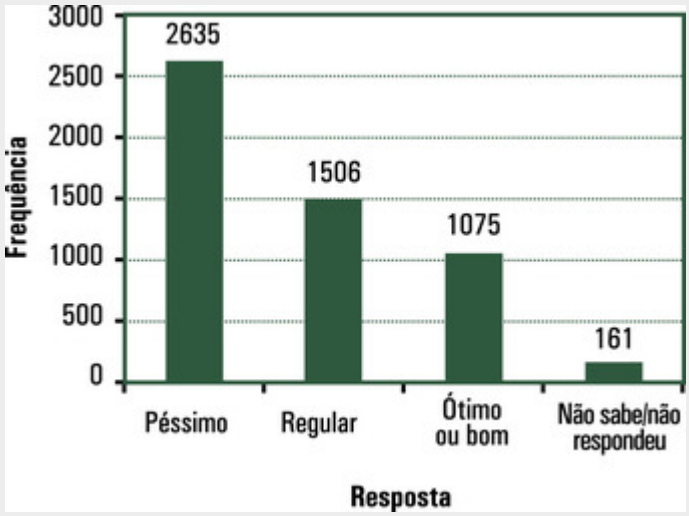


FIGURA 2.5 Opinião dos brasileiros sobre o técnico de futebol

As barras do gráfico podem ser apresentadas na posição horizontal, como mostra o Exemplo 2.6.

Exemplo 2.6 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 realizadas em um Centro de Radiologia, estão na Tabela 2.2. O gráfico de barras, com as *barras em posição horizontal*, está apresentado na Figura 2.6.

Tabela 2.2

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

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

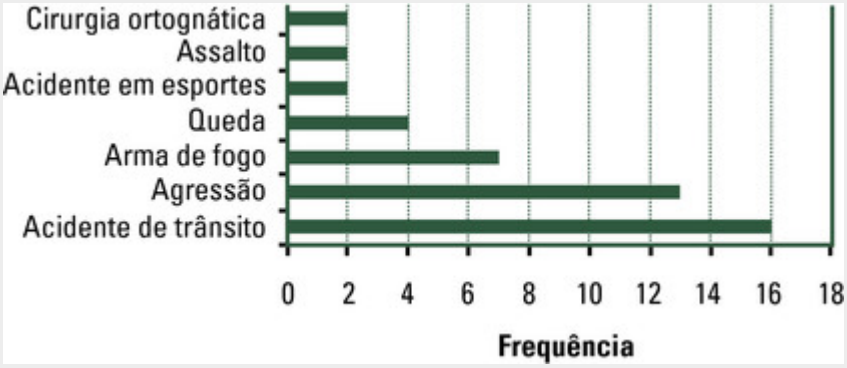


FIGURA 2.6 Pacientes quanto à etiologia da fratura ou à presença de corpo estranho diagnosticada por radiografia panorâmica

Aqui, cabe esclarecer que o programa Excel denomina gráfico de barras somente aqueles que apresentam as barras na posição horizontal. Gráficos com barras verticais são denominados, no Excel, de *gráfico de colunas*. No entanto, o termo técnico, em ambos os casos, é gráfico de barras. Cabe também considerar que gráficos com barras na posição vertical (colunas) são mais comuns, porém gráficos com barras na posição horizontal facilitam a leitura dos nomes das categorias. São, portanto, preferíveis quando os nomes são extensos.

2.1.2 Gráfico de setores

O gráfico de setores[3] 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 360º). Essa circunferência representará o total, ou seja, 100%;
- divida a circunferência em tantos setores quantas sejam as categorias da variável em estudo, mas é preciso calcular o ângulo de cada setor: 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 2.7 Gráfico de setores

Por meio de radiografias panorâmicas, foram constatados fraturas e corpos estranhos na face de 46 pacientes, 29 homens e 17 mulheres. Os dados estão apresentados na Tabela 2.3 e o gráfico de setores, na Figura 2.7.

Tabela 2.3

Pacientes com fraturas e corpos estranhos na face segundo o sexo

Sexo	Frequência	Proporção
Homens	29	0,63
Mulheres	17	0,37
Total	46	1,00

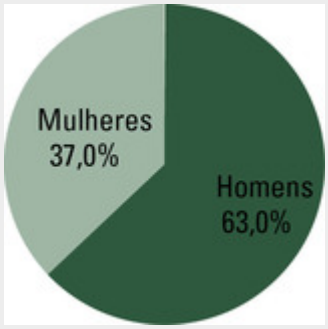


FIGURA 2.7 Pacientes com fraturas e corpos estranhos na face segundo o sexo

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:

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

A fim de destacar melhor a contribuição de cada valor em relação ao total, as "fatias da pizza" podem ser separadas como mostra a Figura 2.15 (na Seção 2.3 deste capítulo). Além disso, os gráficos de setores podem ser feitos em três dimensões, como mostra a Figura 2.8. Esse tipo de apresentação aparece em muitas revistas, mas deve ser evitado porque dificulta a avaliação da proporção de cada categoria.

Exemplo 2.8 Gráfico de setores em 3D

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

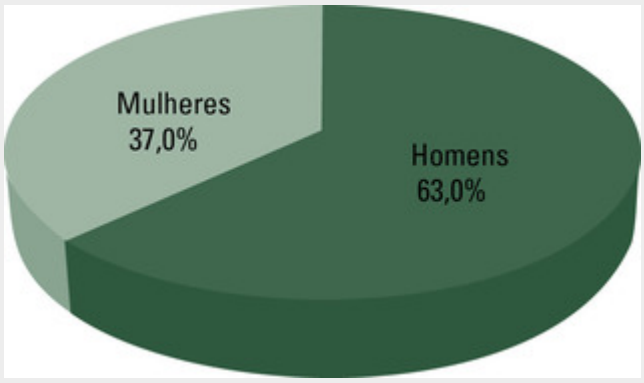


FIGURA 2.8 Pacientes com fraturas e corpos estranhos na face segundo o sexo

Você encontra, no programa Excel, várias opções para o desenho do gráfico de setores. Todas estão corretas, a escolha é sua, mas as opções mais simples são as de mais fácil entendimento por seu leitor.

2.1.2.1 Uma variação do gráfico de setores

O programa Excel apresenta uma variação do gráfico de setores, que denomina de gráfico de rosca. Para desenhar esse gráfico, faça primeiro o gráfico de setores. Em seguida, faça uma circunferência com o mesmo centro do gráfico de setores, mas bem menor. Deixe essa circunferência em branco.

Exemplo 2.9 Gráfico de setores (rosca)

Com os dados da Tabela 2.3, foi desenhado o gráfico da Figura 2.9.

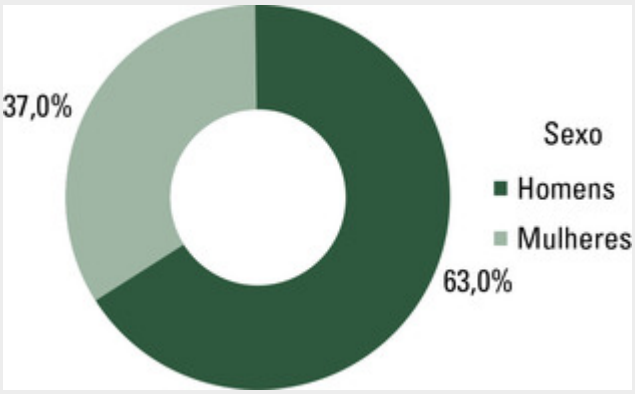


FIGURA 2.9 Pacientes com fraturas e corpos estranhos na face segundo o sexo

2.2 Apresentação de dados quantitativos

2.2.1 Diagrama de linhas

Para apresentar graficamente dados discretos organizados em uma *tabela de distribuição de frequências*, pode-se construir um *diagrama de linhas*, da seguinte forma:

- escreva os valores assumidos pela variável no eixo das abscissas (eixo horizontal);
- escreva as frequências ou as frequências 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 frequências ou pelas frequências relativas (em geral, em porcentagem);
- coloque legendas nos dois eixos e título na figura.

Exemplo 2.10 Diagrama de linhas

A Tabela 1.9 apresenta a distribuição de frequências para o número de faltas dos funcionários da Clínica ABC, no segundo semestre de 2014, ao trabalho. O diagrama de linhas está na Figura 2.10.

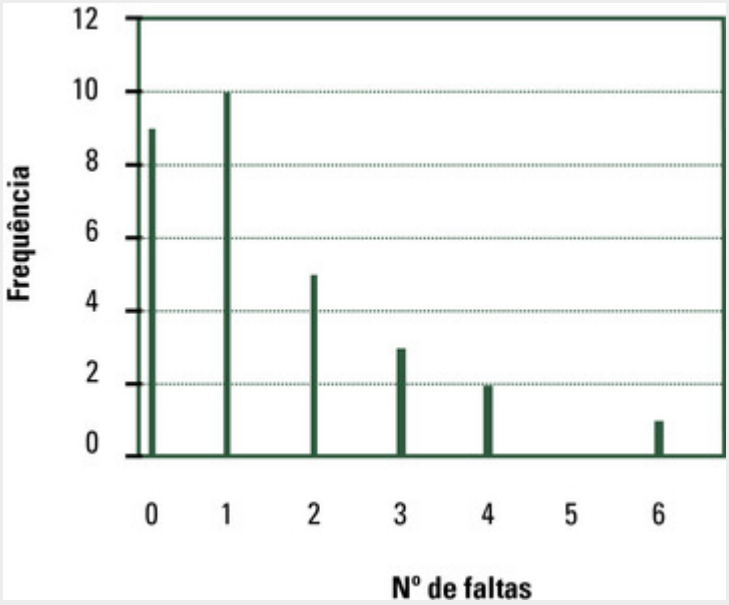


FIGURA 2.10 Diagrama de linhas para o número de faltas dos funcionários da Clínica ABC, no segundo semestre de 2014, ao trabalho

2.2.2 Gráfico de pontos

Os dados contínuos – ao contrário dos discretos – são, na maioria das vezes, diferentes uns dos outros. Veja o Exemplo 2.11: os valores são todos diferentes entre si. Dados contínuos em pequeno número 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 2.11 Gráfico de pontos

O tempo de sobrevivência de sete pacientes submetidos a transplante renal em determinado hospital foi, em dias, de: 17, 5, 48, 120, 651, 64, 150. Para apresentar esses dados em 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, você põe os pontos que vão representar os dados sobre ela, sempre obedecendo à escala, como mostra a Figura 2.11.

FIGURA 2.11 Tempo de sobrevivência, em dias, após transplante renal

2.2.3 Histograma

Quando os dados contínuos são em grande número, não se pode fazer um gráfico de pontos. É mais conveniente organizar os dados em uma tabela de distribuição de frequências,[4] como mostrado no Capítulo 1, e desenhar um *histograma*. Para construir um histograma:

- trace, primeiro, o sistema de eixos cartesianos;
- 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 frequências (ou às frequências relativas) das respectivas classes. As barras devem ser justapostas, a fim de evidenciar a natureza contínua da variável;
- coloque legendas nos dois eixos e título na figura.

Exemplo 2.12 Histograma

Os dados apresentados na Tabela 1.11 do Capítulo 1 estão no histograma da Figura 2.12.

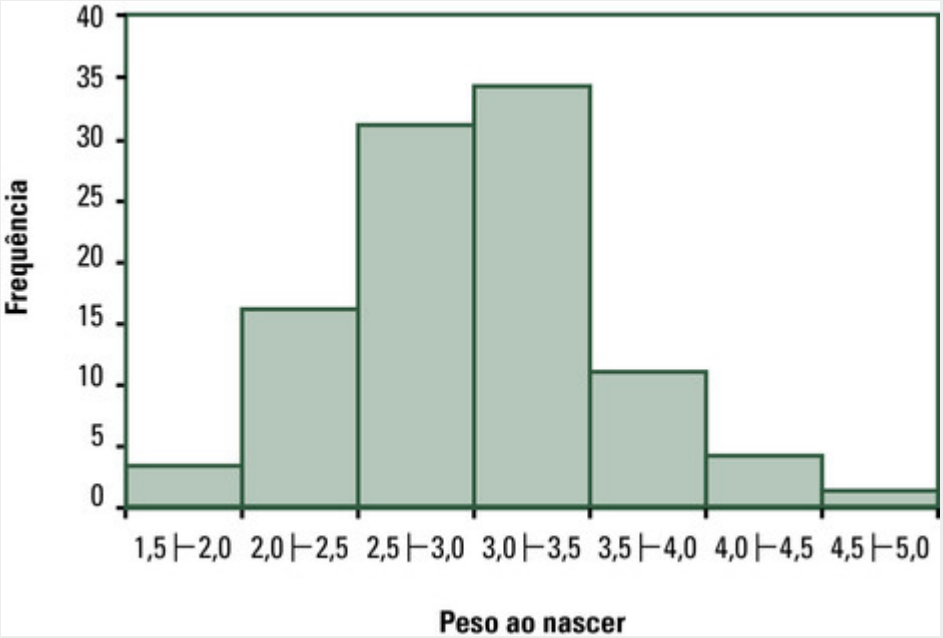


FIGURA 2.12 Histograma para peso ao nascer, em quilogramas, de nascidos vivos

2.2.4 Polígono de frequências

Dados contínuos apresentados em uma tabela de distribuição de frequências também podem ser apresentados em *polígonos de frequências*. Para fazer esse tipo de gráfico:

- trace o sistema de eixos cartesianos;
- marque, no eixo das abscissas, pontos exatamente no meio dos extremos de classe;
- marque, no eixo das ordenadas, as frequências de classe;
- una os pontos por segmentos de reta;
- feche o polígono unindo os extremos da figura com o eixo horizontal;
- coloque legendas nos dois eixos e título na figura.

Exemplo 2.13 Polígono de frequências

O polígono de frequências da Figura 2.13 apresenta os dados da Tabela 1.11 do Capítulo 1.

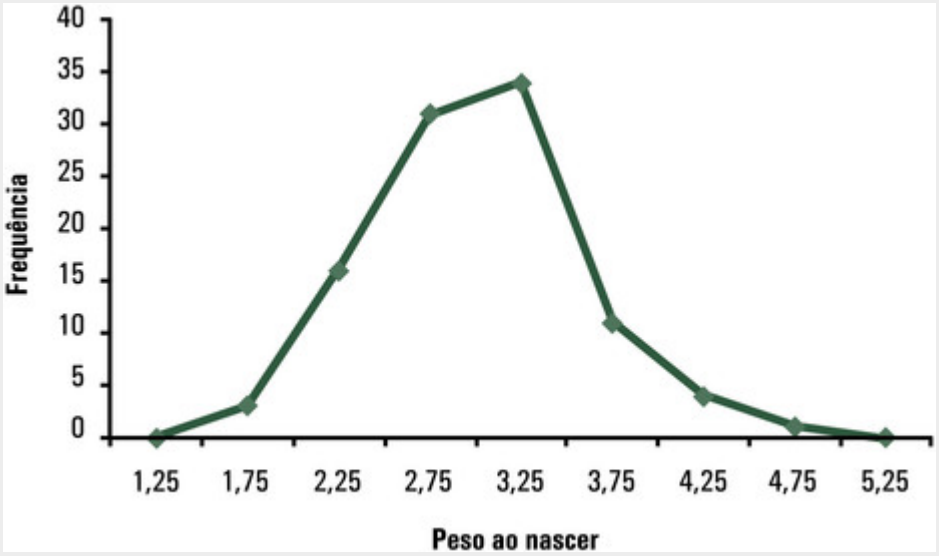


FIGURA 2.13 Polígono de frequências para peso ao nascer de nascidos vivos, em quilogramas

2.3 Exercícios resolvidos

2.3.1. Faça um gráfico de barras e um gráfico de setores para apresentar os dados da Tabela 1.17 do Capítulo 1.

O gráfico de barras está na Figura 2.14 e o gráfico de setores está na Figura 2.15.

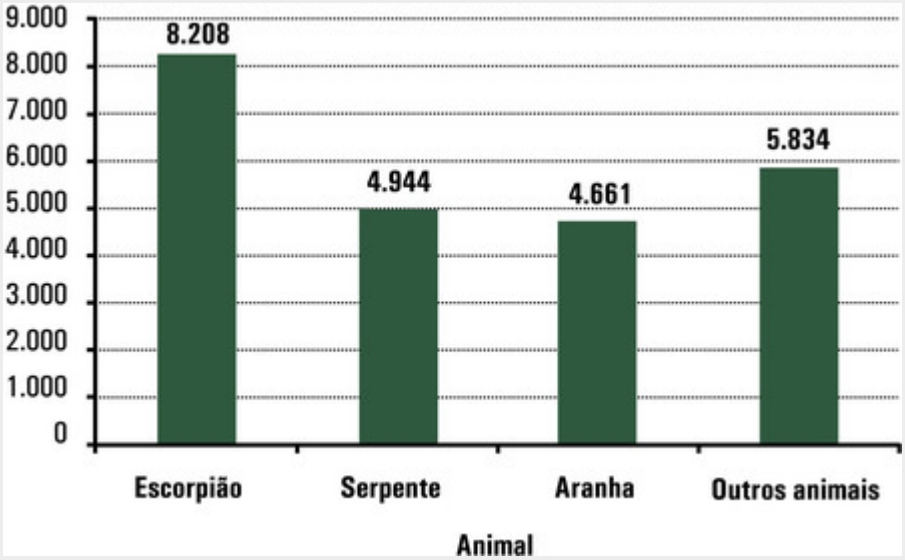


FIGURA 2.14 Casos de intoxicação humana por animal peçonhento, ocorridos no Brasil em 2005, segundo o animal

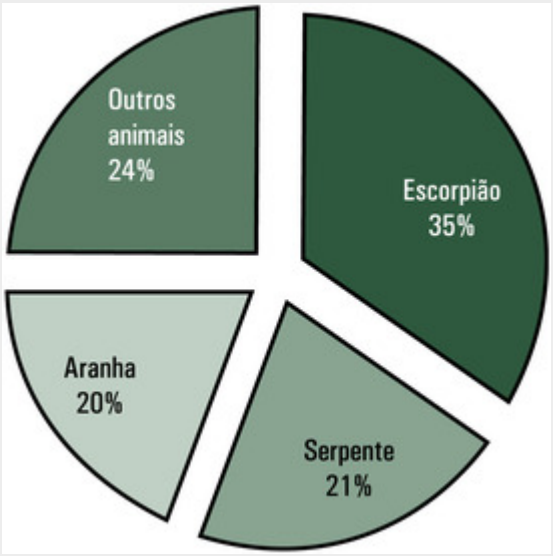


FIGURA 2.15 Casos de intoxicação humana por animal peçonhento, ocorridos no Brasil em 2005, segundo o animal

2.3.2. Faça um polígono de frequências para apresentar os dados da Tabela 1.19 (Cap. 1).

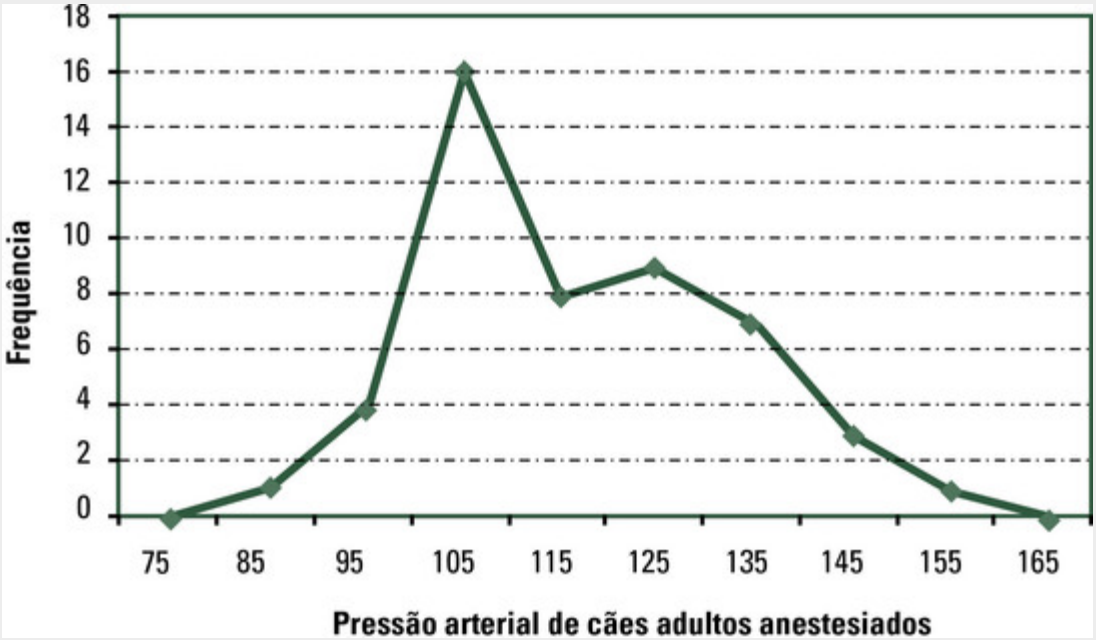


FIGURA 2.16 Pressão arterial, em milímetros de mercúrio, de cães adultos anestesiados

2.3.3. Por que uma pessoa que conhece determinado assunto preferiria olhar uma tabela de distribuição de frequências a olhar um gráfico? Qual seria um argumento razoável contra essa postura?

Como é possível construir gráficos muito diferentes com base nos mesmos dados, a interpretação, com base apenas neles, às vezes não é confiável. Por outro lado, é a apresentação gráfica que ressalta determinadas características dos dados. Em geral, é melhor observar tanto os dados como o gráfico.

2.3.4. Quando um gráfico deve ser grande? Ou 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.

2.4 Exercícios propostos

2.4.1. Uma doença pode ser classificada em três estágios (leve; moderada; severa). Foram examinados vinte pacientes e obtidos os seguintes dados: moderado, leve, leve, severo, leve, moderado, moderado, moderado, leve, leve, severo, leve, moderado, moderado, leve, severo, moderado, moderado, moderado, leve. Com base nesses dados, desenhe um gráfico de setores para apresentar a distribuição de frequências que você já construiu, conforme pedido no Exercício 1.5.4 (Cap. 1).

2.4.2. São dados os tipos de sangue de quarenta 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; A; O; 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 frequências. Desenhe um gráfico de barras para apresentar a distribuição de frequências, que você já construiu conforme pedido no Exercício 1.5.6 (Cap. 1).

2.4.3. Foram avaliadas, por cirurgiões-dentistas com especialização em Ortodontia, crianças no estágio de dentadura decídua, na faixa etária de 3 a 6 anos. Dessas crianças, 615 não tinham hábitos de sucção, 190 tinham o hábito de sucção do polegar, 588 usavam chupeta e 618 usavam mamadeira. Apresente os dados em tabela. Desenhe um gráfico de barras horizontais para apresentar a distribuição de frequências que você construiu conforme pedido no Exercício 1.5.8 (Cap. 1).

2.4.4. Desenhe um histograma para apresentar a distribuição de frequências que você já construiu usando intervalos de classes iguais, conforme pedido no Exercício 1.5.11.

2.4.5. Com base nos dados apresentados no Exercício 1.5.13 (Cap. 1), você construiu uma distribuição de frequências. Desenhe dois gráficos de setores (um para cada zona de moradia) para apresentar essa distribuição.

2.4.6. Você calculou as frequências relativas para o número de óbitos por grupos de causa, Brasil, 2004, no Exercício 1.5.15 (Cap. 1). Agora, faça um gráfico de barras (as barras na posição horizontal) para apresentar os percentuais, por sexo.

2.4.7. No Exercício 1.5.15 (Cap. 1), você calculou as frequências relativas. Agora, desenhe um histograma para apresentar essa distribuição de frequências.

2.4.8. Você já calculou o percentual de órgãos aproveitados (taxa de aproveitamento para cada órgão), usando os dados do exercício do Capítulo 1. Agora, desenhe um gráfico de barras (as barras na posição horizontal) para apresentar a taxa de aproveitamento de cada órgão.

2.4.9. Com base nos dados apresentados na Tabela 2.4, faça uma tabela de distribuição de frequências. Desenhe um histograma.

Tabela 2.4

Pressão sanguínea diastólica de 35 enfermeiros que trabalham em um hospital

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

2.4.10. Com os dados apresentados na Tabela 2.4, você construiu uma tabela de distribuição de frequências. Agora, desenhe um polígono de frequências.

---

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

[2]No programa Excel, o gráfico de barras verticais é chamado gráfico de colunas. No entanto, o nome técnico é gráfico de barras.

[3]O gráfico de setores é mais conhecido como gráfico de pizza. Este, contudo, não é o nome técnico.

[4]Se os intervalos de classe forem diferentes, não se pode fazer o histograma como ensinado aqui. Consulte textos mais avançados.

CAPÍTULO 3

Medidas de Tendência Central

Para entender as características gerais de um conjunto de dados, muitas pessoas preferem olhar uma figura.[1] Daí a importância dos métodos gráficos descritos no Capítulo 2. No caso das *variáveis quantitativas ou numéricas* – mais usadas na pesquisa científica, por serem mais exatas –, os gráficos são, porém, menos informativos, porque, para desenhar um histograma ou um polígono de frequências para uma grande quantidade de dados, é preciso agrupar valores exatos em classes.

Mas já foram propostas, há muito tempo, *medidas estatísticas* que sumarizam as informações contidas em um grande conjunto de dados. Essas medidas apontam características específicas do conjunto de dados e permitem, a quem conhece suas propriedades e limitações, uma visão geral do comportamento dos dados. Neste capítulo, veremos *as medidas de tendência central*. Antes, porém, de descrever essas medidas, precisamos apresentar alguns símbolos matemáticos.

3.1 Símbolos matemáticos

Para representar os valores numéricos de n unidades, escrevemos:

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

O subscrito i indica a posição da medida, portanto x_i é a i-ésima observação; x_1 representa a primeira observação, x_2 representa a segunda e os três pontos são lidos como "e assim por diante".

Exemplo 3.1 Representação de dados

Os pesos, em quilogramas, de cinco recém-nascidos são:

3,500	2,750	3,250	2,250	3,750

Em termos de símbolos, podemos escrever:

$$x_1 = 3,500; \quad x_2 = 2,750; \quad x_3 = 3,250; \quad x_4 = 2,250; \quad x_5 = 3,750.$$

A sequência x_1, x_2, x_3, x_4, x_5 *não* é ordenada pela grandeza dos dados. Veja o Exemplo 3.1: o primeiro bebê da amostra não é o menor, ainda que o bebê maior seja o último. Quaisquer que sejam os dados, os valores $x_1, x_2, x_3, \ldots x_n$ são registrados na ordem em que foram observados.

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

$$x_1 + x_2 + x_3 + \ldots + x_n$$

ou de forma muito mais compacta:

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

que se lê *somatório de χ índice i, i de 1* a *n*. O símbolo Σ, que indica o somatório, é a letra grega sigma maiúscula. Sob o símbolo Σ, está o subscrito $i = 1$, e, sobre o símbolo Σ, está n, indicando que o somatório se estende de x_1 até x_n.

Exemplo 3.2 Notação de somatório

No Exemplo 3.1, são dados os pesos de cinco bebês:
$x_1 = 3{,}500$; $x_2 = 2{,}750$; $x_3 = 3{,}250$; $x_4 = 2{,}250$; $x_5 = 3, 750$

A soma desses pesos, usando a notação de somatório, fica como segue:

$$\sum_{i=1}^{5} x_i = 3, 500 + 2, 750 + 3, 250 + 2, 250 + 3, 750 = 15, 500$$

Quando é fácil saber o número de parcelas que devem ser somadas pelo próprio texto, é usual escrever apenas Σχ em vez de $\sum_{i=1}^{n} x_i$

3.2 Média aritmética

A média aritmética, ou simplesmente média do conjunto de dados, é obtida somando-se todos os dados e dividindo-se o resultado da soma pelo número deles.

$$\text{Média} = \frac{\text{Soma de todos os dados}}{\text{Número de dados}}$$

A fórmula da média é:

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

que se lê *x-traço (ou x-barra) é igual ao somatório de x, dividido por* n.

A média aritmética é uma medida de tendência central. É o centro de equilíbrio do conjunto de dados. Para entender isso, imagine que os dados estejam apresentados no eixo das abscissas e que esse eixo represente os braços de uma balança. A média fica no fulcro da balança, ou seja, no centro de equilíbrio.

Exemplo 3.3 Cálculo da média

Um professor de Educação Física mediu a circunferência abdominal de dez homens que se apresentaram em uma academia de ginástica. Então, obteve os seguintes valores, em centímetros: 88; 83; 79; 76; 78; 70; 80; 82; 86; 106. A média é:

$$\overline{x} = \frac{88 + 83 + 79 + 76 + 78 + 70 + 80 + 82 + 86 + 106}{10} = \frac{827}{10} = 82,8$$

ou seja, a média da circunferência abdominal desses homens é 82,8cm.

Agora, observe a Figura 3.1. Imagine que o eixo das abscissas seja o braço de uma balança e que cada ponto tenha uma unidade de massa. Para haver equilíbrio, é preciso que o fulcro da balança esteja localizado onde está a média (ou seja, no ponto em que foi desenhada uma flecha).

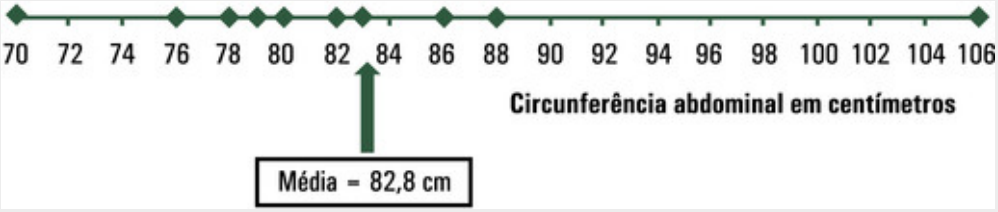


FIGURA 3.1 Distribuição de dados de circunferência abdominal, em centímetros, sobre um eixo, e a respectiva média

3.2.1 Média de dados agrupados em tabelas de distribuição de frequências

Quando os dados são discretos e em grande número, pode haver repetição de valores. Nesses casos, como vimos no Capítulo 1, é razoável organizar os dados em uma *tabela de distribuição de frequências*. Veja a Tabela 3.1.

Tabela 3.1

Tabela de distribuição de frequências

Dados	Frequência
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 frequências é dada por

$$\overline{x} = \frac{\Sigma xf}{\Sigma f}$$

Exemplo 3.4 Média de dados agrupados

Uma psicóloga que trabalha em Recursos Humanos de uma empresa perguntou aos vinte funcionários qual era o número de filhos em idade escolar. Os dados estão apresentados na Tabela 3.2.

Tabela 3.2

Número de filhos em idade escolar de vinte funcionários

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

Para calcular a média, a psicóloga construiu a Tabela 3.3, que é uma distribuição de frequências.

Tabela 3.3

Distribuição de frequências para o número de filhos em idade escolar de vinte funcionários

Número de filhos em idade escolar	Frequência
0	6
1	8
2	4
3	1
4	0
5	1

A Tabela 3.4 apresenta os cálculos intermediários para obter a média: cada valor (x) foi multiplicado pela respectiva frequência (f). A soma foi dividida pela soma das frequências (Σf).

Tabela 3.4

Cálculos auxiliares

Número de filhos em idade escolar	Frequência	Produto
(x)	(f)	(xf)
0	6	0
1	8	8
2	4	8
3	1	3
4	0	0
5	1	5
Total	$\Sigma f = 20$	$\Sigma xf = 24$

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

Quando os dados são contínuos e em grande quantidade, é comum *não* apresentar os dados brutos, mas apenas as tabelas de distribuição de frequências. Veja o Exemplo 3.5. Para calcular a média de dados agrupados em classes, é preciso calcular o *ponto médio* (ou *valor central*) de cada classe.

O ponto médio da classe é a média dos dois extremos da classe.

Exemplo 3.5 Média de dados contínuos agrupados

Os dados apresentados no Exemplo 1.10 (Cap. 1) foram agrupados em faixas de peso na Tabela 1.11, reproduzida na Tabela 3.5.

Tabela 3.5

Nascidos vivos segundo o peso ao nascer em quilogramas

Classe	Frequência
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

Para calcular a média, é preciso obter o ponto médio de cada classe. A classe 1,5 ⊢ 2,0 tem dois extremos: o inferior, que é 1,5, e o superior, que é 2,0. O ponto médio dessa classe é:

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

Os demais pontos médios são obtidos da mesma forma. Agora, construa uma tabela com os cálculos auxiliares. Escreva as classes, os pontos médios (x^*), as frequências (f) de classe e os produtos x^*f, como mostra a Tabela 3.6.

Tabela 3.6

Cálculos auxiliares

Classe	Valor central (x^*)	Frequência (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		$\Sigma f = 100$	$\Sigma x^*f = 300{,}00$

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

A média é, de longe, a medida de tendência central mais usada e, talvez por isso, a mais conhecida[2]. Quem nunca ouviu falar na *média de aprovação* em determinada disciplina ou no *tempo médio de uma viagem (de São Paulo ao Rio de Janeiro, por exemplo)* ou na *idade média dos jogadores de futebol*? Em certas circunstâncias, porém, outras medidas de tendência central, como a *mediana* ou a *moda*, dão melhor informação. Mas o que é mediana e o que é moda?

3.3 Mediana

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 e 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 estão 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 3.6 Cálculo da mediana

Para obter a mediana do peso dos cinco bebês do Exemplo 3.1, coloque os dados em ordem crescente, como segue:

2,250; 2,850; 3,250; 3,500; 3,970

A mediana está no centro dos dados ordenados. Corresponde a 3,250 kg, mostrado na Figura 3.2.

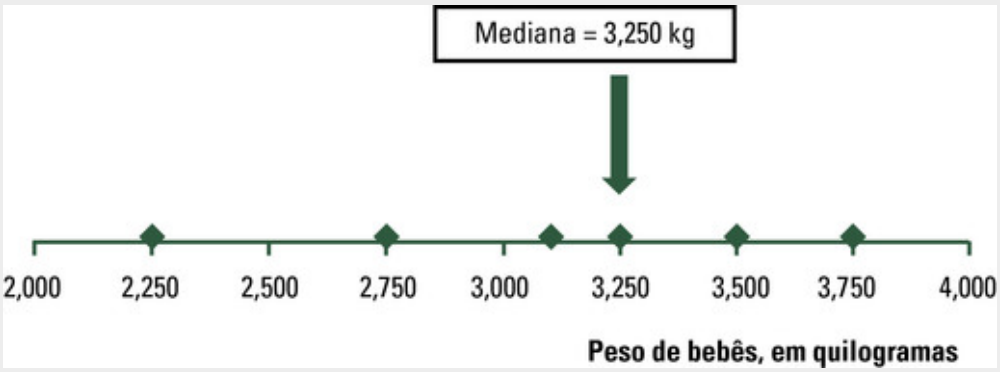


FIGURA 3.2 Distribuição dos pesos de bebês em quilogramas, sobre um eixo, e a respectiva mediana

Em algumas circunstâncias, a mediana descreve, melhor do que a média, a tendência central dos dados. É o caso dos conjuntos com *dados discrepantes*, ou seja, de conjuntos de dados que têm um ou alguns valores bem maiores ou bem menores que os demais. Veja o Exemplo 3.7.

Exemplo 3.7 Decidindo entre média e mediana

São dados: 42, 3, 9, 5, 7, 9, 1, 9. Para obter a média, calcule:

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

Para obter a mediana, é preciso ordenar os dados:

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

e calcular a média aritmética dos valores 7 e 9, que ocupam a posição central dos dados ordenados. Então, a mediana é 8.

A mediana descreve melhor o conjunto de dados porque o valor 42, que é discrepante, "puxa" a média para cima. Entretanto, o valor discrepante não afeta a mediana.

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 = R$ 0,00;
- na segunda vez, x_2 = R$ 0,00;
- na terceira vez, x_3 = R$1.000.000,00.

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

3.4 Moda

Moda é o valor que ocorre com maior frequência.

Exemplo 3.8 Determinando a moda

A moda dos dados 0, 0, 2, 5, 3, 7, 4, 7, 8, 7, 9, 6 é 7, porque é o valor que ocorre maior número de vezes.

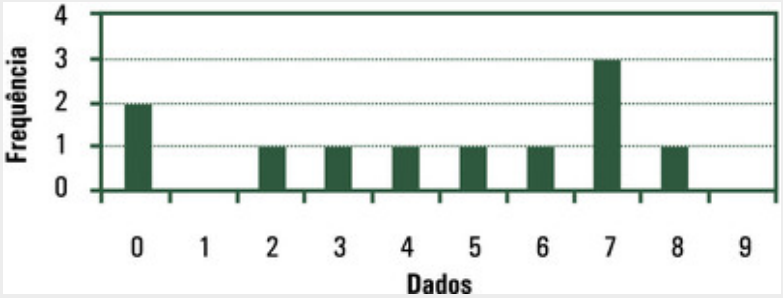


FIGURA 3.3 Distribuição dos dados sobre um eixo e a respectiva moda

Um conjunto de dados pode não ter moda ou ter duas ou mais modas. Assim, o conjunto de dados 0, 2, 4, 6, 8, 10 não tem moda, enquanto o conjunto 1, 2, 2, 3, 4, 4, 5, 6, 7 tem duas modas: 2 e 4.

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

Exemplo 3.9 Classe modal

A distribuição da população brasileira no Censo 2010, segundo o sexo e o grupo de idade, é apresentada na Tabela 3.7. Para homens, a classe modal (com maior frequência, em negrito na tabela) é o grupo de 10 a 14 anos. Para mulheres, a classe modal (em negrito na tabela) é o grupo de 25 a 29 anos.

Tabela 3.7

Distribuição da população brasileira segundo sexo e grupo de idade. Brasil, 2010

Grupo de idade	Sexo	
	Homens	Mulheres
0 a 4 anos	7.016.987	6.779.172
5 a 9 anos	7.624.144	7.345.231
10 a 14 anos	**8.725.413**	8.441.348
15 a 19 anos	8.558.868	8.432.002
20 a 24 anos	8.630.227	8.614.963
25 a 29 anos	8.460.995	**8.643.418**
30 a 34 anos	7.717.657	8.026.855
35 a 39 anos	6.766.665	7.121.916
40 a 44 anos	6.320.570	6.688.797
45 a 49 anos	5.692.013	6.141.338
50 a 54 anos	4.834.995	5.305.407
55 a 59 anos	3.902.344	4.373.875
60 a 64 anos	3.041.034	3.468.085
65 a 69 anos	2.224.065	2.616.745
70 a 74 anos	1.667.373	2.074.264
75 a 79 anos	1.090.518	1.472.930
80 a 84 anos	668.623	998.349
85 a 89 anos	310.759	508.724
90 a 94 anos	114.964	211.595
95 a 99 anos	31.529	66.806
Mais de 100 anos	7.247	16.989

Fonte: IBGE.[3]

[3]Disponível em www.ibge.gov.br/.../caracteristicas_da_populacao_tab_**brasil**. Acesso em: 8 set. 2014.

A moda é a única medida de tendência central que também pode ser usada para descrever *dados qualitativos*. Nesse caso, a moda é a *categoria* da variável que ocorre com maior frequência.

Exemplo 3.10 Determinação da moda

Veja os dados apresentados na Tabela 3.8. O grupo sanguíneo O ocorreu com maior frequência, então é a moda.

Tabela 3.8

Distribuição de indivíduos segundo o grupo sanguíneo

Grupo sanguíneo	Frequência
O	550
A	456
B	132
AB	29
Total	1.167

A moda é bastante informativa quando o conjunto de dados é grande. Se o conjunto de dados for relativamente pequeno (menos de trinta 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.

3.5 Exercícios resolvidos

3.5.1. Com base nos dados da Tabela 3.9, calcule o peso médio dos ratos em cada idade.

Tabela 3.9

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

Número do rato	Idade				
	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:

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

As médias para as demais idades, obtidas da mesma maneira, estão apresentadas na Tabela 3.10 e mostram que o peso médio dos ratos aumenta com a idade.

Tabela 3.10

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

Idade	Média
30	63,6
34	74,6
38	81,4
42	94,6
46	105,6

3.5.2. Determine a mediana dos dados apresentados na Tabela 1.8 (Cap. 1).

Para obter a mediana, os dados da Tabela 1.8 foram arrumados em ordem crescente na Tabela 3.11.

Tabela 3.11

Número de faltas de trinta funcionários ao trabalho. Clínica ABC, segundo semestre de 2014 (em ordem crescente)

0	1	2
0	1	2
0	1	2
0	1	2
0	**1**	3
0	**1**	3
0	1	3
0	1	4
0	1	4
1	2	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 não faltou ou faltou apenas um dia.

3.5.3. Foi feito um ensaio clínico randomizado para testar o efeito de um analgésico em cinco pacientes com osteoartrite. Os pacientes foram designados para receber placebo (2 × ao dia) ou droga (60mg 2 × ao dia), em datas diferentes, por processo aleatório.

Os dados, apresentados na Tabela 3.12, correspondem às medidas da dor à noite relatadas pelos pacientes (0 = nenhuma dor; 100 = dor extrema). Calcule, para cada paciente, as diferenças entre os valores obtidos no final e no início da pesquisa, para placebo e para a droga. Calcule as médias dessas diferenças. Discuta.

Tabela 3.12

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

Nº do paciente	Placebo		Nova droga	
	Início	Final	Início	Final
1	80	70	80	60
2	70	50	75	50
3	75	50	45	25
4	75	85	50	20
5	65	65	60	30

Tabela 3.13

Diferenças entre início e final do tratamento

Nº do paciente	Diferença	
	Placebo	Droga
1	–10	–20
2	–20	–25
3	–25	–20
4	10	–30
5	0	–30
Soma	–45	–125

As médias das diferenças são -9,0 para placebo e -25,0 para o anti-inflamatório. Os pacientes relataram maior alívio da dor quando receberam a droga com efeito analgésico.

3.6 Exercícios propostos

3.6.1. 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

3.6.2. Imagine que você esteja dirigindo um carro em uma estrada e observe que o número de veículos que você ultrapassa é igual ao número de veículos que ultrapassam você. Nesse caso, a velocidade de seu carro corresponde – considerando a velocidade de todos esses carros – a qual medida de tendência central?

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

3.6.4. 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?

3.6.5. Na Tabela 3.14, são apresentadas 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 ensaio pré-clínico para o teste de determinada droga. Encontre média e mediana.

Tabela 3.14

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

N° do rato	Taxa de glicose
1	101
2	98
3	97
4	104
5	95
6	105

3.6.6. Na Tabela 3.15, são 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 3.15

Estaturas (em metros), pesos (em quilogramas) e pressão arterial (em milímetros de mercúrio) de 11 pacientes hospitalizados com AVC

Nº do paciente	Estatura	Peso	Pressão arterial
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

3.6.7. Com os dados apresentados na Tabela 3.16, calcule o número médio de dentes cariados, para cada sexo.

Tabela 3.16

Estudantes de 12 anos, segundo o número de dentes cariados e o sexo

Número de dentes cariados	Sexo	
	Masculino	Feminino
0	16	13
1	2	5
2	3	3
3	2	2
4	2	2

3.6.8. Para estudar o tempo de latência de um sonífero usando ratos de laboratório, um pesquisador administrou o sonífero a dez ratos e determinou o tempo que levavam para dormir. Dos dez ratos, dois precisaram de meio minuto, quatro, de 1 minuto, três, de 1,5 minuto, e 1 não dormiu. Calcule o tempo médio de latência.

3.6.9. Determine média, mediana e moda para cada sexo, em relação aos dados apresentados na Tabela 3.17.

Tabela 3.17

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

Sexo	
Masculino	Feminino
6	4
9	10
6	6
8	8
7	6
6	8

3.6.10. Determine média, mediana e moda para cada sexo, em relação aos dados de volume diário de urina, apresentados na Tabela 3.18.

Tabela 3.18

Volume diário de urina (em litros), por sexo

Sexo	
Masculino	Feminino
0,5	0,9
1,4	0,6
0,9	0,5
0,8	1,3
1,3	0,8
0,5	0,7

3.6.11. Determine mediana e moda para os dados apresentados na Tabela 3.19 e interprete-as.

Tabela 3.19

Tempo de retorno (em dias) às atividades diárias de pacientes submetidas a histerectomia

N° da paciente	Tempo de retorno
1	20
2	30
3	15
4	20
5	40
6	50
7	25
8	30
9	15
10	35

3.6.12. Determine a média dos dados apresentados na Tabela 3.20.

Tabela 3.20

Teor de vitamina C (miligramas de ácido ascórbico em 100 mL) em dez caixas de 100 mL de suco de maçã encontradas no mercado

N° da caixa	Teor de vitamina C
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

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

3.6.14. Qual das medidas de tendência central não pode ser calculada para os dados da Tabela 3.21? Por quê?

Tabela 3.21

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

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

[1]Já disse alguém: “um desenho vale por mil palavras”.

[2]Há quem pretenda ser engraçado dizendo que a média não faz sentido porque, por exemplo, se alguém tem os pés na geladeira e a cabeça no forno, na média está em temperatura agradável. O fato é que, para relatar o comportamento de uma variável, a média não basta. É necessária, mas não é suficiente. Veja o Capítulo 4.

CAPÍTULO 4

Medidas de Dispersão para uma Amostra

As medidas de tendência central resumem a informação contida em um conjunto de dados, mas não contam toda a história. Por exemplo, observa-se, diariamente, que, na mesma cidade, a temperatura varia ao longo do dia. Então, a temperatura média do dia não dá toda a informação. 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 3 *não são suficientes* para descrever um conjunto de dados: informam apenas a *tendência central,* ou seja, onde está o centro, mas nada dizem sobre a variabilidade.

Para entender esse ponto, imagine dois domicílios: no primeiro, moram sete pessoas, todas com 22 anos. 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 e outro com 3 anos –, a mãe da moça – com 38 anos –, um irmão da moça – com 8 anos – e a avó da moça – com 65 anos. A média de idade nesse segundo domicílio 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 descritivas de um conjunto de dados quanto menor é a *variabilidade*. Então, quando você apresentar um conjunto de dados, deve fornecer não apenas medidas de tendência central, mas também uma *medida de variabilidade ou dispersão*. Veremos, neste capítulo, algumas formas de medir variabilidade.

4.1 Mínimo, máximo e amplitude

Mínimo de um conjunto de dados é o número de menor valor.
Máximo de um conjunto de dados é o número de maior valor.

Para medir variabilidade, você pode fornecer o valor mínimo e o valor máximo do conjunto de dados. Pode, também, calcular a *amplitude.*

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

$$amplitude = m\acute{a}ximo - m\acute{i}nimo$$

Exemplo 4.1 Mínimo, máximo e amplitude

A idade das crianças que estão no pátio de uma escola é, respectivamente: 3, 6, 5, 7, 9 anos. É fácil apresentar, em uma tabela, o número de crianças, a mediana, o mínimo, o máximo e a amplitude. Você primeiro ordena os dados como segue: 3, 5, 6, 7, 9. A mediana é 6 e a amplitude é:

$$amplitude = 9 - 3 = 6$$

Tabela 4.1

Estatísticas da idade das crianças

Estatísticas	Resultados
Número de crianças	5
Mediana	6
Mínimo	3
Máximo	9
Amplitude	6

A amplitude de variação é uma ideia básica em Estatística, mas um valor discrepante – por ser muito grande ou muito pequeno – aumenta muito a amplitude. Como dizem os estatísticos, a amplitude é muito *sensível* aos valores discrepantes.

Exemplo 4.2 Comparação de amplitudes

É dado o barulho do tráfego em duas esquinas, medido em decibéis durante os cinco dias úteis de determinada semana. Vamos calcular as amplitudes dos dados de cada conjunto.

1ª esquina: 56; 54; 51; 58; 52; 60.

2ª esquina: 56; 54; 58; 52; 51; 67.

1ª esquina: $amplitude = 60 - 51 = 9$

2ª esquina: $amplitude = 67 - 51 = 16$

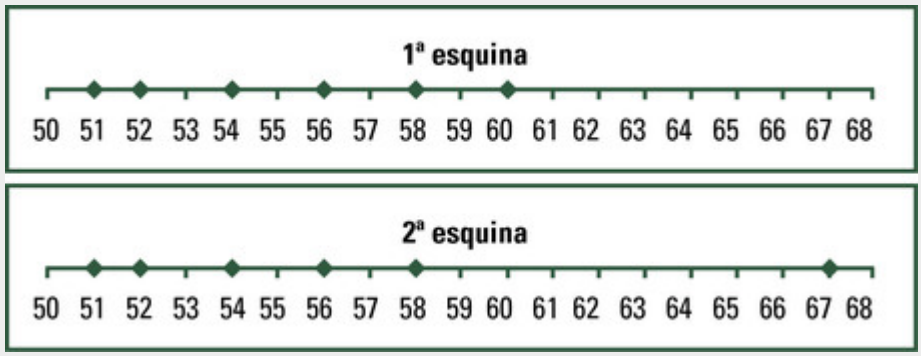


FIGURA 4.1 Distribuição de dados de barulho, em decibéis, sobre um eixo, para os dois conjuntos

Note que a amplitude maior na segunda esquina é explicada por 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 (67) aumentou, em muito, a amplitude dos dados obtidos na segunda esquina.

A amplitude é bastante usada como medida de variabilidade, mas, para descrever um conjunto de dados, muitas vezes é melhor fornecer mínimos e máximos. Por exemplo, se alguém informar que os policiais que estão na ativa em certa corporação têm idades entre 18 e 52 anos, estará fornecendo uma informação mais útil do que se disser que a amplitude das idades é 34 anos.

4.2 Quartil

A mediana, que você viu no Capítulo 3, divide um conjunto de dados em dois subconjuntos com o mesmo número de dados:

- o que antecede a mediana (dados iguais ou inferiores à mediana);
- o que sucede a mediana (dados iguais ou superiores à mediana).

Se o número de observações for grande (digamos, maior de trinta), o conceito de mediana pode ser entendido 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 são pontos que dividem o conjunto de dados ordenados em quatro partes, de modo que cada parte contenha 25% dos dados. O primeiro quartil (Q_1) ocupa a posição central entre a mediana e o dado de menor valor. O segundo quartil é a *mediana* do conjunto de dados. O terceiro quartil (Q_3) ocupa a posição central entre a mediana e o dado de maior valor. Então, se um item está "no quartil superior", significa que está entre os 25% de itens de maior valor.

Para obter os quartis[1] quando o conjunto tem um *número ímpar* de dados:

1. organize os dados em ordem crescente. Encontre a mediana, que é o segundo quartil; marque esse valor;
2. se o número de dados for ímpar, a mediana é um número que está no conjunto. Para achar o primeiro quartil, tome o conjunto de dados iguais ou menores que a mediana; o primeiro quartil é a mediana do novo conjunto de dados;
3. para encontrar o terceiro quartil, tome o conjunto de dados iguais ou maiores do que a mediana; o terceiro quartil é a mediana do novo conjunto de dados.

Exemplo 4.3 Obtendo os quartis de conjunto com número ímpar de dados

O número de dados no conjunto 1, 2, 3, 4, 5, 6, 7, 9, 10 é ímpar. Então, a mediana é o valor central dos dados ordenados, ou seja, 5.

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

Para obter o primeiro quartil, separe os dados *iguais ou menores* do que a mediana. Primeiro quartil é a mediana do novo conjunto de dados, ou seja, 3.

1, **2,** 3, 4, 5.

Para obter o terceiro quartil, separe os dados *iguais ou maiores* do que a mediana. Terceiro quartil é a mediana do novo conjunto de dados, ou seja, 7.

5, 6, **7,** 9, 10.

Se o conjunto tiver um *número par* de dados, para obter os quartis:

1. organize os dados em ordem crescente. Encontre a mediana, que é o segundo quartil; marque esse valor;
2. a mediana, dada pela média dos dois valores centrais, não é, necessariamente, um número igual a qualquer outro do conjunto de dados. Para encontrar o primeiro quartil, separe o conjunto de dados menores do que a mediana; o primeiro quartil é a mediana do novo conjunto de dados;
3. para achar o terceiro quartil, separe o conjunto de dados maiores do que a mediana; o terceiro quartil é a mediana do novo conjunto de dados.

Exemplo 4.4 Obtendo os quartis de conjunto com número par de dados

A mediana dos dados 0, 1, 2, 3, 4, 5, 5, 7, 9, 10 é a média dos dois valores que estão no centro dos dados ordenados, ou seja, 4,5.

0, 1, 2, 3, **4, 5,** 6, 7, 9, 10.

Para obter o primeiro quartil, separe os dados *menores* do que a mediana. O primeiro quartil é a mediana desse novo conjunto de dados, ou seja, 2.

0, 1, **2,** 3, 4.

Para obter o terceiro quartil, separe os dados *maiores* do que a mediana. O terceiro quartil é a mediana desse novo conjunto de

dados, ou seja, 7.

5, 6, **7,** 9, 10.

Pode parecer que o método apresentado para determinar quartis é confuso, mas é pior do que simplesmente confuso: os estatísticos não se entendem nesse assunto.[2] Existem vários métodos para obter quartis e os programas para computador empregam métodos diferentes. Por isso, se você calcular os quartis para o Exemplo 4.3 usando o Excel, encontrará resultados diferentes dos achados aqui e, se usar o Minitab, encontrará outros resultados. O SAS permite escolher entre cinco métodos. Além disso, os valores aqui calculados são chamados no Brasil de quartis (em inglês, *quartiles*), mas o autor[3] que inventou o *boxplot* os chama de "dobradiças" (em inglês, *hinges*). Felizmente, as diferenças entre resultados são pequenas e não afetam as conclusões de um trabalho.

De qualquer modo, é preciso definir *distância interquartílica,* que é uma medida de dispersão que aparece nos *boxplots*. Como a amplitude é muito sensível aos valores discrepantes, ou seja, muda de valor se for incluída uma observação discrepante, a distância interquartílica descreve melhor a dispersão dos dados.

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

Distância interquartílica = Terceiro quartil - Primeiro quartil.

Exemplo 4.5 Distância interquartílica

Vamos calcular as distâncias interquartílicas para o Exemplo 4.2. Reveja os seguintes dados:

1ª esquina: 56; 54; 51; 58; 52; 60

Para encontrar a distância interquartílica, comece ordenando os dados:

51; 52; 54; 56; 58; 60

O número de dados é par. A mediana é a média de 54 e 56, ou seja, 55. Ache o primeiro e o terceiro quartis. Então:

Mediana: 55

1° quartil: 52
3° quartil: 58
Distância interquartílica = 58 – 52 = 6
2ª esquina: 56; 54; 58; 52; 51; 67

Para encontrar a distância interquartílica, é preciso ordenar os dados, calcular a mediana e achar o primeiro e o terceiro quartis. Então:

51; 52; 54; 56; 58; 67

Mediana: 55
1° quartil: 52
3° quartil: 58
Distância interquartílica = 58 – 52 = 6

4.2.1 Diagrama de caixa (*Boxplot*)

As medidas que acabamos de ver – mínimo, primeiro quartil, mediana, terceiro quartil, máximo – permitem traçar o *diagrama de caixa*, que ajuda a entender a informação contida em um conjunto de dados.

Para desenhar um diagrama de caixa:

1. desenhe um segmento de reta em posição vertical, para representar a amplitude dos dados;
2. marque, nesse segmento, o primeiro, o segundo e o terceiro quartis;
3. 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;
4. faça um ponto para representar a mediana, obedecendo à escala, e sobre o segmento de reta anteriormente traçado.

Exemplo 4.6 Diagrama de caixa (*boxplot*)

A Figura 4.2 apresenta um diagrama de caixa para o conjunto de dados: 1; 2; 3; 4; 5; 6; 7; 8; 9; 10. Foram calculados:

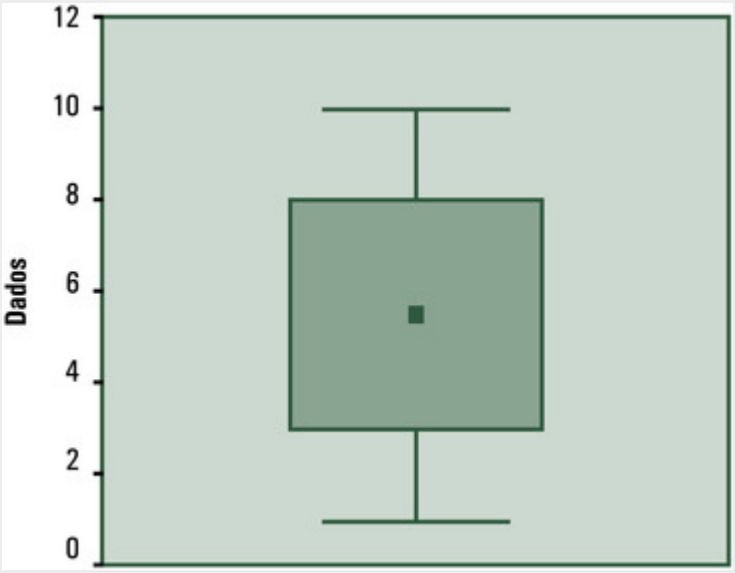


FIGURA 4.2 Diagrama de caixa

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

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.

4.3 Desvio padrão

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. Para calcular o desvio padrão, é preciso, primeiro, calcular a variância. Vamos, então, entender o que é variância.

4.3.1 Cálculo da variância

Quando a média é usada como medida de tendência central, podemos calcular o desvio de cada dado em relação à média, como segue:

$$\text{desvio} = \text{dado} - \text{média}$$

$$d_i = x_i - \bar{x}$$

Exemplo 4.7 Desvios em relação à média

No Exemplo 4.1, são dadas as idades de cinco crianças: 3, 6, 5, 7 e 9 anos. Para calcular os desvios em relação à média, subtraímos a média de cada observação. Como a média é 6, os desvios são os valores apresentados na Tabela 4.2.

Tabela 4.2
Cálculo dos desvios

Observação	Desvio
x	
3	3 – 6 = –3
6	6 – 6 = 0
5	5 – 6 = –1
7	7 – 6 = 1
9	9 – 6 = 3

Desvios pequenos significam dados *aglomerados* em torno da média, enquanto desvios grandes significam dados *dispersos* em torno da média. Mas esses desvios precisam ser resumidos em *um só número*, para que você possa olhar esse número e julgar o grau de variabilidade dos dados. Como é possível fazer isso?

À primeira vista, parece possível calcular a *média dos desvios*. Mas a média seria sempre igual a zero, porque a soma dos desvios negativos é sempre igual à soma dos desvios positivos. O "peso" dos desvios negativos é igual ao "peso" dos desvios positivos, uma vez que a média dá a *tendência central dos dados.* Isso pode ser verificado em qualquer conjunto de dados. No Exemplo 4.7:

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

É preciso eliminar os sinais antes de somar. É intuitivo pensar em calcular a média dos valores absolutos dos desvios. Essa medida realmente existe. É o *desvio médio*, pouco encontrado nos trabalhos de Estatística.[4] Mas existe outra maneira de eliminar os sinais: elevam-se os valores ao quadrado. A soma assim obtida é denominada *soma*

de quadrados dos desvios. A partir dessa soma, obtém-se a *variância*. Veja a definição de variância da amostra, que é indicada por s^2.

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

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

Para calcular a variância:

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

Exemplo 4.8 Calculando a variância

A Tabela 4.3 apresenta os cálculos intermediários para obter a variância dos dados do Exemplo 4.1.

Tabela 4.3

Cálculo da variância

Observação x	Desvio	Desvio ao quadrado
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	$9 - 6 = 3$	$3^2 = 9$
$\Sigma x = 30$	$\Sigma (x\text{-}x) = 0$	

A variância é

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

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

A variância quantifica a variabilidade dos dados. O divisor, n-1, recebe o nome de *graus de liberdade*.[5]

4.3.1.1 Outra fórmula para calcular a variância

A fórmula dada na Seção 4.3.1 para calcular a variância da amostra pode ser algebricamente desenvolvida. Obtém-se, então, uma segunda fórmula que, embora, à primeira vista, pareça mais complicada, permite que o cálculo da variância seja feito com menor número de operações aritméticas.[6] Prefira usar esta segunda fórmula se você fizer cálculos à mão, o que é pouco provável.

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

Exemplo 4.9 Calculando a variância

São dados os tempos em minutos que seis meninos permaneceram sobre seus skates: 4; 6; 4; 6; 5; 5. Para calcular a variância, foram feitos os cálculos intermediários que estão na Tabela 4.4.

Tabela 4.4
Cálculo da variância

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

A variância é

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

4.3.1.2 Variância de dados agrupados em tabelas de distribuição de frequências

A variância de dados agrupados em uma tabela de distribuição de frequências, ou seja, de $x_1, x_2, \ldots x_n$ que se repetem $f_1, f_2, \ldots f_n$ vezes na amostra, é

$$s^2 = \frac{\Sigma x^2 f - \frac{(\Sigma xf)^2}{\Sigma f}}{\Sigma f - 1}$$

Exemplo 4.10 Calculando a variância de dados agrupados

Reveja o Exemplo 3.4 (Cap. 3). Foi construída a Tabela 3.3, reapresentada aqui como Tabela 4.5.

Tabela 4.5

Distribuição de frequências para o número de filhos em idade escolar de vinte funcionários

Número de filhos em idade escolar	Frequência
0	6
1	8
2	4
3	1
4	0
5	1

A Tabela 4.6 apresenta os cálculos intermediários para se obter a variância.

Tabela 4.6

Cálculos auxiliares para obtenção da variância

Número de filhos em idade escolar	Frequência	Produto	Produto
(x)	(f)	(xf)	(x^2f)
0	6	0	0
1	8	8	8
2	4	8	16
3	1	3	9
4	0	0	0
5	1	5	25
Total	$\Sigma f = 20$	$\Sigma xf = 24$	$\Sigma x^2f = 58$

Aplicando a fórmula:

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

$$s^2 = \frac{58 - \frac{24^2}{20}}{20 - 1} = \frac{58 - 28,8}{19} = 1,54$$

4.3.2 Desvio padrão

Lembre-se de que, para calcular a variância, os desvios em relação à média foram *elevados ao quadrado*. Então, a unidade de medida da variância é igual ao *quadrado* da medida das observações. Logo, extraindo a raiz quadrada da variância, você obtém uma medida de variabilidade com a mesma unidade de medida dos dados. É o *desvio padrão*.

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

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

Exemplo 4.11 Calculando o desvio padrão

A Tabela 4.7 apresenta a duração, em minutos, das chamadas telefônicas realizadas em três consultórios médicos. As médias, as

variâncias e os desvios padrões são apresentados na Tabela 4.8. As observações foram medidas em minutos, a variância é dada em minutos ao quadrado, o que não tem sentido prático, mas o desvio padrão é dado em minutos.

Tabela 4.7

Tempo (em minutos) das 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 4.8

Estatísticas obtidas para os dados da Tabela 4.7

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 realizadas nos três consultórios médicos foi, em média, a mesma, ou seja, 5 minutos. No entanto, a duração das chamadas variou significativamente entre os consultórios. Compare, por exemplo, o desvio padrão 0,89 minuto, do consultório A, com o desvio padrão 4,05 minutos, do consultório C.

4.4 Coeficiente de variação

Coeficiente de variação é a razão entre o desvio padrão e a média.

O resultado do cálculo do coeficiente de variação é multiplicado por 100, para ser apresentado 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 de 3, 1 e 5 anos; a média é, evidentemente, 3 anos. No segundo grupo, as pessoas têm idades de 55, 57 e 53 anos: portanto, a média é 55 anos.

Verifique que, nos dois grupos, a dispersão dos dados é idêntica: ambos têm variância $s^2 = 4$. No entanto, as diferenças de dois 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, em que a média é de 3 anos (o coeficiente de variação é 66,67%), do que no segundo grupo, que tem média de 55 anos (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*, ou seja, não tem unidade de medida. Isso acontece porque média e desvio padrão são medidos na mesma unidade – que, então, 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 4.5.3.

4.5 Exercícios resolvidos

4.5.1. Vamos calcular a média e a variância do nível de colesterol de cinco pessoas: 260; 160; 200; 210; 240.

A média é

$$\overline{x} = \frac{260 + 160 + 200 + 210 + 240}{5} = \frac{1070}{5} = 214,0$$

Para obter a variância, foram feitos os cálculos intermediários apresentados na Tabela 4.9.

Tabela 4.9

Cálculos intermediários para obtenção da variância

Nível de colesterol	Desvio em relação à média	Desvio ao quadrado
260	46	2116
160	–54	2916
200	–14	196
210	–4	16
240	26	676
Soma	0	5.920

A variância é:

$$s^2 = \frac{5920}{4} = 1480,00$$

4.5.2. Observe os conjuntos A; B; C; D de dados. Sem fazer cálculos, qual deles apresenta menor variância? Quais têm

maior variância?

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 todos iguais entre si. O conjunto B tem variância maior do que o conjunto A, pois os dados variam de 1 em 1. Os conjuntos C e D têm variâncias maiores do que as dos outros, mas iguais entre si (em ambos os conjuntos, os dados variam de 2 em 2).

4.5.3. Calcule a média, o desvio padrão e o coeficiente de variação dos dados apresentados na Tabela 4.10. Comente os resultados.

Tabela 4.10

Peso (em quilogramas) e comprimento (em centímetros) de dez cães

Peso	Comprimento
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 20,3kg e o desvio padrão é 3,74kg. O coeficiente de variação é 18,42%.

b. Para comprimento: a média é 102,3cm e o desvio padrão é 4,85cm. O coeficiente de variação é 4,74%.

Não se podem comparar desvios padrões de peso e comprimento, porque as unidades de medida são diferentes. No

entanto, 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%). Isso significa que os dados de peso variam mais em relação à média do que os dados de comprimento. Lembre-se de que isso também acontece em humanos adultos e normais: provavelmente, você conhece duas pessoas tais que uma tem o dobro de peso da outra (104 kg e 52 kg, por exemplo), mas não uma com o dobro da altura da outra.

4.5.4. Determine os quartis[7] 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.

Para obter o primeiro quartil, separe os dados menores do que a mediana (6). O primeiro quartil é a mediana desses dados, 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, 10.

7, 8, 10, 11, 11.

4.5.5. Foi feito um experimento para comparar dois programas de treinamento para a execução de um serviço especializado. Vinte homens foram selecionados para esse treinamento. Dez dos vinte foram escolhidos ao acaso e treinados pelo método A. Os outros dez foram treinados pelo método B. Concluído o período de treinamento, todos os homens executaram o serviço e foi medido o tempo de cada um. Os dados são apresentados na Tabela 4.11. Vamos calcular as estatísticas (apresentadas na Tabela 4.12) e desenhar diagramas de caixa (na Fig. 4.3) para comparar os métodos.

Tabela 4.11

Tempo (em minutos) despendido na execução do serviço, segundo o método de treinamento

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

Tabela 4.12

Tempo (em minutos) despendido na execução do serviço, segundo o método de treinamento

Estatística	Método	
	A	B
Mínimo	11	13
Primeiro quartil	16	19
Mediana	19	24
Terceiro quartil	23	28
Máximo	27	31

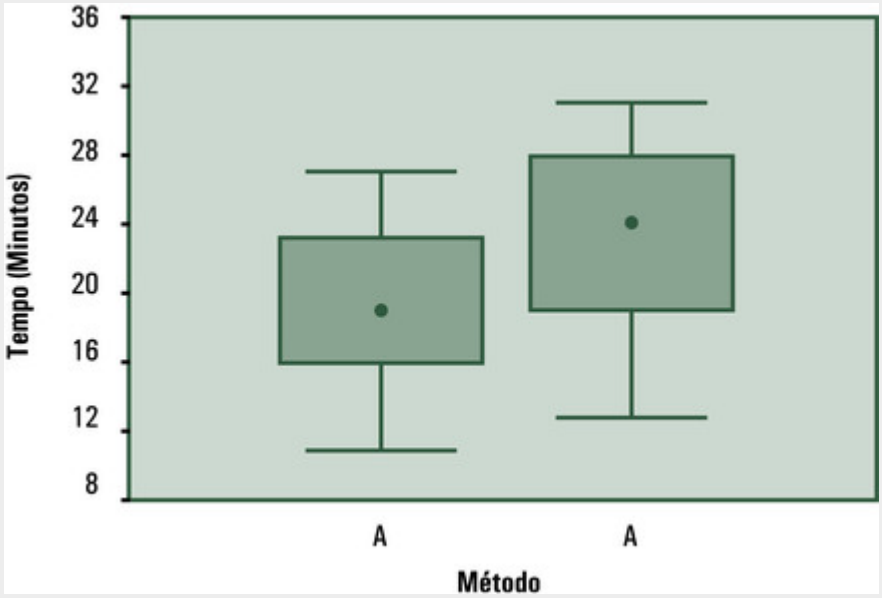


FIGURA 4.3 Comparação de dois diagramas de caixa.

A Figura 4.3 mostra que a variabilidade é praticamente a mesma para os dois métodos. No entanto, a mediana do tempo despendido por homens treinados pelo método A foi menor.

4.5.6. Vamos calcular a variância e o desvio padrão dos dados apresentados na Tabela 3.9 (Cap. 3) em cada idade e comentar o resultado.

A variância é dada pela seguinte fórmula:

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

Usando uma calculadora ou o computador, é possível obter as somas apresentadas na Tabela 4.13.

Tabela 4.13

Resultados parciais

Soma	30 dias	34 dias	38 dias	42 dias	46 dias
$\sum x$	509	597	651	757	845
$\sum x^2$	33.305	46.043	54.765	74.417	92.041
$(\sum x)^2$	259.081	356.409	423.801	573.049	714.025

As variâncias e os desvios padrões estão apresentados na Tabela 4.14. Os desvios padrões aumentam com a idade, ou seja, a dispersão dos dados em torno da média aumenta com a idade.

Tabela 4.14

Variância e desvio padrão do peso (em gramas) de ratos machos da raça Wistar, segundo a idade

Idade (em dias)				
30	34	38	42	46
131,41	213,13	255,70	397,98	398,27
11,5	14,6	16,0	19,9	20,0

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

4.6 Exercícios propostos

4.6.1. Dados os valores 5, 3, 2 e 1, calcule: a) o mínimo; b) o máximo; c) a amplitude.

4.6.2. Dados os valores 3, 8, 5, 6, 4, 3 e 6, calcule: a) $\Sigma\, x$; b) $\sum(x-\bar{x})^2$

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

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

4.6.5. 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?

4.6.6. São dadas, na Tabela 4.15, as notas de três alunos em cinco provas. Calcule, para cada aluno, a média e o desvio padrão das notas obtidas. Discuta.

Tabela 4.15

Notas de quatro alunos em cinco provas

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

4.6.7. Responda às seguintes 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?

4.6.8. Calcule a variância, o desvio padrão e o coeficiente de variação para os dados apresentados no Exercício 3.6.5 (Cap. 3).

4.6.9. 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.

4.6.10. Responda às seguintes 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?

4.6.11. Um professor de Odontologia queria saber se alunos que começam a atender pacientes em disciplinas clínicas têm aumento na frequência do batimento cardíaco. Então, mediu a frequência dos batimentos cardíacos de cinco alunos de primeiro ano (que não cursam disciplinas clínicas) e de cinco alunos do segundo ano, pouco antes do primeiro atendimento de pacientes. Os dados estão apresentados na Tabela 4.16. Calcule as médias e os desvios padrões. Discuta.

Tabela 4.16

Frequência de batimento cardíaco, medida em batimentos por minuto (bpm), de alunos de primeiro e segundo anos

1° ano	2° ano
87	106
70	100
76	86
71	96
69	90

4.6.12. 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 aferiu a perda de peso (em quilogramas) de cada paciente de cada grupo. Os dados estão apresentados na Tabela 4.17. Calcule as médias e as variâncias. Discuta.

Tabela 4.17

Perda de peso (em quilogramas), segundo a dieta

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

[1]Os métodos empregados para calcular os quartis apresentam 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*).

[2]Disponível em Defining Quartiles - Math Forum - Ask Dr. Math mathforum.org/library/drmath/view/60969.html. Acesso em: 4 ago. 2014.

[3]John Wilder Tukey.

[4]A introdução do valor absoluto numa fórmula torna muito mais complicado fazer o cálculo analítico posteriormente, em deduções teóricas.

[5]A soma dos desvios é sempre zero. Então, tendo os valores de $n - 1$ desvios, você pode calcular o valor do n-ésimo desvio que está faltando. Reveja o Exemplo 5.6. Dados os desvios -3, 0,-1 e 1, é fácil verificar que a soma deles é -3. Para que a soma seja zero, falta o desvio de valor 3. 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.

[6]Essa fórmula está sendo apresentada aqui porque é encontrada em muitos textos, mas corresponde à mesma fórmula dada na definição. Facilita os cálculos, mas, hoje, isso não tem sentido.

CAPÍTULO 5

Noções sobre Correlação

Você já ouviu falar que o número de pontos no Enem está relacionado ao grau de conhecimento dos alunos. Também já ouviu falar que o bom desempenho do atleta está relacionado a um bom treinamento. Essas afirmativas mostram que temos consciência de que pode haver *relação entre duas variáveis*. E você sabe que o risco de câncer de pulmão aumenta com o tempo do hábito de fumar e que a pressão arterial aumenta com a idade. Tais assertivas mostram que temos consciência da *evolução de uma variável ao longo do tempo*. Neste capítulo vamos ver como se estudam, em conjunto, duas variáveis.

5.1 Diagrama de dispersão

Vamos pensar em duas variáveis numéricas e chamar, como é habitual em Estatística, uma de X e a outra de Y. Se você medir essas duas variáveis em 22 pessoas, ou em 22 animais, ou em 22 objetos, terá 22 pares de valores dessas variáveis. Se X e Y têm a tendência de variar conjuntamente, dizemos que existe *correlação* entre ambas. Neste Capítulo, vamos ver como se responde às seguintes questões:

- É razoável considerar que existe correlação entre X e Y?
- Que tipo de correlação existe entre ambas?
- Qual é o grau dessa correlação?

É preciso desenhar gráficos e fazer alguns cálculos. Começaremos desenhando um diagrama de dispersão.

Diagrama de dispersão (*scaterplot*) é um gráfico feito para mostrar o grau de correlação entre duas variáveis.

Para desenhar o diagrama de dispersão:

) trace um sistema de eixos cartesianos e represente cada uma das variáveis em um dos eixos;

) estabeleça as escalas de maneira a dar ao diagrama o aspecto de um quadrado;

) escreva os nomes das variáveis nos respectivos eixos e, em seguida, faça as graduações;

) desenhe um ponto para representar cada um dos pares de valores das variáveis.

Exemplo 5.1 Diagrama de dispersão

Um fisioterapeuta mediu a altura (X) e o peso (Y) de 22 universitários. Os dados estão apresentados na Tabela 5.1 e o diagrama de dispersão na Figura 5.1. Observando a figura, você "vê" a variação conjunta de altura e peso: os pesos tendem a ser maiores para as alturas maiores.

Tabela 5.1

Altura (em metros) e peso (em quilogramas) de 22 universitários

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

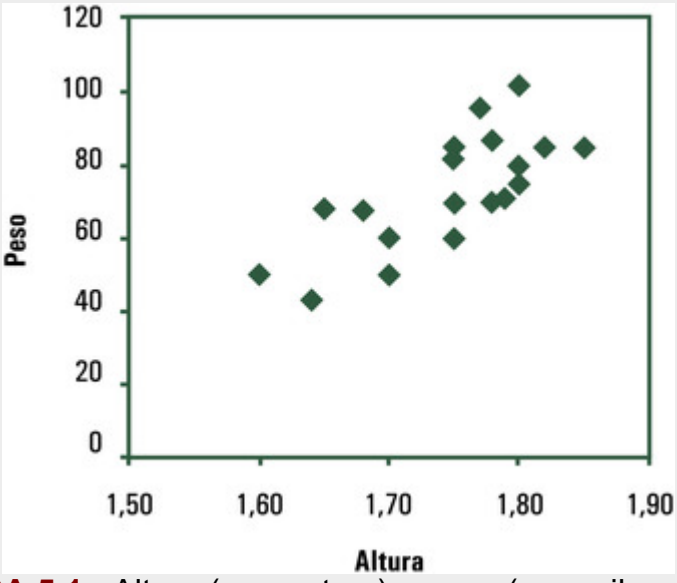


FIGURA 5.1 Altura (em metros) e peso (em quilogramas) de 22 universitários

Podemos considerar que existe *correlação entre X e Y* quando os dados apresentados no diagrama de dispersão formam uma nuvem de pontos que, de alguma forma, mostra a *variação conjunta das variáveis*. Veja o Exemplo 5.2.

Exemplo 5.2 Correlação forte, correlação fraca, correlação nula

Os dados apresentados na Tabela 5.2 estão apresentados nos diagramas da Figura 5.2. Veja que:

Tabela 5.2

Correlação forte, fraca e nula

Conjunto A		Conjunto B		Conjunto C	
X	Y	X	Y	X	Y
1	2	1	6	1	3
2	6	2	3	2	1
3	5	3	5	3	7
4	8	4	8	4	12
5	6	5	4	5	3
6	9	6	12	6	7
7	10	7	9	7	3
8	8	8	3	8	4
9	12	9	6	9	3
10	10	10	12	10	6

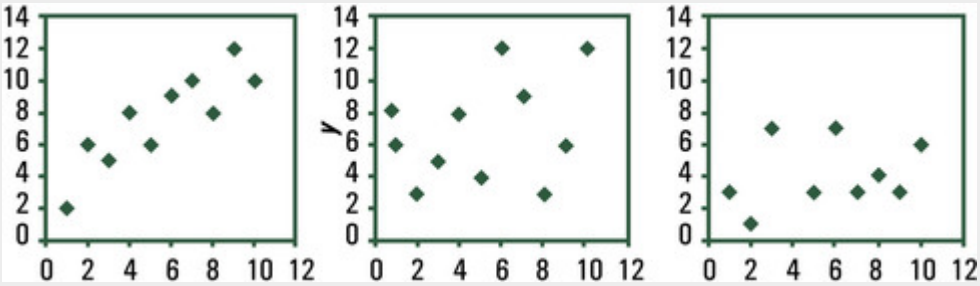


FIGURA 5.2 Correlação forte, fraca e nula

- no Conjunto A, os pontos estão distribuídos em torno e bem próximos de uma reta, mostrando variação conjunta: a correlação é *forte;*
- no Conjunto B, os pontos estão *espalhados* em torno de uma reta; embora exista variação conjunta, a correlação é *fraca*;
- no Conjunto C, X cresce e Y varia ao acaso; como a variação não é conjunta, não existe correlação entre as variáveis, ou seja, a correlação é *nula*.

Dizemos que a *correlação* entre duas variáveis é *positiva* quando X cresce e Y, em média, também cresce; dizemos que a *correlação* é *negativa* quando X cresce e Y, em média, decresce.

Exemplo 5.3 Correlação positiva e correlação negativa

A simples observação dos diagramas apresentados na Figura 5.3 deixa claro que, no Conjunto A, a correlação é *positiva*, enquanto, no Conjunto B, a correlação é *negativa*.

Tabela 5.3

Correlação positiva e correlação negativa

Conjunto A		Conjunto B	
X	*Y*	*X*	*Y*
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

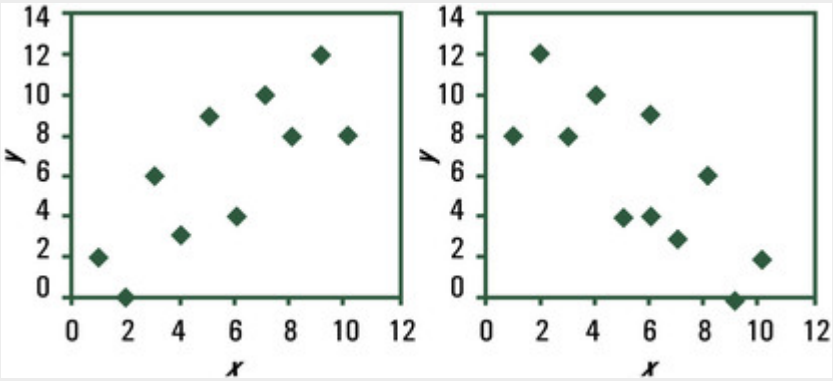


FIGURA 5.3 Correlação positiva e correlação negativa

A correlação entre duas variáveis pode ser *linear* ou *não linear*. Dizemos que a correlação é linear quando a nuvem de pontos que representam os dados se dispersa em torno de uma reta. A correlação é não linear quando a nuvem de pontos se dispersa em torno de uma curva. Neste livro, são estudadas apenas as relações lineares entre duas variáveis.

Exemplo 5.4 Relação linear e relação não linear entre duas variáveis

Tabela 5.4
Relação linear e relação não linear entre duas variáveis

Conjunto A		Conjunto B	
X	Y	X	Y
0	2	0	–0,084
1	3	1	1,764
2	4	2	2,844
3	5	3	3,156
4	6	4	2,7
5	7	5	1,476
6	8	6	–2,7

A Figura 5.4 exibe *correlações perfeitas*: no Conjunto A, os pontos estão sobre uma reta, enquanto, no Conjunto B, os pontos estão sobre uma parábola.

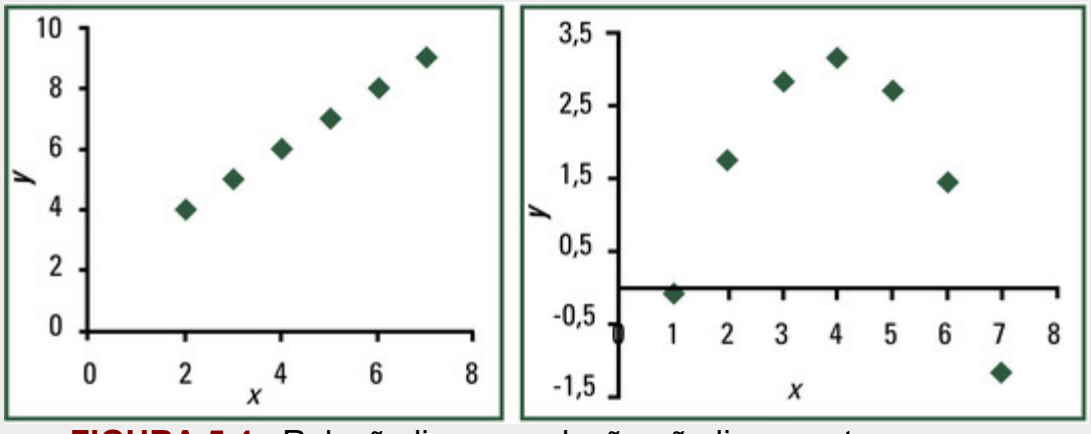


FIGURA 5.4 Relação linear e relação não linear entre duas variáveis

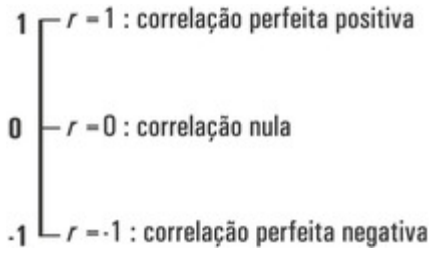
5.2 Cálculo do coeficiente de correlação

O *grau de correlação linear* entre duas variáveis numéricas X e Y é medido pelo *coeficiente de correlação de Pearson*,[1] que se representa por r e é definido pela seguinte fórmula:

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

Coeficiente de correlação de Pearson é a medida do grau de relação linear entre duas variáveis numéricas.

O valor de r só pode variar entre –1 e +1, inclusive, ou seja, $-1 \leq r \leq +1$. Então:



Para julgar o valor de r, existe uma regra prática, embora rudimentar:[2]

- correlação pequena
  $0 < r < 0{,}25$ ou $-0{,}25 < r < 0$
- correlação fraca
  $0{,}25 < r < 0{,}50$ ou $-0{,}50 < r < -0{,}25$
- correlação moderada

$0{,}50 < r < 0{,}75$ ou $-0{,}75 < r < -0{,}50$

- correlação forte

  $0{,}75 < r < 1{,}00$ ou $-1 < r < -0{,}75$

Nas ciências físicas, os coeficientes de correlação têm valores relativamente altos. Nas ciências da saúde, os coeficientes de correlação são menores, devido à grande variabilidade dos fenômenos biológicos. Nas ciências do comportamento, coeficientes de correlação iguais ou maiores que 0,70 são extremamente raros. Mas é importante saber que, para julgar o valor do coeficiente de correlação (r), deve ser aplicado um *teste estatístico*,[3] que leva em conta o *tamanho da amostra* (n).

Exemplo 5.5 Cálculo do coeficiente de correlação

Vamos calcular o coeficiente de correlação para os dados apresentados na Tabela 5.3 (Conjunto A). Os cálculos intermediários são apresentados na Tabela 5.5.

Tabela 5.5

Cálculos intermediários para a obtenção do coeficiente de correlação (Conjunto A da Tabela 5.3)

Conjunto A				
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
$\Sigma X = 55$	$\Sigma Y = 62$	$\Sigma XY = 423$	$\Sigma X^2 = 385$	$\Sigma Y^2 = 518$

Substituindo, na fórmula, os somatórios pelos valores calculados na Tabela 5.5 e lembrando que o tamanho da amostra é $n = 10$, obtemos:

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

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

$$r = 0,781$$

Usando a regra prática, podemos dizer que a correlação entre X e Y é positiva e moderada.

Exemplo 5.6 Cálculo do coeficiente de correlação

Vamos calcular o coeficiente de correlação para os dados do Conjunto B, apresentado na Tabela 5.3. Os cálculos intermediários são apresentados na Tabela 5.6.

Tabela 5.6

Cálculos intermediários para obter o coeficiente de correlação (Conjunto B da Tabela 5.3)

Conjunto B				
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 5.6 e lembrando que o tamanho da amostra é $n = 10$, obtemos:

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

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

$$r = 0,781$$

Aplicando a regra prática, dizemos que a correlação entre X e Y é negativa e moderada.

É necessário pressupor – para que se possa calcular o coeficiente de correlação – que:

1. cada unidade da amostra forneceu valores tanto de X como de Y;
2. as unidades foram selecionadas *ao acaso* – ou, pelo menos, são representativas de uma grande população;
3. as variáveis X e Y foram *medidas de forma independente*. Não tem sentido calcular o coeficiente de correlação se Y tiver sido obtido por meio de uma fórmula que inclui X.

Exemplo 5.7 Pressuposição necessária para o cálculo de *r*

Você pode calcular o coeficiente de correlação entre as notas obtidas pelos alunos de um curso na primeira prova (X) com as notas obtidas na segunda prova (Y). No entanto, não tem sentido correlacionar as notas obtidas na primeira prova (X) com as notas finais de aprovação (Z) se essas notas forem médias de todas as notas (que incluem a nota X da primeira prova).

5.3 Cuidados na interpretação do coeficiente de correlação

O diagrama de dispersão dá ideia da relação entre duas variáveis. Mas, para que o coeficiente de correlação de Pearson tenha significado, é preciso que 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 variam em sentidos contrários, mas não indica que acréscimos em uma das variáveis *causam* decréscimos na outra variável. E cuidado com o chavão: correlação não significa causa! *Pode* existir uma relação de causa e efeito entre as variáveis.

De qualquer forma, um exemplo antigo, mas muito interessante, foi apresentado por um estatístico que mostrou a existência de correlação positiva entre o número de recém-nascidos e o número de cegonhas em pequenas cidades da Dinamarca,[4] nos anos 1940. A correlação entre essas duas variáveis é *espúria*: não indica *relação de causa e efeito*. Existe uma *terceira variável*, tamanho da cidade, que se correlacionava tanto com o número de recém-nascidos (quanto maiores são as cidades, mais crianças nascem) quanto com o número de casas com chaminés, perto das quais as cegonhas faziam seus ninhos.

5.4 Gráfico de linhas

Quem trabalha na área de saúde frequentemente precisa observar a *tendência* da variável, ou seja, *como* uma variável evolui ao longo do tempo. Isso pode ser feito por meio de um *gráfico de linhas*, também chamado *gráfico de série temporal*. Os dados observados referem-se à *variável resposta* e o tempo é a *variável explanatória*.

Variável resposta ou *desfecho* é a variável que estamos estudando.

Variável explanatória ou *fator* é a variável que tem efeito sobre a variável resposta ou desfecho.

Exemplo 5.8 Variável resposta e variável explanatória

A altura de uma criança varia em função da idade (tempo de vida). Então, a *variável resposta* é *altura* e a *variável explanatória* é *idade*.

Para fazer um gráfico de linhas:

1. colete valores da variável Y nos tempos que você quer estudar;
2. trace um sistema de eixos cartesianos; no eixo das abscissas, represente o tempo (X), e, no eixo das ordenadas, coloque a variável resposta Y;
3. estabeleça as escalas e faça as necessárias graduações em cada um dos eixos;
4. escreva os nomes das variáveis nos respectivos eixos;
5. desenhe um ponto para representar cada par de valores (X, Y);
6. una os pontos por segmentos de reta;
7. escreva o título.

Exemplo 5.9 Gráfico de linhas

Tabela 5.7

População residente no Brasil, segundo o ano do censo demográfico

Ano do censo	População
1940 (1)	41.236.315
1950 (1)	51.944.397
1960 (1)	70.070.457
1970	93.139.037
1980	119.002.706
1991	146.825.475
2000	169.799.170
2010	190.755.799

Nota: População presente.

Fonte: IBGE (2003).[5]

No gráfico, os pontos consecutivos ligados por linhas ajudam a visualizar as mudanças da variável no período em estudo. Assim, a Figura 5.5 mostra, nitidamente, o crescimento da população brasileira entre 1940 e 2010. Nesse período, a população mais do que quadruplicou.

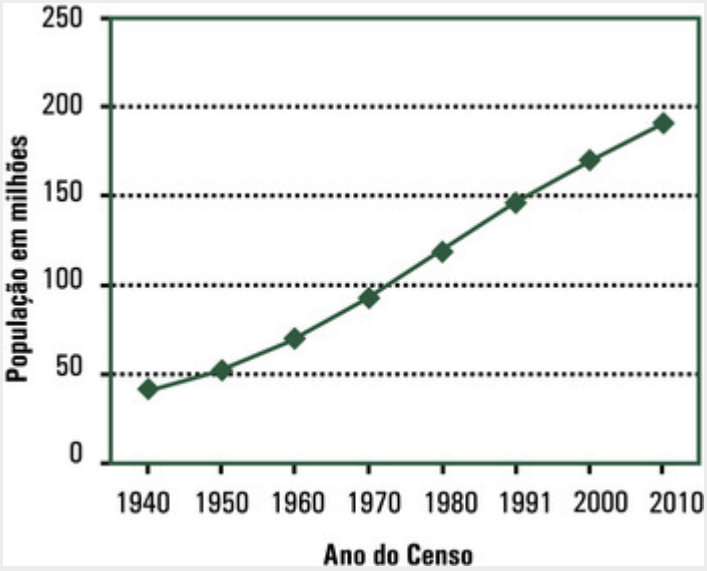


FIGURA 5.5 População do Brasil, segundo o ano do censo demográfico

5.5 Exercícios resolvidos

5.5.1. Calcule os coeficientes de correlação para cada um dos três conjuntos de dados apresentados no Exemplo 5.2.

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, $r = 1{,}000$.

5.5.2. 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 são apresentados na Tabela 5.8. Construa um diagrama de dispersão. Você acha que existe correlação entre as medidas? Em caso positivo, a correlação é linear?

Tabela 5.8

Peso seco (em miligramas) das placas dentais de dez pacientes e índice clínico

Peso seco	Índice clínico
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

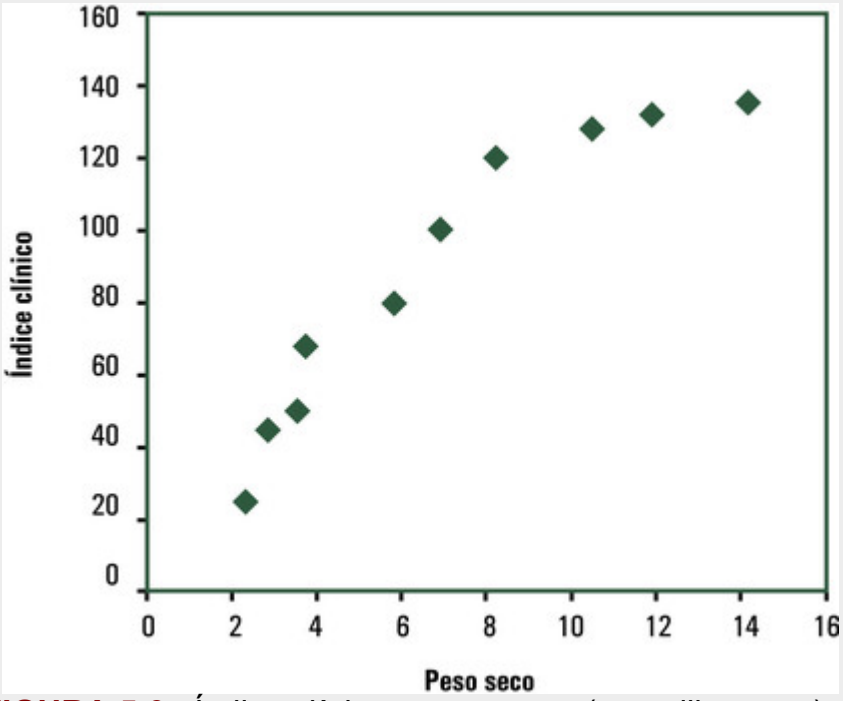


FIGURA 5.6 Índice clínico e peso seco (em miligramas) das placas dentais em dez pacientes

Existe correlação positiva entre as variáveis (duas maneiras de medir placas dentais), pois ambas crescem no mesmo sentido. Aliás, sempre se espera correlação entre duas maneiras de medir uma mesma variável. Observe que a correlação é não linear.[6]

5.5.3. Faça um diagrama de dispersão e calcule o coeficiente de correlação para os dados apresentados na Tabela 5.9. Discuta o resultado.

Tabela 5.9

Peso (em quilogramas) e comprimento (em centímetros) de sete recém-nascidos

Peso	Comprimento
3,5	51
3,7	49
3,1	48
4,2	53
2,8	48
3,5	50
3,2	49

Tabela 5.10

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
$\Sigma X = 24$	$\Sigma Y = 348$	$\Sigma X^2 = 83,52$	$\Sigma Y^2 = 17.320$	$\Sigma XY = 1.197,4$

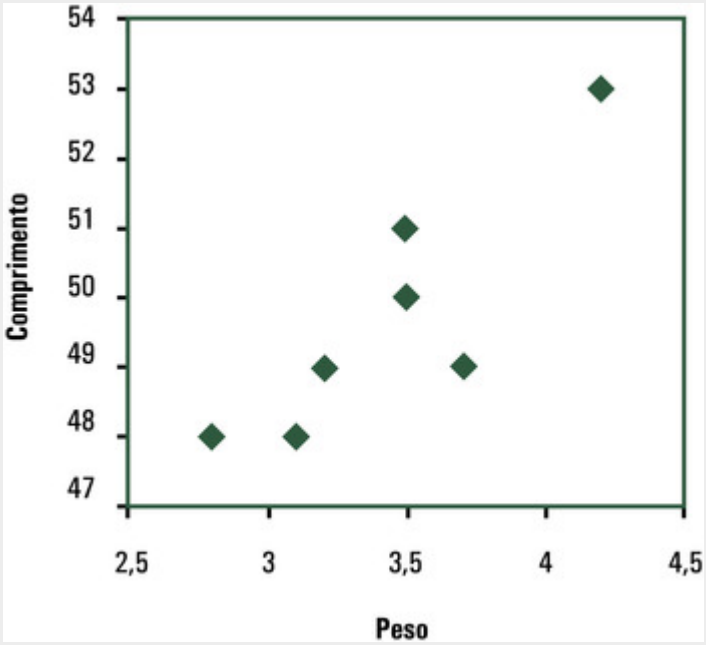


FIGURA 5.7 Peso (em quilogramas) e comprimento (em centímetros) de sete recém-nascidos

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

5.5.4. A Tabela 5.11 fornece o peso, a estatura e o IMC (índice de massa corporal) de dez 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 5.11

Peso (em quilogramas), estatura (em centímetros) e IMC de dez 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	64,0	23,51
1,72	57,5	19,44
1,73	67,0	22,39
1,74	66,0	21,80
1,79	77,0	24,03
1,80	66,0	20,37

O IMC é dado pela seguinte fórmula:

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

e indica a condição da pessoa, como segue:

IMC	Condição
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 essa 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.

5.5.5. Faça um gráfico de linhas para os dados apresentados no Exercício 5.5.2, para mostrar como o índice clínico varia em função do peso seco das placas. Discuta.

A Figura 5.8 mostra que o índice clínico usado para medir a quantidade de placa aumenta linearmente (e de forma acelerada) com o peso seco das placas, em miligramas, até cerca de 8mg. Depois, tende a estabilizar. Isso 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 das placas, que aumenta quando o acúmulo de placas é grande.

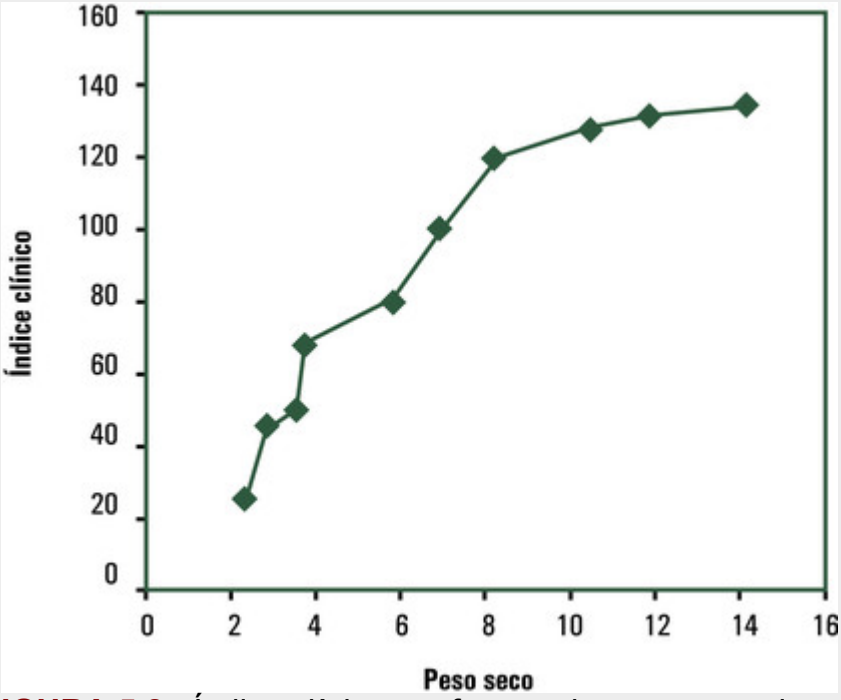


FIGURA 5.8 Índice clínico em função do peso seco das placas bacterianas

5.5.6. Reveja os dados apresentados no Exemplo 5.1, relativos ao fisioterapeuta que mediu o peso (Y) e a altura (X) de 22 universitários. O valor do coeficiente de correlação para esses

dados é $r = 0{,}747 \approx 0{,}75$. Verifique. A correlação é forte e positiva, indicando relação entre as variáveis.

[6] Existe uma explicação para o fato de a curva se estabilizar: o índice clínico mede apenas a extensão da área coberta pelas placas, e não o volume, que determina o peso.

5.6 Exercícios propostos

5.6.1. 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) $r = -0{,}90$.

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

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

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

5.6.5. Indique a afirmativa que melhor descreve os diagramas (a), (b) e o (c), apresentados na Figura 5.9.

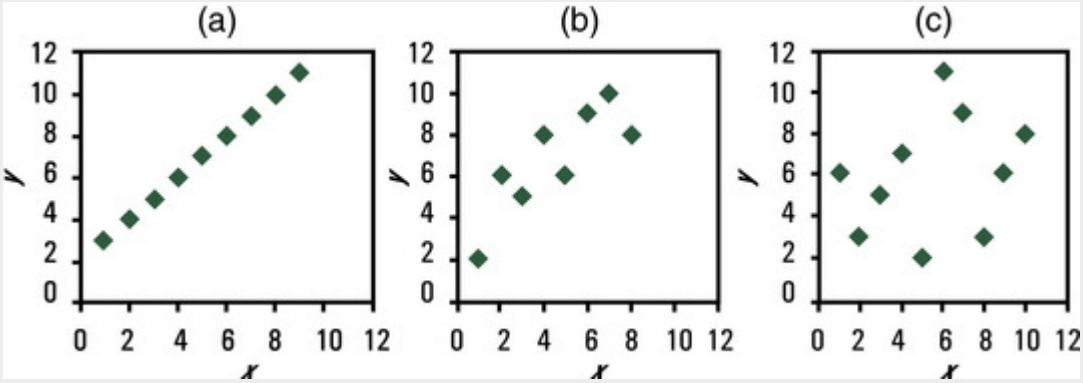


FIGURA 5.9 Diagramas de dispersão.

a) Forte correlação positiva
b) Forte correlação negativa
c) Correlação nula ou próxima de nula
d) Correlação positiva fraca

e) Correlação negativa fraca
f) Correlação perfeita positiva
g) Correlação perfeita negativa

5.6.6. 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*.

5.6.7. A correlação entre idade e expectativa de vida é:
a) positiva
b) nula
c) negativa
d) irregular

5.6.8. O diagrama de dispersão deve 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

5.6.9. Faça um diagrama de dispersão e calcule o coeficiente de correlação para os dados apresentados na Tabela 5.12. Discuta o resultado.

Tabela 5.12

Dados relativos a duas variáveis *X* e *Y*

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

5.6.10. Faça diagramas de dispersão e calcule os valores de r para os conjuntos de dados da Tabela 5.13.

Tabela 5.13

Dois conjuntos de pares de valores de duas variáveis

Conjunto A		Conjunto B	
X	Y	X	Y
1	1	1	1
2	3	1,5	2
3	6	3	3
4	5	4,5	2
5	8	5	1

5.6.11. Se todos os valores de Y forem iguais entre si, qual será o valor de r?

5.6.12. Calcule o coeficiente de correlação para os dados apresentados na Tabela 5.14.

Tabela 5.14

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

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

5.6.13. Calcule os coeficientes de correlação de Pearson para os dados dos dois conjuntos de dados apresentados na Tabela 5.15. Discuta a razão de os valores de r serem tão diferentes, embora os dados sejam tão semelhantes.

Tabela 5.15

Dois conjuntos de pares de valores de duas variáveis

Conjunto A		Conjunto B	
X	Y	X	Y
1	2	1	2
2	4	2	4
3	6	3	6
4	8	4	8
5	10	5	0

5.6.14. Suponha que foram obtidos, de pacientes com enfisema[7], o número de anos que o paciente fumou (X) e a avaliação do médico (uma nota, medida numa escala de zero a 100) sobre a diminuição da capacidade pulmonar do paciente (Y). Os resultados para dez pacientes são apresentados na Tabela 5.16. Calcule o valor do coeficiente de correlação.

Tabela 5.16

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

N.° do paciente	X	Y
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

Saiba que $\Sigma Y = 18055$; $\Sigma X^2 = 11053$; $\Sigma Y^2 = 30600$.

5.6.15. O volume máximo de oxigênio inalado (VO_2MAX) tem sido usado como medida da situação cardíaca tanto de indivíduos saudáveis como de pessoas que sofrem de doenças cardíacas. Os dados[8] de VO_2MAX em mililitros por quilograma por minuto e o tempo de exercício em minutos para 12 voluntários, homens saudáveis, depois da prática de exercícios, estão apresentados na Tabela 5.17. Desenhe um diagrama de dispersão. Olhando o diagrama, você diria que VO_2 MAX diminui quando aumenta o tempo de atividade?

Tabela 5.17

Duração do exercício (em minutos) e VO2 MAX em mililitros por quilograma por minuto para 12 homens saudáveis

Voluntário	Duração do exercício	VO_2 MAX
1	10,0	82
2	9,5	73
3	10,2	68
4	10,5	74
5	11,0	66
6	11,3	63
7	11,6	58
8	12,0	54
9	12,1	56
10	12,5	51
11	12,8	55
12	13,0	44

5.6.16. Faça um gráfico de linhas para os dados apresentados na Tabela 5.18. Discuta o resultado.

Tabela 5.18

Taxas de fecundidade total no Brasil, segundo o ano do censo

Ano do censo	Taxa de fecundidade total
1940	6,16
1950	6,21
1960	6,28
1970	5,76
1980	4,35
1991	2,89
2000	2,38
2010	1,90

[7]Ott, L e Mendenhall, W. *Understanding Statistics*. 6 ed. Belmont: Wadsworth, 1994, p. 487.

[8]Ott, L e Mendenhall, W. *Understanding Statistics*. 6 ed. Belmont: Wadsworth, 1994, p. 503.

[5]IBGE. Dados Históricos dos Censos de 1940 a 1996 – IBGE. Instituto Brasileiro de Geografia e Estatística. www.ibge.gov.br/home/estatistica/populacao/.../1940_1996.shtm. Acesso em: Abr. 2014. Resultados do Universo do Censo Demográfico 2010. www.ibge.gov.br/. Acesso em: Abr. 2014.

[1]Para estudar a correlação entre variáveis ordinais, calcula-se o coeficiente de correlação de Spearman. Ver em: Vieira, Sonia. *Bioestatística: tópicos avançados*. Rio de Janeiro: Elsevier, 2003.

[2]A regra é imprecisa, mas serve como primeira aproximação. Além disso, valores de *r* entre -0,30 e +0,30, embora possam apresentar significância estatística, não são perceptíveis nos diagramas. (Colton, T. *Statistics in Medicine*. New York: Little, Brown and Company, 1974. p. 209-11.)

[3]Ver o teste *t* no Capítulo 12.

[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 um desses anos. (Box, G. E. P., Hunter, W. G., Hunter, J. S. *Statistics for experimenters.* New York: Wiley, 1978.)

CAPÍTULO 6

Noções sobre Regressão

Como vimos no Capítulo 5, a configuração dos pontos no diagrama de dispersão pode sugerir *correlação entre duas variáveis*, mas também pode sugerir *relação linear* entre elas. Se a variação da variável resposta Y em função da variação da variável explanatória X for aproximadamente linear, é razoável buscar a equação da reta que descreve os dados.

Exemplo 6.1 Uma relação linear

Um pesquisador colocou, em oito tubos de ensaio, a mesma quantidade de plasma humano e depois reuniu, nos oito tubos, a mesma quantidade de procaína.[1] O pesquisador, então, analisou o conteúdo de cada tubo em tempos diferentes (variável X) e obteve a quantidade de procaína que já estava hidrolisada (Y) em cada um. Os dados são apresentados na Tabela 6.1 e o diagrama de dispersão, na Figura 6.1. A relação entre a quantidade de procaína hidrolisada (Y) e o tempo decorrido (X) após o início da pesquisa parece linear. Então, tem lógica traçar uma reta para mostrar como Y varia em função de X, nas condições estudadas.

Tabela 6.1

Quantidade de procaína hidrolisada (em 10 moles/litro) no plasma humano em função do tempo decorrido desde que foi colocada no tubo de ensaio contendo plasma humano (em minutos)

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

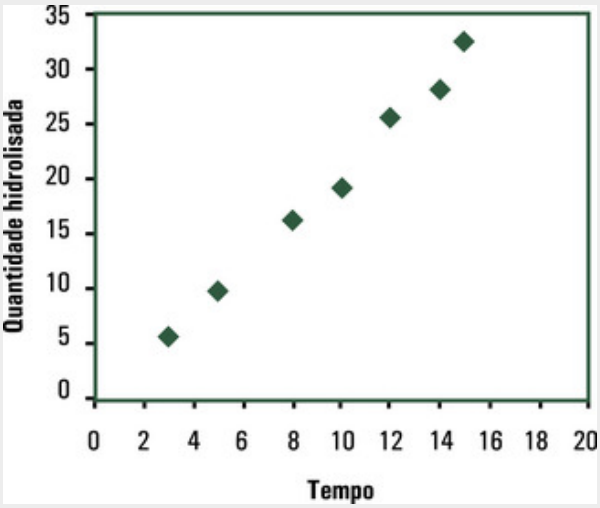


FIGURA 6.1 Quantidade de procaína hidrolisada (em 10 moles/litro) no plasma humano em função do tempo, em minutos, decorrido desde que foi colocada no tubo de ensaio contendo plasma humano

[1]Procaína é um anestésico local.

6.1 Regressão linear simples

Regressão é um termo antigo em Estatística, mas ainda usado para relatar que um modelo matemático foi ajustado aos dados para explicar a variação da variável resposta Y em função da variação da variável explanatória X.

Exemplo 6.2 Ideia de regressão

Reveja o Exemplo 6.1. Como se explica a *variação* da quantidade de procaína nos oito tubos de ensaio que continham plasma humano? Pelo passar do tempo. Veja bem: a procaína se hidrolisa no plasma humano, ou seja, a água do plasma quebra a molécula de procaína, por meio de reação química. À medida que o tempo passa, mais procaína é hidrolisada. Agora, observe a Figura 6.1: a *variação* da quantidade de procaína hidrolisada em função da *variação* do tempo decorrido desde que foi colocada no tubo de ensaio contendo plasma humano é linear.

Vamos estudar neste capítulo apenas a *regressão linear simples – linear*, porque o modelo que vamos ajustar é uma reta, e *simples*, porque há apenas uma variável explanatória. A *melhor* reta (*melhor* no sentido de que reúne as propriedades estatísticas desejáveis) recebe o nome de *reta de regressão*.[2] Nesta seção, são fornecidas as fórmulas para se obter essa reta, ou seja, para se obterem o coeficiente linear e o coeficiente angular da reta.

Equação da reta:

$$Y = a + bX$$

a: coeficiente linear
b: coeficiente angular

Vamos entender o significado desses coeficientes no sistema de eixos cartesianos. 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 6.3 Equação da reta: coeficientes lineares diferentes

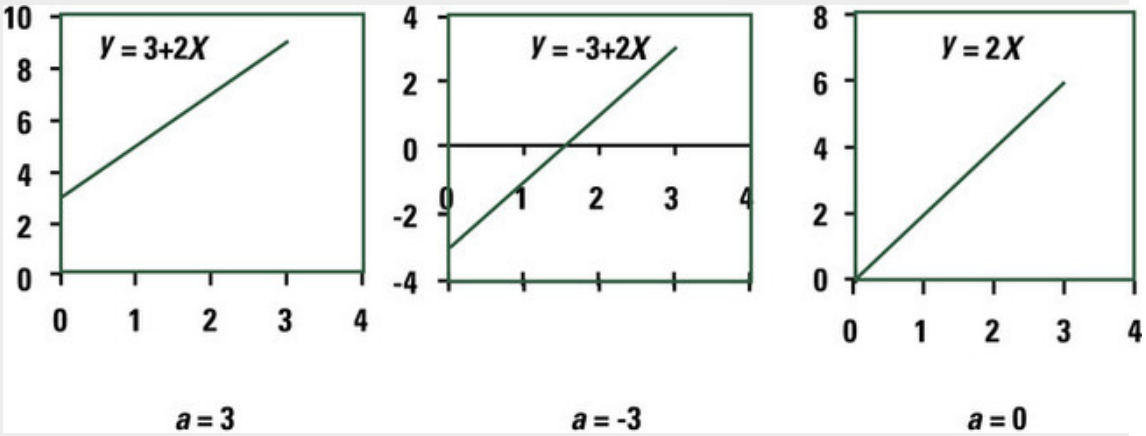


FIGURA 6.2 Apresentação gráfica de retas com diferentes coeficientes lineares

O *coeficiente angular da reta*, aqui indicado por *b*, dá a inclinação da reta.[3] Se *b* for um número:

■ *positivo*, a reta é ascendente;
■ *negativo*, a reta é descendente;
■ *zero*, a reta é paralela ao eixo das abscissas.

Exemplo 6.4 Equação da reta: coeficientes angulares diferentes

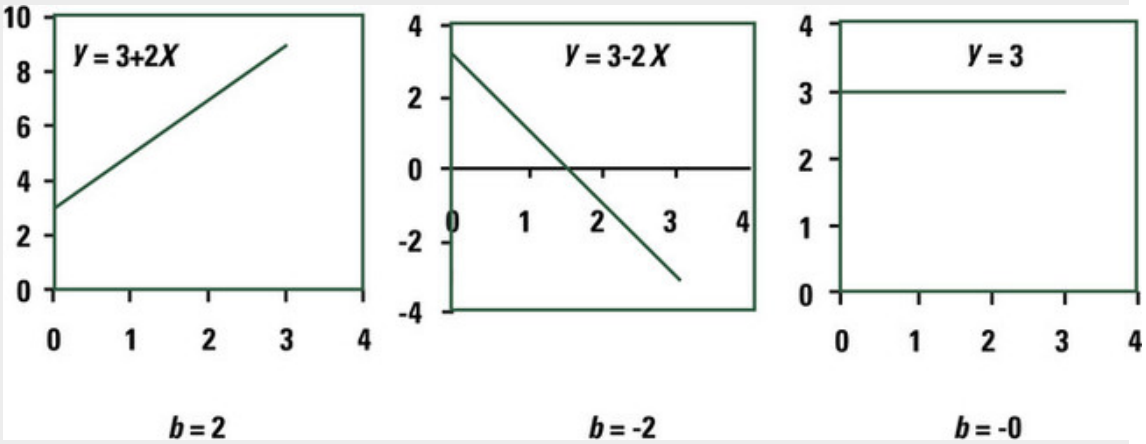


FIGURA 6.3 Apresentação gráfica de retas com diferentes coeficientes angulares

Em Estatística, o coeficiente angular da reta é obtido por meio da seguinte fórmula:

$$b = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}}$$

e o coeficiente linear é obtido por meio desta fórmula:

$$a = \overline{Y} - b\overline{X}$$

em que $\overline{Y}$ e $\overline{X}$ são as médias de Y e X, respectivamente. Veja o Exemplo 6.5.

Exemplo 6.5 Cálculo dos coeficientes de regressão

Vamos obter a reta de regressão para o problema apresentado no Exemplo 6.1.

Tabela 6.2

Cálculos intermediários para a obtenção de *a* e de *b*

X	Y	XY	X^2
2	3,5	7,0	4
3	5,7	17,1	9
5	9,9	49,5	25
8	16,3	130,4	64
10	19,3	193,0	100
12	25,7	308,4	144
14	28,2	394,8	196
15	32,6	489,0	225
69	141,2	1589,2	767

Aplicando as fórmulas, obtemos:

$$b = \frac{1589,2 - \frac{69 \times 141,2}{8}}{767 - \frac{69^2}{8}} = \frac{371,35}{171,875} = 2,16$$

$$a = \frac{141,2}{8} - 2,16 \times \frac{69}{8} = -0,98$$

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 no diagrama de dispersão. Veja a Figura 6.4.

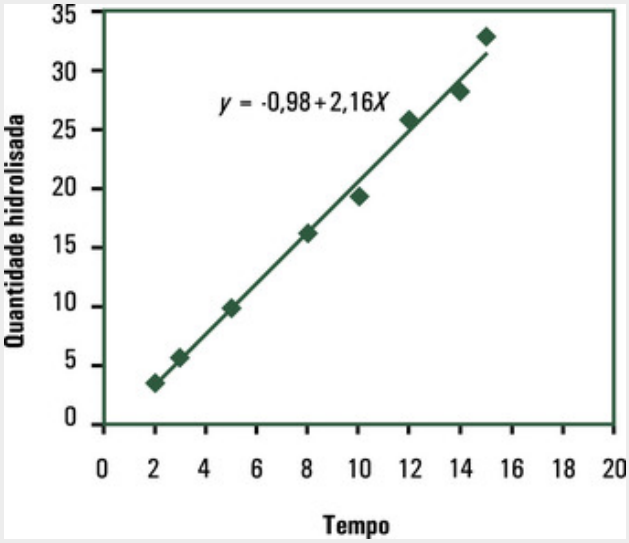


FIGURA 6.4 Reta de regressão: quantidade de procaína hidrolisada (em 10 moles/litro) no plasma humano, em função do tempo (em minutos) decorrido após sua administração

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.

Exemplo 6.6 Estimativas da variável resposta

Observe os dados apresentados na Tabela 6.1. Não existe o valor $X = 13$, mas é possível estimar o valor da variável resposta Y para $X = 13$. Basta fazer:

$$\hat{Y} = -0,98 + 2,16 \times 13 = 27,10$$

O valor $\hat{Y}$ = 27,10 é uma *estimativa*, feita com base na equação da reta de regressão, para a quantidade de procaína que deve estar hidrolisada 13 minutos após sua administração.

6.2 Extrapolação

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. A *extrapolação* pode ser incorreta ou, até mesmo, desastrosa.

É verdade que as pessoas gostariam de prever o que acontecerá em futuro próximo ou longínquo com base no que viram no passado. Mas isso nem sempre dá certo: o fenômeno pode ser modificado por fatores que não foram previstos. Toda extrapolação exige muito cuidado.

Exemplo 6.7 A extrapolação indevida

A Tabela 6.3 apresenta as temperaturas médias mensais, nos primeiros sete meses do ano, de uma cidade do sul do Brasil. Esses dados são apresentados no diagrama de dispersão da Figura 6.5. 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 é cíclico – não é linear além do período estudado.

Tabela 6.3

Temperaturas médias (em graus centígrados), 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

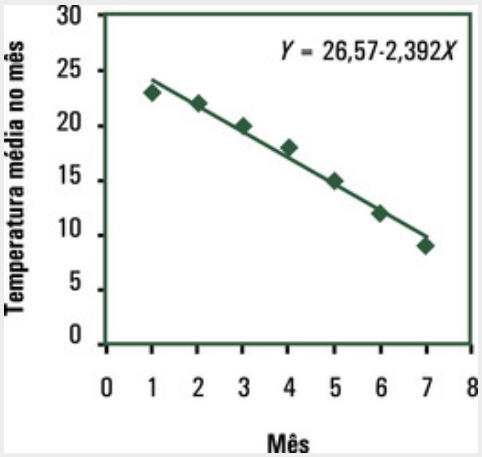


FIGURA 6.5 Reta ajustada às temperaturas médias de uma cidade do sul do Brasil, segundo o mês

6.3 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 6.1, o pesquisador fixou os tempos em que iria observar a quantidade de procaína hidrolisada, 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 da pesquisa. Nesses casos, tanto é possível ajustar a regressão de Y contra X quanto 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 da variável resposta (Y) contra a variável explanatória (X).

Exemplo 6.8 Escolha da variável explanatória

Veja os dados apresentados na Tabela 6.4. Você deve ajustar uma regressão da pressão arterial (Y) contra o peso (X), porque é o peso que pode explicar (explanar) a pressão arterial – e não o contrário.

Tabela 6.4

Pressão arterial (PA) (em milímetros de mercúrio) e peso de cães 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:

$$b = \frac{68733 - \frac{587 \times 3473}{30}}{11907 - \frac{(587)^2}{30}} = \frac{777,9667}{421,3667} = 1,846$$

$$a = \frac{3473}{30} - 1,846 \times \frac{587}{30} = 79,64$$

Então:

$$\hat{Y} = 79,64 + 1,846X$$

A reta de regressão, apresentada na Figura 6.6, 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 cão adulto em função de seu peso apresenta grande margem de erro.

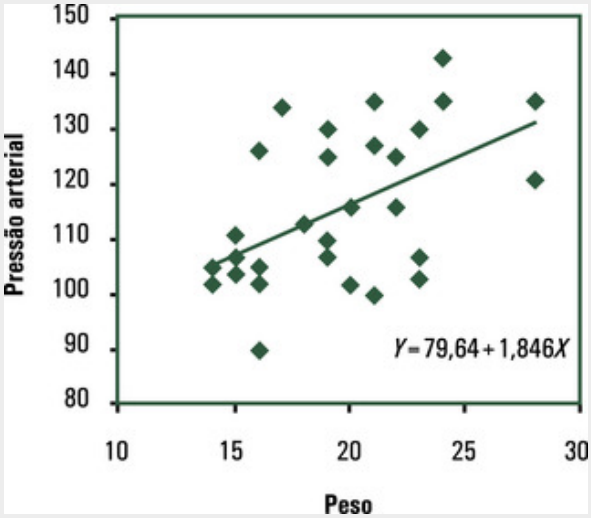


FIGURA 6.6 Reta de regressão para pressão arterial em função do peso

6.4 Coeficiente de determinação

Antes de aprendermos 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 alguém que quer diminuir o peso porque — seu médico lhe disse: obesos tendem 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, ou seja, para cada quilo a mais, haverá aumento fixo na pressão arterial? *Não* é assim. Sabe-se que existe a tendência de a pressão arterial aumentar de acordo 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, por exemplo, o de fumar e de consumir sal em excesso. 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.

Assim, 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 há erro na previsão, ou seja, 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 denominada *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 queira comprar uma camiseta para uma criança. Você chega à loja e pede ajuda à vendedora. O que ela pergunta em primeiro lugar? 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 sua idade – o que é medido pelo R^2. Portanto, saber a idade da criança ajuda na *previsão* do tamanho de sua camiseta.[4]

O *coeficiente de determinação* é a proporção da variação de Y explicada pela variação de X.

O *coeficiente de determinação* é calculado 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 6.9.

Exemplo 6.9 Coeficiente de determinação

Calcule o coeficiente de determinação para os dados apresentados na Tabela 6.1 e para os dados apresentados na Tabela 6.4. Discuta cada um deles.

Usando os cálculos intermediários já apresentados na Tabela 6.2, é possível obter R^2 = 0,994. Isso significa que 99,4% da variação da quantidade de procaína hidrolisada no plasma se explicam pelo tempo decorrido após sua administração. Em outras palavras, se você souber o tempo decorrido desde que a procaína foi colocada no plasma, poderá justificar 99,4% da variação de procaína que se hidrolisou.

Para os dados contidos na Tabela 6.4, com o auxílio de um computador (ou de seu professor), é possível obter R^2 = 0,265, um valor baixo. Se fosse alto, a explicação seria que, dado o peso de um cão, a pressão arterial seria altamente previsível. No entanto, fatores como idade, vida sedentária, hereditariedade e alimentação também são importantes.

Para ajustar uma regressão linear simples de X contra Y, é preciso que os dados dessas duas variáveis tenham sidos *obtidos de forma independente*. 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 6.9: a regressão obtida é uma *falácia* porque não se pode fazer uma regressão da diferença das variáveis contra o valor inicial.

Exemplo 6.10 Uma falácia

Observe os dados da Tabela 6.5, que estão no diagrama de dispersão da Figura 6.7: *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 6.8, com $R^2 = 0{,}582$. Só que isso *não* pode ser feito: a regressão obtida é uma falácia.

Tabela 6.5
Notas de dez alunos em duas provas

1ª prova	2ª prova	Diferença = 2ª prova – 1ª 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

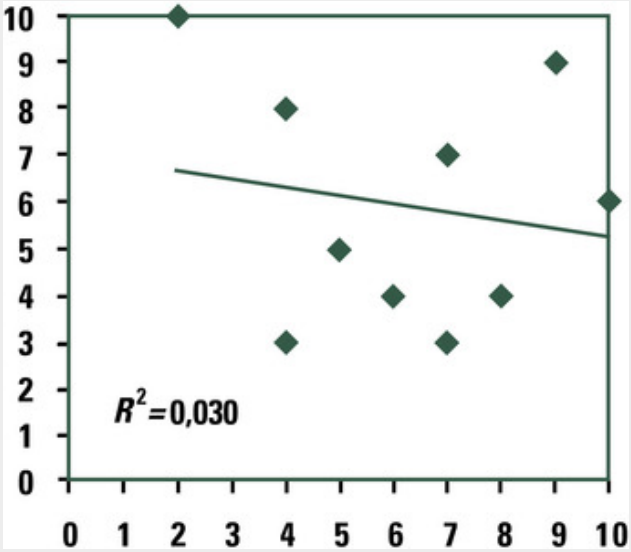


FIGURA 6.7 Nota na segunda prova em função da nota na primeira prova

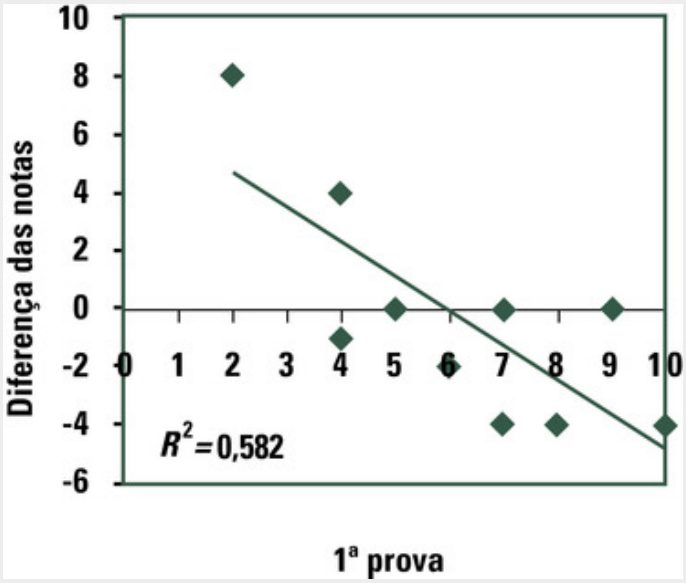


FIGURA 6.8 Diferença das notas de dez alunos em duas provas em função da primeira nota

*6.5 Regressão não linear

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 reta.[5] Veja o Exemplo 6.11.

Exemplo 6.11 Uma regressão não linear

Observe os dados da Tabela 6.6, apresentados em diagrama de dispersão na Figura 6.9: os pontos estão dispersos em torno de uma curva.

Tabela 6.6

Valores de duas variáveis *X* e *Y*

X	Y
0,0	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

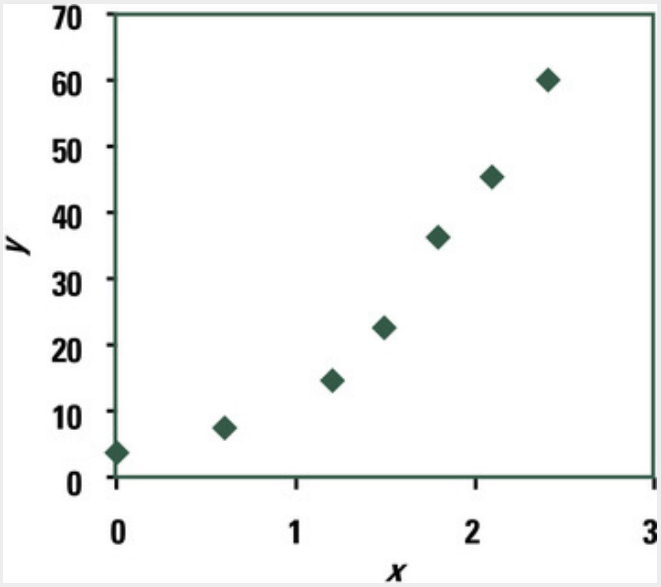


FIGURA 6.9 Diagrama de dispersão para os valores *X* e *Y* apresentados na Tabela 6.6

Quando os pontos apresentados em diagrama de dispersão *não* estão em torno de uma reta, podemos *transformar*[6] a variável *Y*. Por exemplo, é possível desenhar um diagrama de dispersão colocando, no lugar de valores de *Y*, os valores do logaritmo neperiano[7] de *Y*.

Exemplo 6.12 Transformação dos dados

Para os dados apresentados no Exemplo 6.11, os valores de *X* e dos logaritmos neperianos de *Y* estão apresentados na Tabela 6.7 e na Figura 6.10. Note que o diagrama de dispersão apresentado na Figura 6.10 mostra pontos praticamente sobre uma reta.

Tabela 6.7

Valores de *X* e dos logaritmos neperianos de *Y*

X	lnY
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

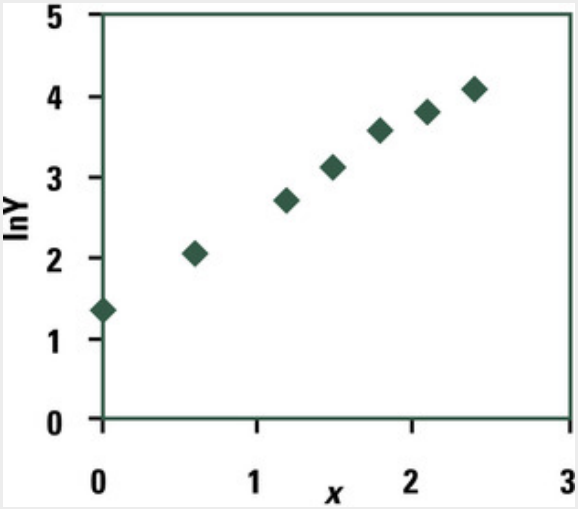


FIGURA 6.10 Diagrama de dispersão

É possível ajustar uma regressão linear de ln$Y$ contra X. Para calcular a e b, nas fórmulas dos coeficientes de regressão, usam-se os logaritmos neperianos de Y, em vez dos valores coletados de Y. Se você quiser voltar ao valor da variável Y, é preciso calcular o antilogaritmo da equação. Essa equação é chamada de *exponencial* porque traz a variável explanatória no expoente.[8]

Exemplo 6.13 Ajuste de regressão não linear

A Tabela 6.8 apresenta os cálculos intermediários para se obter a equação exponencial no Exemplo 6.11.

Tabela 6.8

Cálculos intermediários para obtenção de *a* e *b*

X	lnY	XlnY	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

$$b = \frac{33,4788 - \frac{9,6 \times 20,7940}{7}}{17,46 - \frac{9,6^2}{7}} = 1,1554$$

$$a = \frac{20,7940}{7} - 1,1554 \times \frac{9,6}{7} = 1,3861$$

A equação de reta de regressão de ln$Y$ 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. Você, então, obtém a equação *exponencial*.

$$\hat{Y} = \text{antiln}(1,3861)\ e^{1,1554X}$$

ou:

$$\hat{Y} = 3,999\ e^{1,1554X}$$

Para que uma regressão linear possa ser ajustada aos dados, muitas vezes basta transformar uma das variáveis.[9] Outras vezes, é preciso transformar ambas as variáveis.[10] Também podem ser utilizadas outras transformações, além da *transformação logarítmica*, mostrada neste capítulo. Assim, também são usadas a *extração de raiz quadrada* e a *inversão*, além de outras mais complicadas.

As transformações são, em geral, *empíricas*, ou seja, 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 na qual 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, ou seja, o *logaritmo neperiano da velocidade* e o *inverso da temperatura*. Em seguida, ajusta-se 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 $A = -Rb$.

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: existe uma só variável resposta estudada em função de uma só variável explanatória. Mas a variação da variável resposta, ou o desfecho, pode ser posta em função de diversas variáveis explanatórias. É o caso, por exemplo, da pressão arterial (desfecho), que depende não apenas do fator peso, como mostrado no exemplo, mas também de outros fatores hereditários, de alimentação, de hábitos etc. Nesses casos, ajusta-se aos dados uma *regressão múltipla*, ou seja, uma função com diversas variáveis explanatórias. Mas esse tema não será tratado neste livro.

6.6 Exercícios resolvidos

6.6.1. Ajuste uma reta de regressão aos dados apresentados no Exercício 5.5.3 (Cap. 5), para estudar peso em função do comprimento dos recém-nascidos. Calcule o coeficiente de determinação.

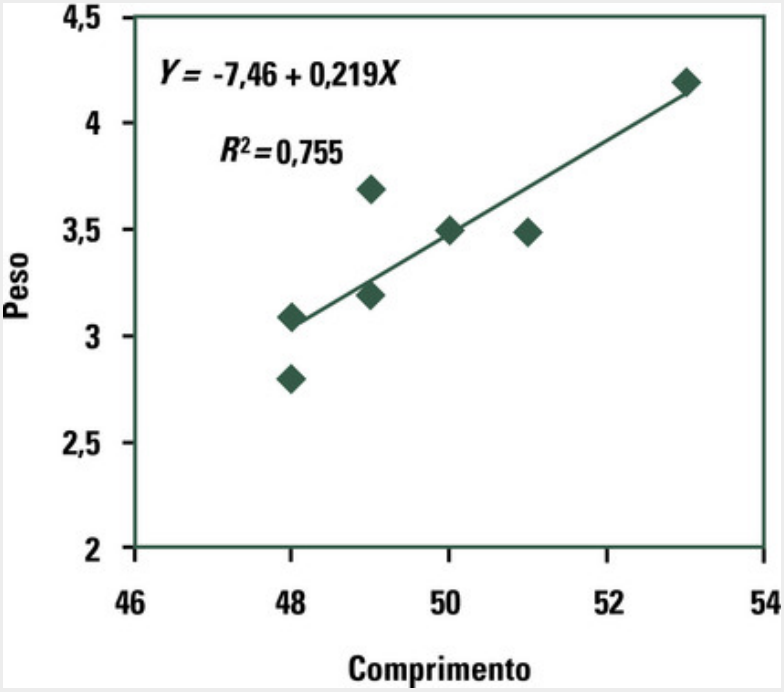


FIGURA 6.11 Reta de regressão para peso de recém-nascidos em função do comprimento

6.6.2. Ajuste uma reta de regressão aos dados apresentados no Exercício 5.5.4 (Cap. 5), para estudar peso em função de altura. Calcule o coeficiente de determinação.

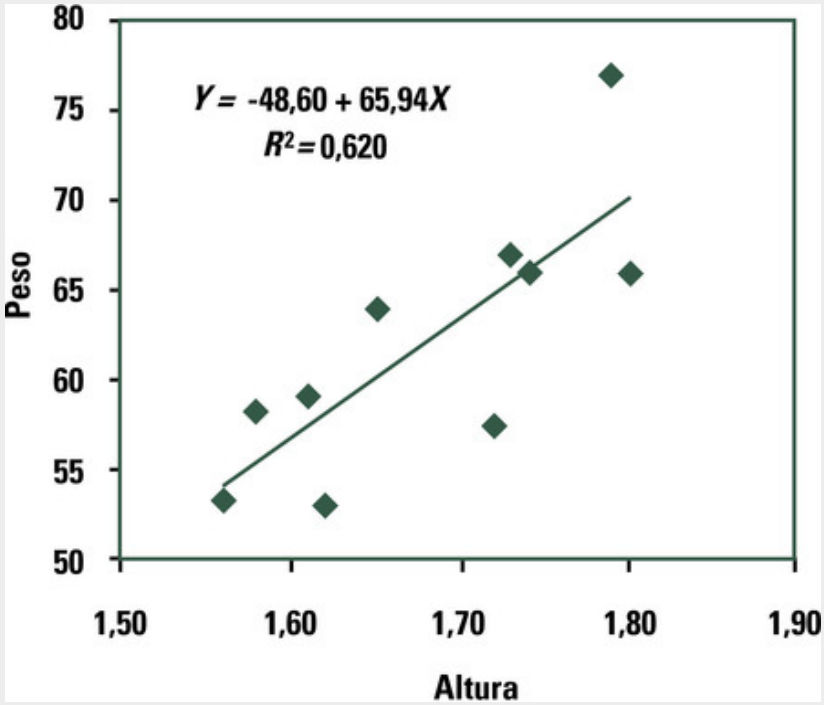


FIGURA 6.12 Reta de regressão para peso em função da altura

6.7 Exercícios propostos

6.7.1. Ajuste uma reta de regressão aos dados apresentados na Tabela 6.9.

Tabela 6.9

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

6.7.2. A reta de regressão será a mesma se você trocar X por Y? O coeficiente de correlação muda?

6.7.3. É preciso que X e Y tenham as mesmas unidades para que seja possível calcular a reta de regressão?

6.7.4. Se os filhos fossem exatamente 5 cm mais altos que seus pais, como ficaria a reta de regressão que daria a altura dos filhos em função da altura de seus pais?

6.7.5. Como seria a reta de regressão se todos os pontos de X tivessem o mesmo valor?

6.7.6. Os dados da Tabela 6.10 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 a ataques de cárie) e a média do número de anos de estudo do responsável pelas crianças. O que você acha?

Tabela 6.10

Número médio de anos de estudo do responsável pelas crianças de uma amostra e CPO-D médio

Anos de estudo do responsável	CPO-D médio
0	1,70
De 1 até 4 anos	1,85
De 5 até 8 anos	0,75
De 9 a 11 anos	0,44

6.7.7. Uma cadeia de padarias queria saber se a quantidade de dinheiro gasto em propaganda faz as vendas aumentarem. Durante seis semanas, fez, em ordem aleatória, gastos com propaganda de valores variados, conforme mostra a Tabela 6.11, e anotou os valores recebidos nas vendas. Calcule a reta de regressão e coloque em forma de gráfico. O que você acha?

Tabela 6.11

Gastos com propaganda (em reais) na semana e valores recebidos (em reais) nas vendas

Gastos	Valores recebidos
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

6.7.8. Com os dados[11] apresentados no Exercício 5.6.14 (Cap. 5), obtidos de pacientes com enfisema, calcule a reta de regressão.

6.7.9. Com os dados[12] apresentados no Exercício 5.6.15 (Cap. 5) sobre o volume máximo de oxigênio inalado (VO_2MAX), você diria que a variável diminui linearmente à medida que a atividade aumenta? Calcule a reta de regressão.

6.7.10. Os dados[13] apresentados na Tabela 6.12 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 de Y são valores da pressão sanguínea diastólica. Desenhe um diagrama de dispersão. Uma reta

de regressão explicaria a variação da pressão sanguínea diastólica em função desse tempo de repouso?

Tabela 6.12

Tempo (em minutos) desde o início do repouso e pressão sanguínea diastólica (em milímetros de mercúrio)

Tempo em minutos desde o início do repouso	Pressão sanguínea diastólica
0	72
5	66
10	70
15	64
20	66

6.7.11. Faça um diagrama de dispersão para apresentar os dados da Tabela 6.13. Calcule a reta de regressão. Coloque a reta no gráfico. Que peso médio deveriam ter dez ratos com 32 dias?

Tabela 6.13

Idade (em dias) e peso médio (em gramas) de dez ratos machos da raça Wistar

Idade	Peso médio
30	64
34	74
38	82
42	95
46	106

6.7.12. Ajuste uma equação exponencial aos dados da Tabela 6.14.

Tabela 6.14

Dados de *X* e *Y*

X	Y
28	1,25
32	1,25
35	1,75
38	2,25
39	3,25
41	3,25
42	4,25

[11]Ott, L e Mendenhall, W. *Understanding Statistics.* 6 ed. Belmont: Wadsworth, 1994, p. 487.

[12]Ott, L e Mendenhall, W. *Understanding Statistics.* 6 ed. Belmont: Wadsworth, 1994, p. 487.

[13]Schork, M. A. e Remington, R. D. *Statistics with applications to the biological and health sciences*. 3 ed. New Jersey: Prentice Hall, 2000, p. 297.

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

[3]O coeficiente angular, chamado neste livro de b, é a tangente trigonométrica do ângulo θ formado pelo eixo das abscissas e pela reta de equação $Y = a + bX$.

[4]A vendedora também pergunta se o presente é para menino ou menina. Essa informação também contribui, embora menos do que a idade, para a escolha do tamanho (na primeira infância, os meninos são maiores), mostrando-se, contudo, decisiva para a escolha do modelo.

[5]No programa Excel, você encontra as seguintes opções para ajuste de regressão: linear (que vimos até o momento), logarítmica, polinomial (que não será vista neste livro), potência, exponencial, média móvel (que não será vista neste livro).

[6]Desde que não haja razão teórica para se acreditar que a relação é obrigatoriamente linear.

[7]No Excel, procure a opção *exponencial*.

[8]O programa Excel para computadores faz essa transformação com muita facilidade.

[9]Para ajustar uma regressão *logarítmica*, transforme X, ou seja, ajuste a regressão dos logaritmos de X contra Y. Para ajustar uma regressão *potência*, transforme X e Y, ou seja, ajuste a regressão dos logaritmos de X contra os logaritmos de Y.

[10]Veja mais sobre o assunto em Vieira, Sonia. *Bioestatística: tópicos avançados*. 2 ed. Rio de Janeiro: Campus, 2003.

CAPÍTULO 7

Noções sobre Amostragem

Até o momento, vimos a *Estatística Descritiva*, que mostra como relatar os dados que temos em mãos. A interpretação do material coletado é feita por meio de gráficos e da apresentação de estatísticas como médias e desvios padrões e – se for o caso – coeficientes de correlação e reta de regressão. Então, se você medir o peso e a altura de cem crianças com 7 anos, saberá apresentar e resumir os dados, ou seja, *descrever o que encontrou nesse grupo de crianças*.

É possível *generalizar* as observações feitas nessas cem crianças (uma amostra) para todas as crianças com 7 anos da região (a população). Mas, para isso, é preciso usar um conjunto de técnicas de Estatística que permitem, com base em uma amostra, fazer *inferência* para a população de onde a amostra foi retirada. Veremos um pouco dessas técnicas nos próximos capítulos. Neste, vamos estudar população e amostra.

7.1 População e amostra

População ou universo é o conjunto de unidades sobre o qual desejamos informação.

Amostra é todo subconjunto de unidades retiradas da população para obter a informação desejada.

A chave para o bom entendimento da Estatística é saber distinguir entre os dados observados (amostra) e a vasta quantidade de dados que poderiam ter sido observados (população). O uso de amostras permite obter respostas para a questão estudada, com *margens de erro* conhecidas.

Os termos *população ou universo* não se restringem, porém, ao conjunto de pessoas, referindo-se, sim, a *qualquer conjunto grande de unidades* que têm algo em comum, como, por exemplo, radiografias feitas pelos alunos de uma faculdade em determinado curso, prontuários de pacientes atendidos pelo SUS durante todo um ano, laudos de necropsia encaminhados à Justiça por um dado serviço, auditorias das contas hospitalares de uma maternidade ou certidões de óbito registradas numa cidade em determinado período.

Também é preciso distinguir entre *população-alvo* e *população configurada.* Para isso, imagine que um instituto de pesquisa queira saber a proporção de moradores de uma cidade favoráveis à proposta do prefeito de implantar ciclovias. A *população-alvo* da pesquisa é constituída por todos os moradores da cidade. No entanto, nem toda a população-alvo estará disponível para ser amostrada: há os que não estão circulando nas ruas porque estão hospitalizados ou estão em casa cuidando de uma criança ou um doente, os muito velhos, os presidiários, os que não sabem responder, como é o caso de crianças pequenas e deficientes mentais, indecisos, pessoas que não aceitam responder etc. Logo, a *população configurada para amostragem* é necessariamente menor do que a população-alvo. Veja a Figura 7.2.

FIGURA 7.2 Configuração da amostra

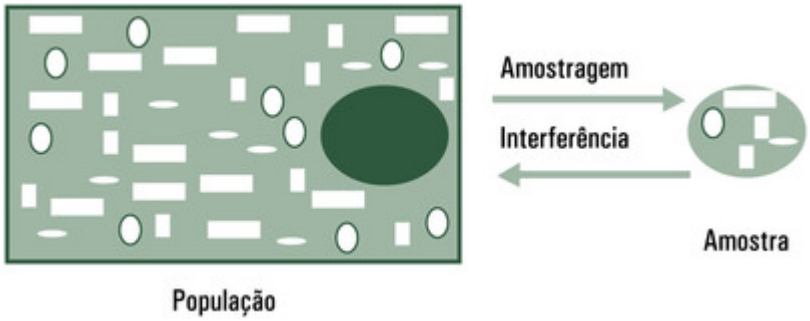


FIGURA 7.1 A ideia de amostragem

7.2 Parâmetros e estatísticas

Parâmetro é um valor em geral desconhecido (e, portanto, que precisa ser estimado) que representa determinada característica da população. Em dada população e em dado momento, o parâmetro não varia, ou seja, é um valor fixo.[1]

Estatística é uma quantidade calculada com os dados de uma amostra. É usada para estimar o parâmetro correspondente, na população de onde foi retirada.[2]

---

[1]Statistics Glossary. Disponível em *www*.stats.gla.ac.uk/steps/glossary. Acesso em 22 de janeiro de 2015.

[2]Statistics Glossary. *www*.stats.gla.ac.uk/steps/glossary. Acesso em 22 de janeiro de 2015.

É importante entender, quando se faz pesquisa por amostragem, que é possível tirar diferentes amostras da mesma população e os valores das estatísticas variarão de amostra para amostra. Por exemplo, no Brasil a média de idade dos universitários é um parâmetro. Diferentes amostras retiradas ao acaso da população de alunos darão estimativas diferentes desse parâmetro, mas todas serão *estatísticas*.

7.3 Razões para o uso de amostras

Chama-se de *censo* o levantamento de dados de toda a população. A Fundação Instituto Brasileiro de Geografia e Estatística (IBGE) faz o *Censo Demográfico do Brasil* a cada dez anos, por exigência da Constituição da República Federativa do Brasil. São coletadas, por exemplo, informações sobre sexo, idade e nível de renda de todos os residentes no país. Mas os pesquisadores da área de saúde não fazem censos, embora, às vezes, usem os dados neles coletados. As razões para se trabalhar com amostras – e não com toda a população – são poucas, mas absolutamente relevantes.

A primeira razão é a questão *do custo e da demora dos censos*. Por exemplo, qual é a média de 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 as estudar 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 – 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. Por exemplo, imagine 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. De qualquer modo, a amostra deve refletir as características da população da qual foi retirada.

7.4 Métodos de amostragem

Antes de obter uma amostra, é preciso definir quais serão os *critérios para selecionar* as unidades que a comporão. De acordo com o critério, tem-se o tipo de amostra, como apresenta o digrama da Figura 7.3.

FIGURA 7.3 Tipos de amostra

7.4.1 Amostra probabilística

A *amostra probabilística* é constituída por unidades *retiradas da população por procedimento casual ou aleatório*. Vamos definir dois tipos de amostra probabilística: a casual simples e a estratificada.

7.4.1.1 Amostra casual simples

Para obter uma *amostra casual simples*, também chamada amostra aleatória simples, confira um número a cada unidade da população e depois selecione *ao acaso* os números das unidades que irão formar a amostra. Veja a Figura 7.4, que exibe quatro pessoas selecionadas ao acaso de um conjunto de doze pessoas.

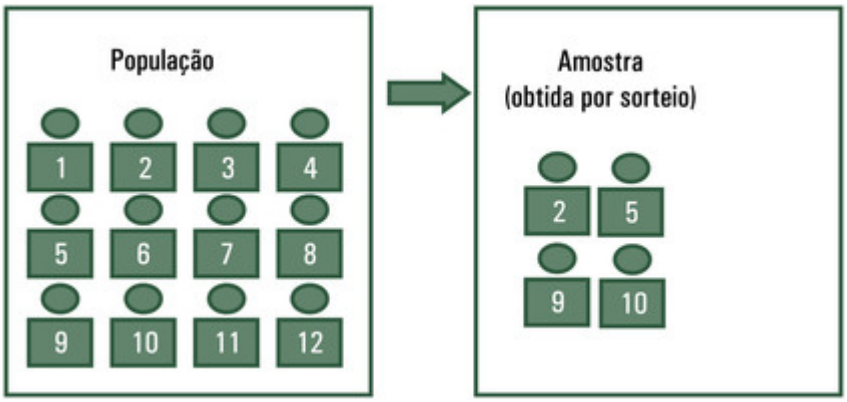


FIGURA 7.4 População e amostra casual simples

Os números das unidades que serão chamadas para a amostra devem ser obtidos por um *gerador de números aleatórios*, que é encontrado em computador.[3] É o que chamamos informalmente de sorteio. Também se faz o procedimento aleatório retirando papeizinhos de uma caixa ou bolas de uma urna (usados em programas de auditório na televisão). Para lembrar esse procedimento, veja a Figura 7.5 e o Exemplo 7.1, que ajudam a entender as regras do procedimento, que deve ser evitado porque é mais sujeito ao viés.

FIGURA 7.5 Procedimento aleatório

Exemplo 7.1 Amostra aleatória simples

Um dentista quer obter uma amostra de 2% dos quinhentos pacientes de sua clínica para entrevistá-los sobre a qualidade de atendimento da secretária. Para obter uma amostra aleatória de

2% dos quinhentos pacientes, é preciso sortear dez. Isso pode ser feito da maneira mais antiga e mais conhecida (e também mais trabalhosa): escrevem-se os nomes de todos os pacientes em pedaços de papel, colocam-se todos os pedaços de papel em uma urna, misturando-os bem, e retira-se um nome. O procedimento é repetido até serem retirados os nomes dos dez pacientes que comporão a amostra. Seria, porém, melhor que o dentista tivesse usado um gerador de números aleatórios, que pode ser encontrado em um computador.

7.4.1.2 Amostra estratificada

Se a população estiver naturalmente dividida em grupos distintos de pessoas, o pesquisador deve obter uma *amostra aleatória estratificada.* Para isso, agrupa as pessoas similares em *estratos* e obtém, de cada estrato, uma amostra casual simples proporcional ao tamanho do estrato, formando, então, uma só amostra. Veja a Figura 7.6 e o Exemplo 7.2.

FIGURA 7.6 Amostra estratificada

Exemplo 7.2 Amostra estratificada

Um dentista quer obter uma amostra de 2% dos quinhentos pacientes de uma clínica para entrevistá-los sobre a qualidade de atendimento da secretária. Ele suspeita que homens estejam sendo

mais bem atendidos do que mulheres. Aproximadamente ⅔ dos pacientes são do sexo feminino. Para obter dados de ambos os grupos, o dentista deve separar as fichas de homens e de mulheres, formando, assim, dois estratos. Em seguida, obtém uma amostra aleatória de cada estrato e reúne os dados dos dois estratos numa só amostra aleatória estratificada.

A amostra aleatória simples é, em tese, a preferida pelos estatísticos. No entanto, só a amostra estratificada garante a representação de todos os estratos (as categorias) da população na amostra coletada.

7.4.2 Amostra semiprobabilística

Para retirar da população uma *amostra semiprobabilística*, usa-se o *procedimento parcialmente aleatório*. Vamos definir três tipos de amostra probabilística: amostra sistemática, amostra por conglomerados e amostra por quotas.

7.4.2.1 Amostra sistemática

A *amostra sistemática* é constituída por unidades retiradas da população seguindo um *sistema* preestabelecido. Você ordena as unidades, numera e retira para a amostra a *k*-ésima unidade. O número *k* é obtido por sorteio. Por exemplo, se você quiser uma amostra constituída por ⅓ dos prontuários de um hospital, deve sortear um número entre 1 e 3. Se sair o número 1, selecione a primeira unidade (número 1) para a amostra. A partir de então, tome, *sistematicamente*, a primeira unidade de cada três, em sequência. No caso do exemplo, como a primeira unidade é 1, seguem, de três em três, as unidades de números 4, 7, 10 etc. Veja a Figura 7.7 e o Exemplo 7.3.

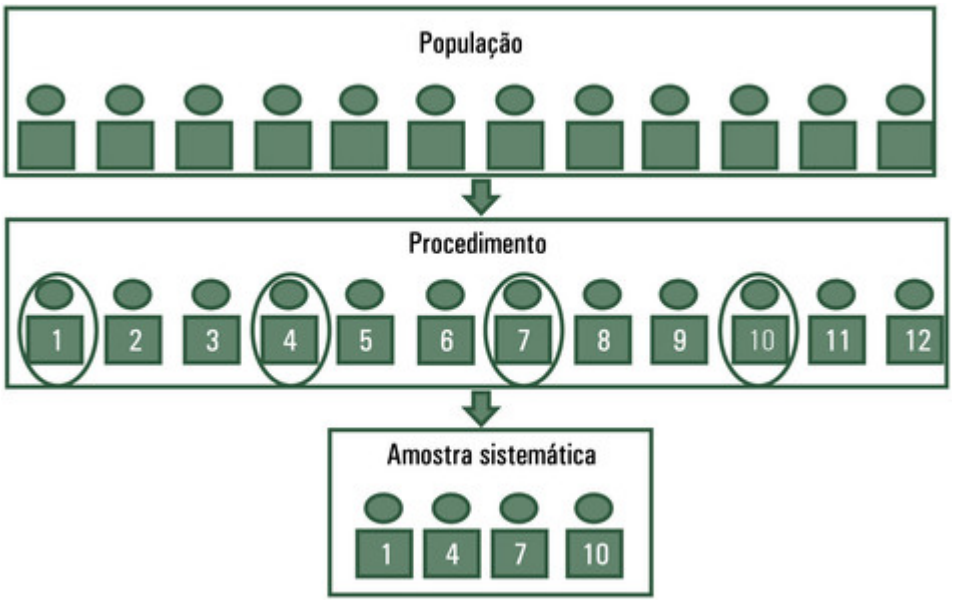


FIGURA 7.7 Amostra sistemática

Exemplo 7.3 Amostra sistemática

Imagine que você precise obter uma amostra de 2% dos quinhentos pacientes de uma clínica para entrevistá-los sobre a qualidade de atendimento da secretária. Dois por cento de quinhentos pacientes significam uma amostra de dez. Para obter essa amostra, você pode dividir 500 por 10, obtendo 50. Sorteie, então, um número entre 1 e 50, inclusive. Se sair o número 27, esse será o número do primeiro paciente a ser incluído na amostra. Depois, a partir do número 27, conte 50 e chame esse paciente. Proceda dessa forma até completar a amostra de dez pacientes.

7.4.2.2 Amostra por conglomerados

Conglomerados são grupos de unidades que já existem na população por alguma razão. Um asilo é um conglomerado de idosos; uma escola de ensino médio é um conglomerado de adolescentes; um hospital é um conglomerado de doentes. Na *amostragem por conglomerados*, um conglomerado é selecionado ao acaso da população. Veja a Figura 7.8, que mostra uma população com três conglomerados, da qual foi sorteado um, e o Exemplo 7.4.

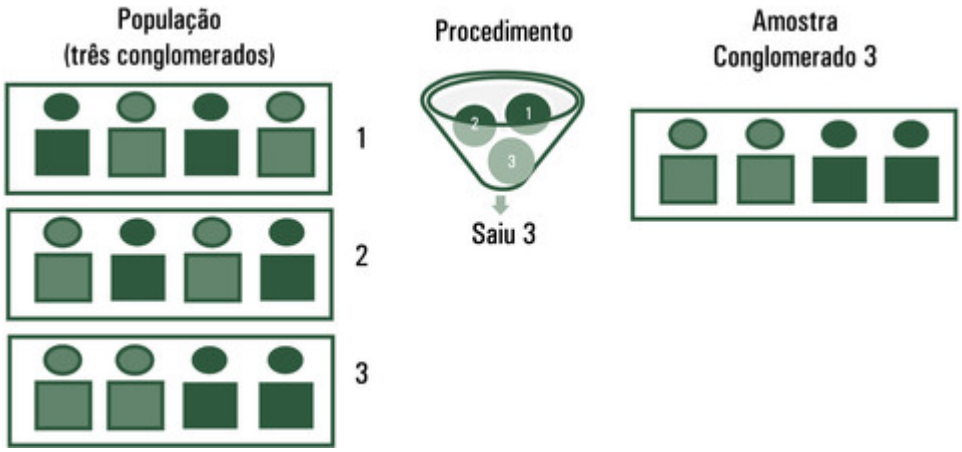


FIGURA 7.8 Amostra por conglomerados

Exemplo 7.4 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. Para obter uma amostra por conglomerados, o professor pode sortear duas academias similares (conglomerados) de ginástica da cidade, avaliar o desempenho das mulheres que frequentam essas duas academias e comparar o desempenho das que fazem com o daquelas que não fazem uso da terapia de reposição hormonal na pós-menopausa.

Não confunda *amostra aleatória estratificada* com *amostra por conglomerados*. Embora ambas envolvam grupos, são muito diferentes. Os conglomerados existem na população e, embora haja diferença dentro deles, são similares entre si, de tal maneira que cada um deles pode representar a população. Os estratos, por sua vez, são formados pelo pesquisador porque a população que examina é constituída por unidades diferentes. Então, embora haja similaridade dentro dos estratos, existem diferença entre eles.

7.4.2.3 Amostra por quotas

Na *amostragem por quotas*, as pessoas são selecionadas para a amostra porque têm uma característica bem específica. A ideia de quota é semelhante à de estrato, com uma diferença básica: a amostra estratificada é selecionada ao acaso da população, enquanto a *amostra por quotas* não é aleatória. A grande vantagem é ser relativamente barata. Por essa razão, é muito usada em levantamentos de opinião e pesquisas de mercado. Veja a Figura 7.9: ⅔ da população é negra, ⅓ da população é branca. Para constituir a amostra, percorre-se a população; tomam-se ⅔ dos primeiros negros encontrados e ⅓ dos primeiros brancos encontrados – não se faz sorteio.

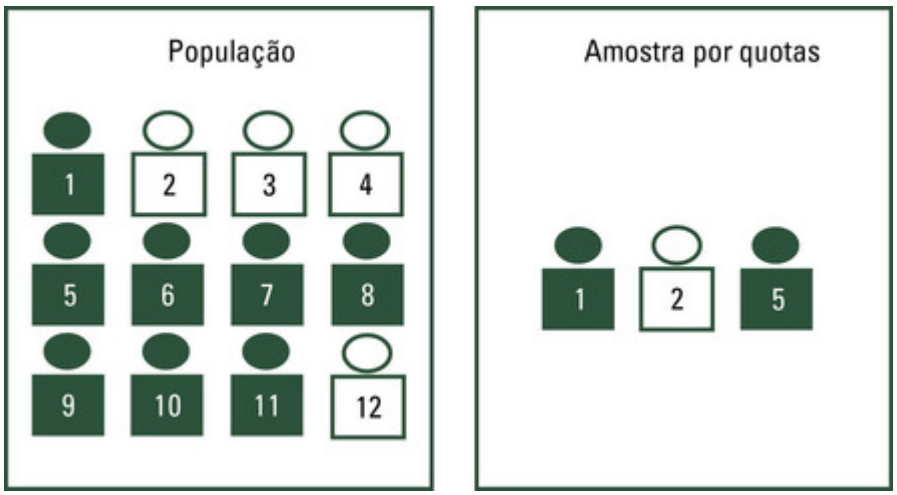


FIGURA 7.9 Amostra por quotas

Exemplo 7.5 Amostra por quotas

Considere uma pesquisa de opinião referente a serviços públicos de saúde. Como se faz uma amostra por quotas? 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. 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 trinta homens com "mais de 50 anos que recebam mais de seis e menos de dez salários mínimos". Você terá de julgar, pela aparência da pessoa, se ela se enquadra nas características descritas – homem de mais de 50 anos que ganha entre seis e dez salários mínimos. Se achar que viu a pessoa certa, deve fazer a abordagem e, em seguida, confirmar as características com perguntas. Você continua o procedimento, até preencher a quota.

7.4.3 Amostra não probabilística ou de conveniência

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

Exemplo 7.6 Amostra não probabilística

Um nutricionista quer entrevistar mães de cinquenta crianças de 3 e 4 anos, a fim de conhecer os hábitos alimentares dessas crianças. Se o nutricionista trabalha em uma escola em que estão matriculadas crianças dessa faixa etária, provavelmente procurará as mães das crianças matriculadas na escola para obter a amostra de que precisa.

Não confunda *amostra de conveniência* com *amostra por conglomerados.* Embora ambas envolvam grupos, são muito diferentes. Os conglomerados existem na população e, embora haja diferença dentro deles, *são similares entre si*, de tal maneira que cada um deles pode representar a população. Então, o pesquisador sorteia um deles. Já a amostra de conveniência é tomada pelo pesquisador porque tem acesso a essas unidades – sem considerar a falta de representatividade.

7.4.4 Avaliação das técnicas de amostragem

As *amostras aleatórias* exigem que o pesquisador tenha a listagem com todas as unidades da população, porque, dessa listagem, 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 se toma, 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 que usam algum tipo de procedimento aleatório são praticamente obrigatórias quando o objetivo da pesquisa é estimar probabilidades. É o caso das prévias eleitorais, que perguntam aos respondentes a probabilidade de voto em cenários hipotéticos de eleição. Os respondentes são escolhidos de maneira planejada, para que seja caracterizada a casualização. No caso de pesquisas de opinião, as amostras constituídas por voluntários são especialmente ruins. Tendem a responder voluntariamente a determinadas questões pessoas que são extremamente favoráveis ou contrárias à ideia apresentada.

Do ponto de vista do estatístico, as amostras probabilísticas são preferíveis, embora, na prática, nem sempre sejam possíveis. Na área de saúde, o pesquisador trabalha, *necessariamente*, com unidades às quais tem *acesso.* Nos ensaios clínicos,[4] os participantes são escolhidos de acordo com *critérios de elegibilidade*. Um pesquisador da área de saúde não pode procurar pacientes com determinada patologia e usar procedimento aleatório para trazê-los para sua clínica, por exemplo. Pode, no entanto, buscar pacientes com determinadas características (elegíveis) tratados na instituição em que trabalha. O interesse nessas pesquisas está centrado não nas estimativas de probabilidade, mas nas diferenças relativas que podem ser bem estimadas com um bom delineamento.[5]

7.5 Noções sobre o tamanho das amostras

Do ponto de vista do estatístico, as amostras devem ser grandes, para trazer maior confiança às conclusões obtidas. Para entender as razões desse ponto de vista, imagine que em uma cidade existam dois hospitais.[6] Em um deles, nascem, em média, 120 bebês por dia e, no outro, 12. A razão de meninos para meninas é, em média, 50% nos dois hospitais.

Em certa ocasião, nasceram, 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 compõem 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? Muitas vezes, 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. Mas, se seu orçamento for curto, não tente enquadrar nele uma pesquisa ambiciosa. Um pesquisador sempre precisa levar em conta o que é usual na área. Então, você tem a regra de ouro para determinar o tamanho da amostra: veja o que se faz na sua área consultando a literatura e verifique o que seu orçamento permite fazer. De qualquer forma, o certo é calcular o tamanho da amostra por critério estatístico.[7]

7.6 A questão da representatividade

A amostra só traz informações sobre a *população de onde 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. Além disso, a amostra deve ter o tamanho usual da área em que a pesquisa se enquadra. Amostras demasiadamente 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 dez vezes, *podem* ocorrer dez caras. Provável? Não. Possível? Sim.

Conclusões e decisões tomadas com base em amostras só têm sentido quando 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 é representativa da população. Você precisará 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 7.7 Amostra tendenciosa

Em 1988, Shere Hite[8] levantou, por meio de questionários inseridos em revistas femininas americanas, dados sobre a sexualidade feminina. Estima-se que cerca de 100.000 mulheres tenham sido colocadas em contato com o questionário, mas só 4.500 responderam. Mesmo assim, a amostra é grande. Mas os

estatísticos consideraram a amostra tendenciosa. O comportamento dos voluntários é diferente do comportamento dos não voluntários. Então – embora seja difícil ou até mesmo impossível estudar o comportamento de pessoas que não respondem a um questionário –, *não* se pode concluir que a amostra de respondentes represente 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 de sua vida pessoal.

[8]O exemplo é de Silver, M. *Business statistics*. Londres: McGraw, 1997.

Finalmente, algumas pessoas afirmam não acreditar em resultados obtidos de pesquisas porque elas próprias nunca foram chamadas para opinar. Se você é um daqueles 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 e você, possivelmente, também não participou das pesquisas. É verdade que ocorrem erros, é verdade que existem fraudes e é verdade também que o improvável acontece, mas daí a achar que não existem acertos vai uma enorme distância. O Brasil tem excelentes institutos de pesquisa.

7.7 Exercícios resolvidos

7.7.1. Os prontuários dos pacientes de um hospital estão organizados em um arquivo, por ordem alfabética. Qual é a maneira mais rápida de amostrar ⅛ do total de prontuários?

Seleciona-se, para a amostra, um de cada oito prontuários ordenados (por exemplo, o terceiro de cada oito, desde que três tenha sido o número escolhido por procedimento aleatório).

7.7.2. Na metade do século passado, uma colunista muito conhecida por sua seção de aconselhamento em um jornal americano perguntou a seus leitores: "se você tivesse de começar de novo, teria filhos?" Ela recebeu cerca de 10.000 respostas, cerca de 70% dizendo "Não". Você acha que as respostas foram tendenciosas?

Pessoas que escrevem para a "Seção dos Leitores" de jornais e revistas normalmente têm respostas fortes, que refletem opinião polarizada. Este exemplo mostra quanto pode ser tendenciosa uma amostra de voluntários que se dão ao trabalho de escrever a um jornal expondo uma situação pessoal de desconforto.

7.7.3. Para levantar dados sobre o número de filhos por casal em uma comunidade, um pesquisador organizou um questionário e, em seguida, enviou-o, 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 casal morador na residência. Você acha que os dados assim obtidos seriam tendenciosos?

Os dados devem ser tendenciosos porque é razoável esperar que: a) os casais com muitos filhos responderiam pensando na possibilidade de algum tipo de ajuda, como, por exemplo, instalação de uma creche no bairro; b) os casais que recentemente tiveram o primeiro filho também responderiam; c) muitos dos casais que não têm filhos não responderiam.

7.7.4. 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. Essa 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.

7.7.5. Muitas pessoas acreditam que as famílias se tornaram menores. Suponha que, para estudar essa questão, tenha sido selecionada uma amostra de 2.000 mulheres. O pesquisador, então, perguntou a elas quantos filhos tinham, quantos filhos tinham seus pais e quantos filhos tinham suas avós. O procedimento produz dados tendenciosos. Por quê?

Mulheres de gerações anteriores sem filhos não têm possibilidade de serem selecionadas para a amostra. Por outro lado, mulheres de gerações anteriores com muitos filhos terão grande probabilidade de serem amostradas.

7.7.6. Para estudar atitudes religiosas, um sociólogo sorteia dez 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! Praticamente 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.

7.7.7. 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 sua maior preocupação 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%). Além disso, tendem a responder pais que querem seus filhos na universidade e estão preocupados com os custos.

7.7.8. 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 e o envia, por e-mail, a todos os profissionais inscritos no Conselho de Odontologia. O dentista provavelmente não receberá respostas de todos. Você saberia dizer algumas das razões para 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 e-mail respondem, seja porque não têm tempo, seja porque têm preguiça ou inércia, ou ainda imaginam razões espúrias para terem sido contatadas, entre outras. 3. Não dão respostas profissionais que não contam com boa documentação de casos ou não a têm em ordem. 4. Provavelmente também não respondem profissionais que estejam enfrentando problema de ordem financeira, legal, de admissão em cursos etc.

7.7.9. 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 um conglomerado. Podem ser sorteados quatro setores. Em seguida, em cada setor, escolhe-se um ponto ao acaso e, a partir de então, 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.

7.7.10. A Tabela 7.1 apresenta os resultados parciais de um levantamento de altura e peso de brasileiros, feito pelo IBGE. Nessa tabela, são apresentados: número de participantes na pesquisa, tamanho da amostra e as medianas de altura e peso, segundo o grupo de idade. Por que não foi feito um levantamento de altura e peso de todos os brasileiros?

Tabela 7.1

Tamanho da amostra, medianas de altura e peso da população, por sexo, segundo grupos de idade. Brasil, período 2008-2009

Grupo de idade	Homens			Mulheres		
	Número	Altura	Peso	Número	Altura	Peso
20 a 24 anos	8.299	173,0	69,4	7.938	161,1	57,8
25 a 29 anos	8.084	173,0	72,7	7.945	160,7	60,5
30 a 34 anos	7.044	171,6	74,2	7.288	160,0	62,0
35 a 44 anos	12.511	171,0	74,6	13.332	159,4	63,8
45 a 54 anos	9.845	169,9	74,6	10.904	158,3	65,1
55 a 64 anos	6.585	168,2	73,1	7.545	156,6	65,3
65 a 74 anos	4.035	166,9	70,3	4.650	155,0	63,4
75 anos e mais	2.229	165,7	66,8	2.847	152,8	59,2

O levantamento de dados de toda a população (censo) é muito caro. Então, os censos são feitos de dez em dez anos. No decorrer desse período, o IBGE faz diversos levantamentos de dados, como, por exemplo, o apresentado na referida tabela.

Fonte: IBGE, Diretoria de Pesquisas, Coordenação de Trabalho e Rendimento, Pesquisa de Orçamentos Familiares 2008-2009.

7.8 Exercícios propostos

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

7.8.2. 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.

7.8.3. Dada uma população de quarenta alunos, descreva uma forma de obter uma amostra casual simples de seis alunos.

7.8.4. Organize uma lista com dez nomes de pessoas em ordem alfabética. Depois, descreva uma forma de obter uma amostra sistemática de cinco nomes.

7.8.5. Pretende-se obter uma amostra dos alunos de uma universidade para estimar o percentual deles com trabalho remunerado. a) Qual é a população em estudo? b) Qual é o parâmetro que se quer estimar? c) Você acha que seria possível obter uma boa amostra dos alunos no restaurante universitário? d) No ponto de ônibus mais próximo?

7.8.6. A maneira de fazer a pergunta pode influenciar a resposta. Basicamente, existem dois tipos de questão: 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 queira levantar dados sobre hábitos de higiene oral das pessoas de uma comunidade. Escreva, então, uma "questão fechada" e uma "questão aberta".

7.8.7. 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.

7.8.8. 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 lhe pedir para estudar a questão, como você definiria a população do estudo?

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

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

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 o cumprimento do regulamento

Cumprimento do regulamento			
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**

[3]Recomenda-se, enfaticamente, esse procedimento. O Excel pode gerar números aleatórios.

[4]Veja ensaios clínicos em Vieira, S. e Hossne, WS. *Metodologia científica para a área da saúde*. 2 ed. Rio de Janeiro: Elsevier, 2015.

[5]Piantadosi, Steven. *Clinical Trials: A Methodologic Perspective*. Nova York: Wiley, 2005.

[6]Baseado em um exemplo de Kahnemen, D. e Tvesky, A. "Judgement under uncertainty: heuristics and bias", *Science 185*, 27 de setembro de 1974.

[7]Veja, 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.

CAPÍTULO 8

Distribuição Normal

Você sabe que, no jogo de uma moeda, ou sai cara ou sai coroa, ou seja, o acaso determina o resultado. Também sabe que não é apenas nos jogos de azar que os resultados ocorrem ao acaso. Nascer menino ou menina pode ser entendido como obra do acaso. Dois irmãos, filhos dos mesmos pais, podem ter olhos de cores diferentes: um deles pode ter olhos azuis e o outro, olhos castanhos. Você tem ideia, portanto, do que é *casual* ou *aleatório*. Neste capítulo, vamos abordar a *variável casual* ou *aleatória* e sua *distribuição*. Pode parecer difícil, mas tenha em mente que, muitas vezes, a Estatística apenas formaliza o que já intuímos.

8.1 Variável aleatória

Absorver o conceito de aleatoriedade é muito mais importante do que absorver o conceito de causa e efeito, que já pertence ao nosso dia a dia.[1] O fato é que as variáveis assumem valores diferentes em diferentes unidades da mesma população.

Uma variável é aleatória quando o acaso tem influência em seus valores.

Exemplo 8.1 Uma variável aleatória

O tempo despendido para um aluno ler um livro é uma variável aleatória. Há fatores determinísticos, mas também há fatores aleatórios que afetam o tempo de leitura. De qualquer forma, se você anotar o tempo em que cada um de cem alunos lê o mesmo livro, verá grande variabilidade nos valores obtidos, porque esse tempo é uma variável aleatória.

Foi um matemático do século XIX[2] quem primeiro pensou em descrever a variabilidade das medidas biométricas e estudar sua distribuição. Para isso, fez muitas medições em nada menos do que 5.732 soldados escoceses.[3] A Tabela 8.1 apresenta a distribuição de frequências para o perímetro torácico[4] dos soldados em 16 classes, todas com amplitude de uma polegada.

Tabela 8.1

Distribuição de frequências para perímetro torácico de homens adultos, em polegadas

Classe	Perímetro torácico	Frequência	Proporção
1	34	3	0,00052
2	35	19	0,00331
3	36	81	0,01413
4	37	189	0,03297
5	38	409	0,07135
6	39	753	0,13137
7	40	1.062	0,18528
8	41	1.082	0,18876
9	42	935	0,16312
10	43	646	0,11270
11	44	313	0,05461
12	45	168	0,02931
13	46	50	0,00872
14	47	18	0,00314
15	48	3	0,00052
16	49 e mais	1	0,00017

Fonte: Daly, F.; Hand, D; Jones, C; Lunn, AD (1995).

Veja a Tabela 8.1: a proporção de soldados escoceses com 38 polegadas de perímetro torácico (ou seja, entre 37,5 e 38,5 polegadas), por exemplo, era 0,07135, ou seja, praticamente 7%. Agora, veja o histograma apresentado na Figura 8.1: na base do retângulo, é dado o intervalo de 37,5 a 38,5 polegadas; a proporção de soldados escoceses com perímetro torácico entre 37,5 e 38,5 polegadas deve ser lida no eixo das ordenadas (aproximadamente 0,07, ou 7%).

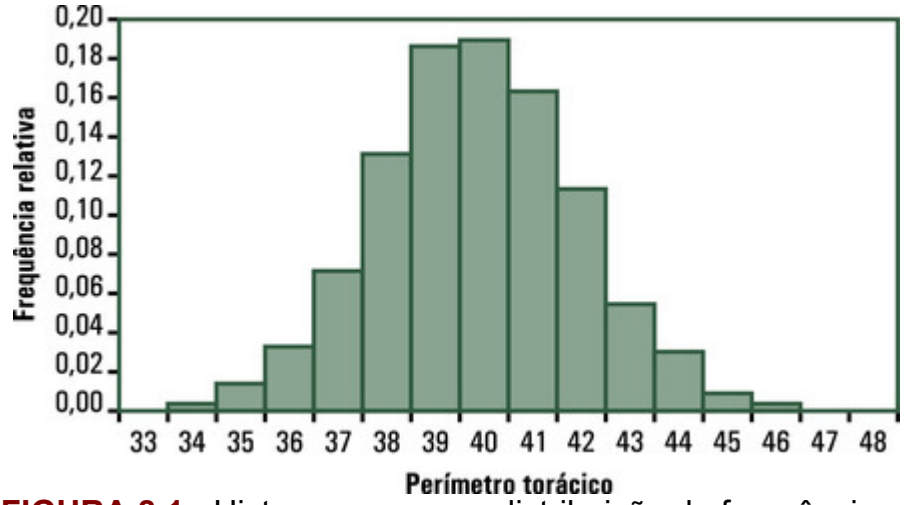


FIGURA 8.1 Histograma para a distribuição de frequências do perímetro torácico de homens adultos, em polegadas

Toda distribuição de frequências é construída com os dados de uma amostra. Se a variável é contínua – como peso ao nascer, quantidade de glicose no sangue, pressão intraocular, comprimento do fêmur –, os histogramas têm, na maioria das vezes, a aparência da Figura 8.1. Eles se assemelham à *distribuição normal*, uma distribuição teórica apresentada em gráfico na Figura 8.2.

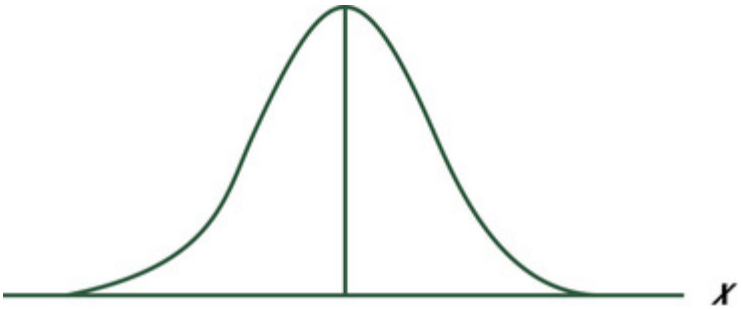


FIGURA 8.2 Gráfico da distribuição normal

Observe agora a Figura 8.3: fica fácil ver que o histograma apresentado na Figura 8.1 tem configuração semelhante à da distribuição normal da Figura 8.2. E é o fato de uma distribuição de frequências ser tão parecida com a distribuição normal que permite resolver muitos problemas de probabilidade em Estatística. Vamos, então, estudar um pouco sobre distribuição normal.

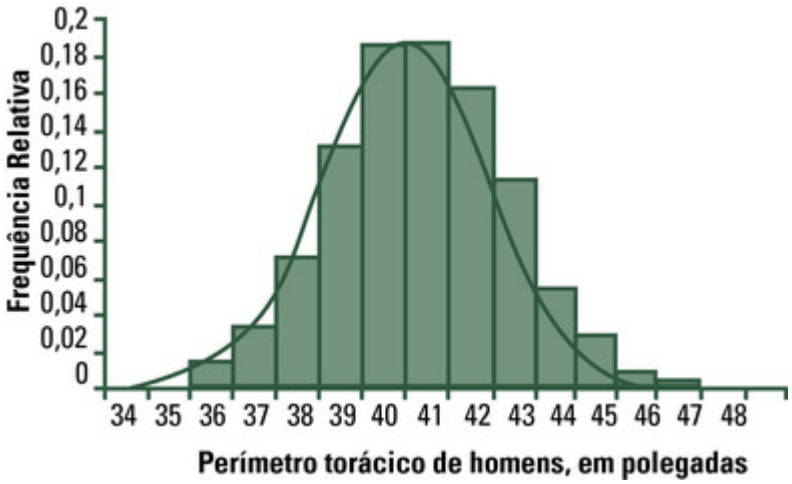


FIGURA 8.3 Gráfico da distribuição normal desenhado sobre um histograma

8.2 Distribuição normal: características

A *distribuição normal*, também chamada distribuição de Gauss, tem características bem conhecidas:

- graficamente, é uma curva em forma de sino, como mostram as Figuras 8.2 e 8.4;

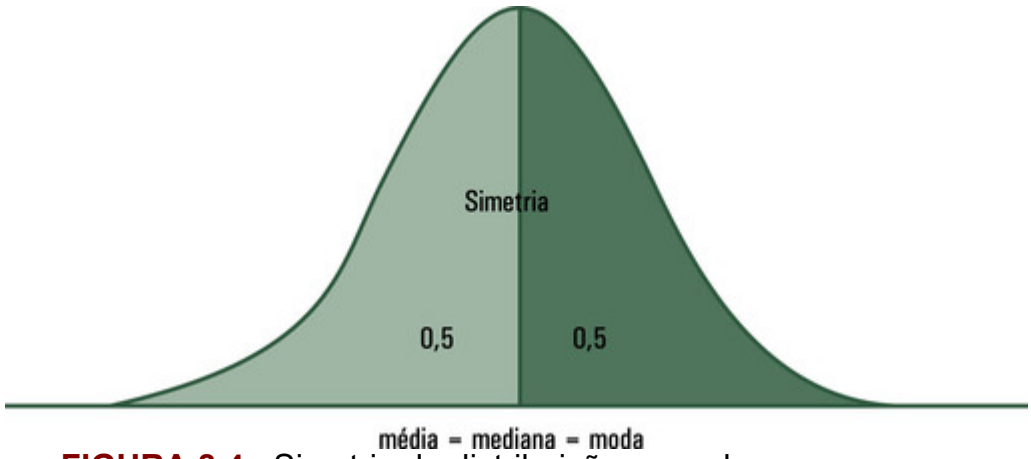


FIGURA 8.4 Simetria da distribuição normal

- a média, a mediana e a moda coincidem e estão no centro da distribuição;
- a curva é simétrica em torno da média. Logo, 50% dos valores são iguais ou maiores do que a média e 50% dos valores são iguais ou menores do que a média;
- a curva abriga 100% da população, ou seja, 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).[5]

Exemplo 8.2 Uma distribuição normal

A escala de inteligência de Weschler[6] pressupõe que inteligência é uma variável com distribuição normal de média $\mu = 100$ e desvio

padrão σ = 15. Dadas as características da distribuição normal, usando escala de inteligência de Weschler:

- metade das pessoas tem QI igual ou maior do que 100; 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 as pessoas com QI muito baixo (na cauda à esquerda da curva).

[6] Existem muitas maneiras de "medir" a inteligência (embora nenhuma delas explique, exatamente, o que está sendo medido). Mas o teste de Weschler foi idealizado pressupondo que a inteligência tem distribuição normal, como mostrado no exemplo.

8.3 Soma de variáveis aleatórias independentes

É necessário, para vários procedimentos em Estatística, pressupor que a variável em análise tem distribuição normal ou aproximadamente normal. Essa pressuposição encontra respaldo no *teorema do limite central*. Expor esse teorema está além dos limites deste livro, mas um exemplo ajuda muito.[7]

Imagine que vamos fazer 150 pães, um a um, seguindo uma receita que produz pães com 500 gramas. Por simples acaso, poderemos colocar mais, ou menos, farinha e/ou leite e/ou açúcar em alguns pães. O forno pode estar mais quente, ou menos quente, quando assarmos alguns dos pães. Pode haver um pouco mais ou um pouco menos de umidade no ar enquanto alguns pães crescem; a temperatura ambiente pode estar um pouco mais alta, ou um pouco mais baixa e assim por diante. O fato é que, no final, teremos alguns pães com mais do que 500 gramas, outros com menos e a maioria com pesos muito próximos de 500 gramas.

O *teorema do limite central* afirma que o peso de nossos pães irá variar de acordo com a distribuição normal. Por quê? Porque, sobre o peso de nossos pães, atuou grande número de variáveis aleatórias independentes – algumas atuaram para *aumentar* o peso dos pães, outras para *diminuir*. Cada variável tem efeito pequeno, mas os efeitos se somam. É pouco comum que um pão só sofra efeitos positivos ou só efeitos negativos – essas seriam as caudas da curva. A maior parte dos pães sofre efeitos positivos e negativos em quantidade que dá origem a uma *distribuição normal*.

As medidas biológicas sofrem o efeito de uma soma de variáveis aleatórias independentes. Cada variável afeta as medidas do que estamos estudando de uma forma – às vezes positiva (por exemplo, colocamos mais farinha no pão) ou negativa (colocamos menos farinha no pão). O efeito da soma de todas essas variáveis aleatórias (quantidade de açúcar, farinha, calor, umidade etc.) sobre o que estamos medindo (peso dos pães) produz uma distribuição normal.

É por isso que um fisioterapeuta está diante da distribuição normal quando monitora o desempenho físico de seus pacientes,

porque desempenho é uma variável aleatória que sofre o efeito de diversas variáveis, como idade, saúde geral, compreensão da situação, simpatia recíproca, ajuda familiar etc., que se somam (com sinais negativos ou positivos). Uma enfermeira também está diante da distribuição normal quando estuda o peso de recém-nascidos (uma variável aleatória que sofre o efeito de diversas outras variáveis aleatórias, como tempo de gestação, genética, saúde da mãe e do bebê, idade da mãe etc.).

8.4 Probabilidades associadas à distribuição normal

Nenhuma distribuição de dados reais tem características *idênticas* às da distribuição normal. No entanto, se você puder pressupor que a variável que estuda tem distribuição aproximadamente normal, pode considerar que os dados obedecem à chamada "regra empírica". Veja a Figura 8.5. De acordo com a "regra empírica", cerca de

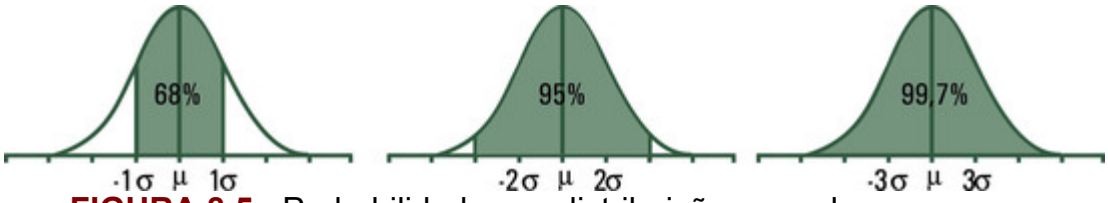


FIGURA 8.5 Probabilidades na distribuição normal: regra empírica

- 68% (pouco mais de ⅔) dos dados estarão a menos de um desvio padrão de distância da média μ;
- 95% dos dados estarão a menos de dois desvios padrões de distância da média μ;
- 99,7% dos dados estarão a menos de três desvios padrões de distância da média μ.

Mais exatamente, se a variável tem distribuição normal:

$$P\{(\mu-\sigma) \le X \le (\mu+\sigma)\} = 0,6827$$

$$P\{(\mu - 2\sigma) \leq X \leq (\mu + 2\sigma)\} = 0,9545$$

$$P\{(\mu - 3\sigma) \leq X \leq (\mu + 3\sigma)\} = 0,9973$$

Exemplo 8.3 Aplicando a regra empírica

De acordo com o teste de inteligência de Weschler, o quociente de inteligência tem distribuição normal de média $\mu = 100$ e desvio padrão $\sigma = 15$. Então, dadas as características da distribuição normal, de acordo com esse teste:

- 68% das pessoas têm quociente de inteligência entre 100 ± 15, ou seja, entre 85 e 115;
- 95% das pessoas têm quociente de inteligência entre 100 ± 2 x 15, ou seja, entre 70 e 130;
- 99,7% das pessoas têm quociente de inteligência entre 100 ± 3 x 15, ou seja, entre 55 e 145.

As probabilidades associadas às variáveis biológicas por meio da distribuição normal são apenas *aproximações*. De qualquer forma, o intervalo $\mu \pm \sigma$ abrange cerca de ⅔ da população e o intervalo $\mu \pm 2\sigma$ engloba praticamente 95% da população, ou seja, a grande maioria. Convencionou-se, assim, definir normalidade na área da saúde – quando se mede uma variável contínua – considerando "normais" todas as pessoas que têm medidas dentro do intervalo $\mu \pm \sigma$. As pessoas que têm medidas fora do intervalo $\mu + 2\sigma$ fogem do padrão de normalidade.

Exemplo 8.4 Uso da distribuição normal

Reveja a Tabela 8.1, na qual os dados estão agrupados em uma tabela de distribuição de frequências. Vamos calcular a média e o desvio padrão. A média é

$$\bar{x} = \frac{\sum xf}{\sum f} = \frac{234146}{5732} \cong 40,85$$

A variância dos dados apresentados na Tabela 8.1 é

$$s^2 = \frac{\sum x^2 f - \frac{(\sum xf)^2}{\sum f}}{\sum f - 1} = \frac{9589248 - \frac{(234146)^2}{5732}}{5732 - 1} \cong 4,2$$

Logo, o desvio padrão é

$$s \cong 2,07$$

Como foi tomada uma grande amostra (n = 5732), podemos tomar a média e o desvio padrão calculados como valores dos parâmetros μ e σ da população. Então:

$$\mu - \sigma = 40,85 - 2,07 = 38,78 \cong 39$$

$$\mu + \sigma = 40,85 + 2,07 = 42,92 \cong 43$$

$$\mu - 2 \times \sigma = 40,85 - 2 \times 2,07 = 36,71 \cong 37$$
$$\mu + 2 \times \sigma = 40,85 + 2 \times 2,07 = 44,99 \cong 45$$

Com base nesses resultados, podemos considerar que o "normal" entre soldados escoceses do século XIX era um perímetro torácico que variava entre 39 e 43 polegadas. Medidas de perímetro torácico abaixo de 37 polegadas ou acima de 45 polegadas fugiam ao padrão.

*8.5 Distribuição normal reduzida ou padronizada

Denomina-se *distribuição normal reduzida* ou *padronizada* a distribuição normal de média $\mu = 0$ e desvio padrão $\sigma = 1$. A variável com distribuição normal reduzida é comumente indicada pela letra Z. Você "transforma" um valor da variável X em Z fazendo o seguinte cálculo:

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

A variável Z é denominada *reduzida* ou *padronizada* e a transformação de X em Z é uma *redução* ou *padronização* da variável. O importante é que, na distribuição normal reduzida, valem as probabilidades dadas na Figura 8.6, que correspondem às medidas das áreas sob a curva.

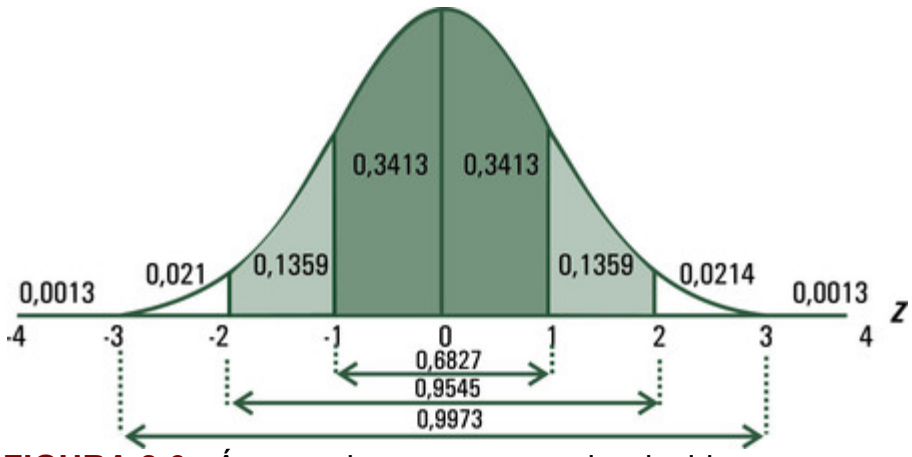


FIGURA 8.6 Áreas sob a curva normal reduzida

Além dos valores de probabilidade exibidos na Figura 8.6, é possível verificar outros valores de probabilidades, associados à distribuição normal reduzida, em tabelas já prontas. Assim, a Tabela

8.2 fornece a probabilidade de a variável normal reduzida assumir valor no intervalo entre a média (zero) e um valor qualquer de Z, até 3. Vamos, então, estudar o procedimento para encontrar probabilidades associadas a diferentes valores de Z, na Tabela 8.2.

Tabela 8.2

Tabela de distribuição normal reduzida

	Último dígito									
Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
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	0,2704	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	0,4656	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

Exemplo 8.5 Probabilidade de Z assumir um valor entre zero e 1,25

Qual é a probabilidade de a variável Z, que tem distribuição normal reduzida, assumir um valor entre zero e 1,25? Veja a Figura 8.7.

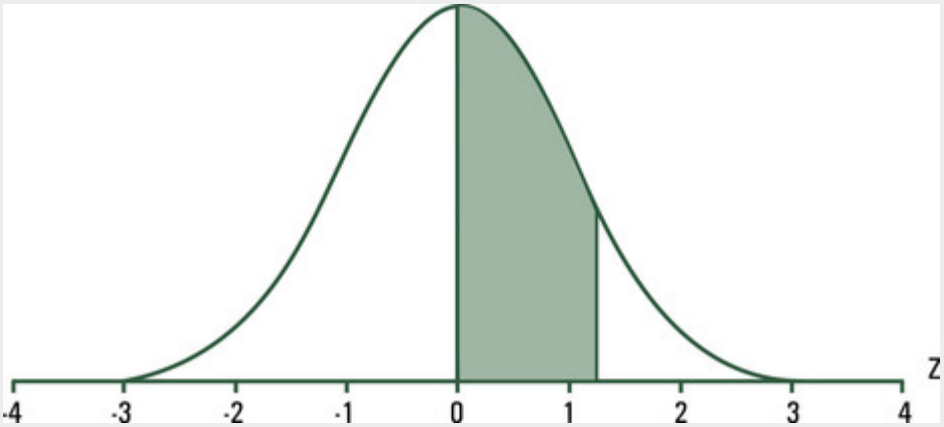


FIGURA 8.7 Probabilidade de Z assumir valor entre zero e 1,25

A probabilidade de Z assumir um valor entre zero e 1,25 corresponde à área escurecida na Figura 8.7. Essa probabilidade é encontrada na Tabela 8.2, também trazida neste livro,[8] em Anexo. Para achar a probabilidade pedida:

- na primeira *coluna* da Tabela 8.2, procure o valor 1,2 (para facilitar, esse valor está em negrito);
- encontrado o valor 1,2, siga na linha que começa com esse valor até a coluna que começa com 0,05. (Para facilitar, esse valor também está em negrito.);
- no cruzamento de 1,2 com 0,05, você encontra 0,3944 (também está em negrito);
- 0,3944 é a probabilidade de Z assumir um valor entre zero e 1,25. Escrevemos:

$$P(0 \leq Z \leq 1,25) = 0,3944$$

[8]Você encontra a tabela de distribuição normal reduzida ou padronizada na Internet, mas verifique como deve proceder para usá-la.

Exemplo 8.6 Probabilidade de *Z* assumir um valor maior que 1,25

Qual é a probabilidade de a variável *Z*, que tem distribuição normal reduzida, assumir um valor igual ou maior que 1,25? Veja a Figura 8.8.

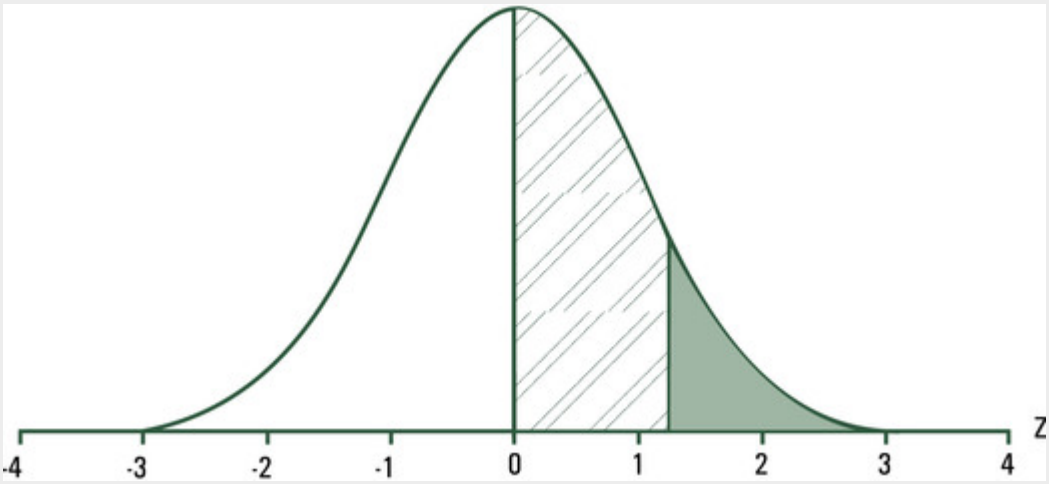


FIGURA 8.8 Probabilidade de Z assumir valor maior que 1,25

A probabilidade de *Z* assumir valor igual ou maior que 1,25 é a medida da área escurecida na Figura 8.8. Então:

- a probabilidade de ocorrer valor entre zero e 1,25, que corresponde à área com hachuras na Figura 8.8, é:
- $P\ (0 \leq Z \leq 1,25) = 0,3944$;

- a probabilidade de Z assumir valor maior ou igual à média zero é 0,5000;

$$P(Z \geq 0) = 0,5000$$

- a probabilidade de ocorrer valor maior ou igual a 1,25 (área escura na Figura 8.8) é

$$P(Z \geq 1,25) = 0,5000 - 0,3944 = 0,1056.$$

Exemplo 8.7 Probabilidade de Z assumir valor menor do que -0,51

Qual é a probabilidade de a variável Z, que tem distribuição normal reduzida, assumir valor menor do que -0,51? Veja a Figura 8.9.

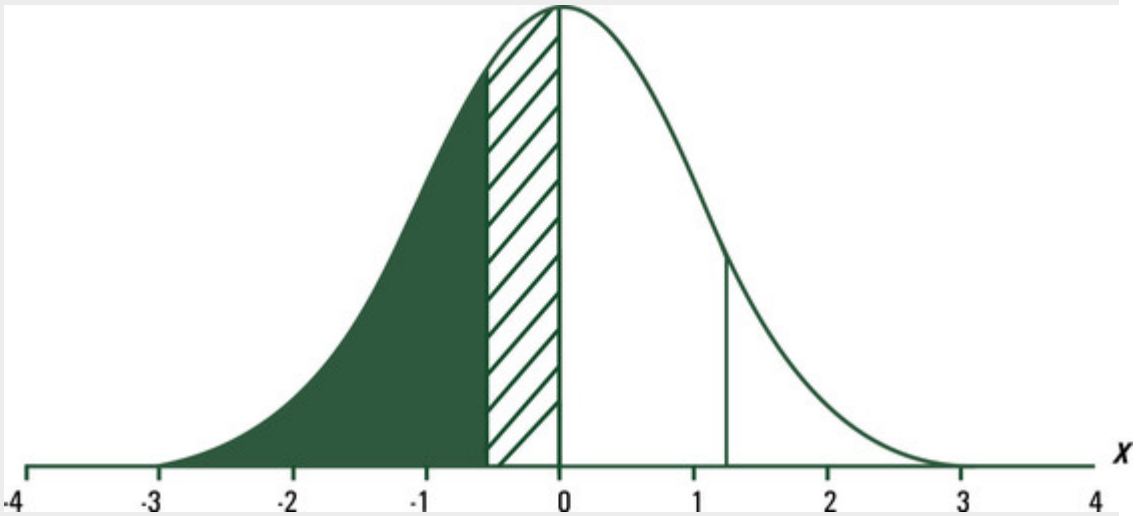


FIGURA 8.9 Probabilidade de *Z* assumir valor menor do que -0,51

Para resolver o problema, pense assim:

- a probabilidade pedida é a área escurecida da Figura 8.9;
- como a curva é simétrica, a probabilidade de ocorrer valor *igual* ou *menor do que -0,51* é igual à probabilidade de ocorrer valor *igual ou maior que 0,51*

$$P\,(Z \leq -0,51) = \mathrm{P}\,(Z \geq 0,51);$$

- a probabilidade de ocorrer valor entre zero e 0,51 é dada na Tabela 8.2; encontre a linha que começa com 0,5 e a siga, até achar a coluna que tem 0,01 no cabeçalho. No cruzamento da linha que começa com 0,5 e da coluna que começa com 0,01, está 0,1950, que corresponde à área com hachuras na Figura 8.9. Escrevemos:

$$P\,(0 \leq Z = 0,51) = P(-0,51 \leq Z \leq 0) = 0,1950;$$

- a probabilidade de ocorrer valor menor ou igual a zero (a média) é 0,5000:

$$P\,(Z \leq 0) = 0,5000;$$

- então,

$$P\,(Z \leq -0,51) = 0,5000 - 0,1950 = 0,3050.$$

Mas você pode estar se perguntando: qual é o interesse em estudar a distribuição normal reduzida – um tipo particular de distribuição? A razão é simples: para encontrar a probabilidade de uma variável com distribuição normal assumir valor em determinado intervalo, você

- *reduz* a variável;
- acha as probabilidades associadas à distribuição normal reduzida, como aprendeu aqui;
- "volta" à variável original.

*8.6 Cálculo das probabilidades sob a distribuição normal

Veja alguns exemplos de cálculo de probabilidades, pressupondo que a variável em estudo tenha distribuição normal.

Exemplo 8.8 Probabilidade – variável com 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? Veja a Figura 8.10.

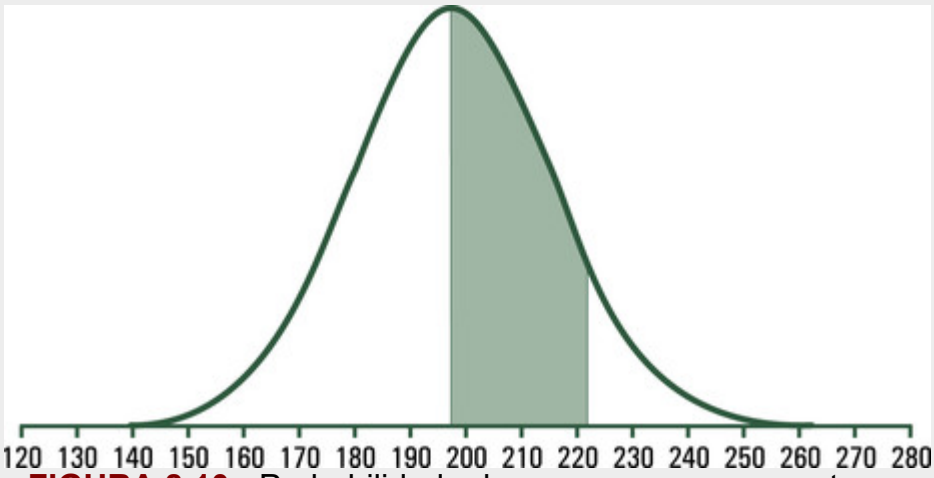


FIGURA 8.10 Probabilidade de uma pessoa apresentar entre 200 e 225 mg de colesterol por 100 mL de plasma

A probabilidade pedida corresponde à área escurecida na Figura 8.10. Para responder à pergunta, pense como segue:

- A quantidade de colesterol em 100 mL de plasma sanguíneo humano, indicada aqui por X, tem distribuição normal com média 200 mg e desvio padrão 20 mg.
- Então, a variável

$$Z = \frac{X - 200}{20}$$

tem distribuição normal reduzida. Nessa distribuição, a média é zero e ao valor $x = 225$ corresponde

$$Z = \frac{225 - 200}{20} = 1,25$$

- A probabilidade de Z assumir valor entre a média zero e $z = 1,25$ é 0,3944, como mostrado na Tabela 9.2.

$$P\,(0 \leq Z \leq 1,25) = 0,3944$$

- A probabilidade de X assumir valor entre a média $\mu = 200$ e 225 (igual à probabilidade de Z assumir valor entre a média zero e $z = 1,25$) é 0,3944.

$$P\,(200 \leq X \leq 225) = P\,(0 \leq Z \leq 1,25) = 0,3944$$

Portanto, a probabilidade de uma pessoa apresentar entre 200 e 225 mg de colesterol por 100 mL de plasma é 0,3944.

Exemplo 8.9 Probabilidade – variável com 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? Veja a Figura 8.11.

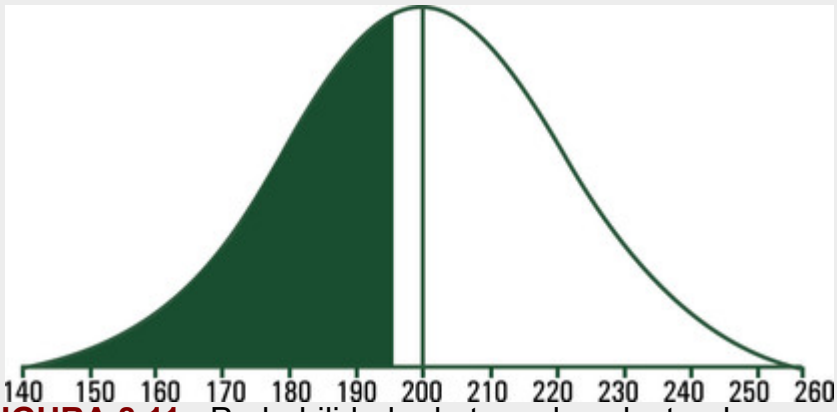


FIGURA 8.11 Probabilidade de taxa de colesterol menor do que 195 mg por 100 mL de sangue

A probabilidade pedida é mostrada pela área escurecida na Figura 8.11. Para resolver o problema:

- pressupondo que a quantidade de colesterol em 100 mL de plasma sanguíneo humano tenha distribuição aproximadamente normal com média de 200 mg e desvio padrão de 20 mg, a variável

$$z = \frac{195 - 200}{20} = -0,25$$

tem distribuição normal reduzida.

- A probabilidade de Z assumir valor menor do que -0,25 é igual à probabilidade de z assumir valor maior do que 0,25:

$$P(Z \leq -0{,}25) = P(Z \geq 0{,}25)$$

- A probabilidade de Z assumir valor entre a média zero e 0,25, dada na Tabela 8.2, é 0,0987.
- A probabilidade de Z assumir valor igual ou menor do que -0,51 é

$$P(Z \leq -0{,}25) = P(Z \geq 0{,}25) = 0{,}5 - 0{,}0987 = 0{,}4013$$

Logo, a probabilidade de uma pessoa apresentar 195 mg de colesterol por 100 mL de plasma ou menos é 0,4013 ou 40,13%.

8.7 Usos da distribuição normal

Imagine que você esteja lendo um artigo que informa que uma amostra de 4.000 jovens forneceu para pressão sistólica a média $\bar{x} = 123,4$ mmHg e desvio padrão s = 14,0 mmHg. Esses valores *estimam* a média μ e o desvio padrão σ, parâmetros da população da qual essa amostra proveio. Por que essa informação é útil?

Primeiro, é razoável assumir que a pressão sistólica tem distribuição normal. Veja o gráfico da Figura 8.9. Depois, leve em conta que você já aprendeu o seguinte:

- a probabilidade de ocorrer valor de X no intervalo $\mu \pm \sigma$ é 0,6826;
- a probabilidade de ocorrer valor de X no intervalo $\mu \pm 2\sigma$ é 0,9544.

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, tem-se que:

- a probabilidade de encontrar pessoas na população da qual a amostra proveio com pressão sistólica entre 109,4 e 137,4 mm de mercúrio é de aproximadamente (porque a distribuição é aproximadamente normal e os parâmetros estão estimados) 68,26%. Ou seja, cerca de ⅔ 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 mostra proveio com pressão sistólica entre 95,4 e 151,4 mm de mercúrio é de 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.

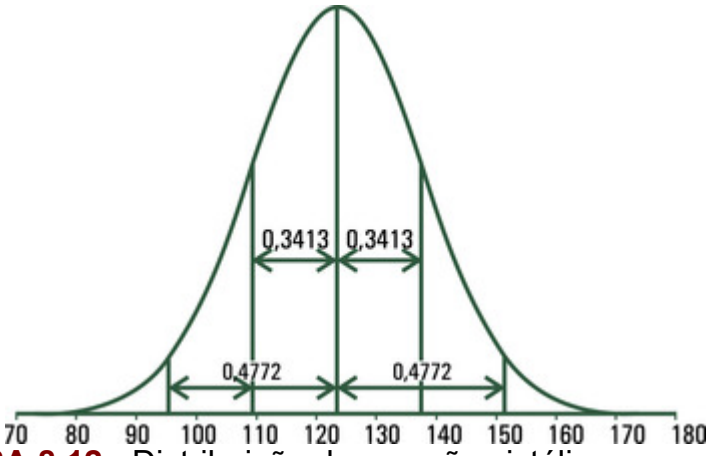


FIGURA 8.12 Distribuição da pressão sistólica

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 4.000 jovens. A média calculada foi 123,4 mmHg. Se fossem obtidas outras cinquenta amostras dessa mesma população, as médias de pressão sistólica variariam. Qual seria a distribuição dessas médias?

As médias de diferentes amostras têm distribuição normal ou aproximadamente normal, de acordo com um teorema da Estatística (o teorema do limite central). A grande aplicação dessa informação – o intervalo de confiança para uma média – será vista no Capítulo 9.

Em exames radiológicos e laboratoriais, o uso da distribuição normal é comum. Veja como isso é feito. Com base em grandes amostras, estimam-se μ e σ. Em seguida, com base na distribuição normal, definem-se critérios de normalidade e não normalidade. Por exemplo, para densidade mineral óssea (BMD, em inglês *bone mineral density*), que é medida em gramas por centímetro ao quadrado, a Organização Mundial de Saúde considera:

- *normal:* de 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 $\mu - 2{,}5\sigma$.

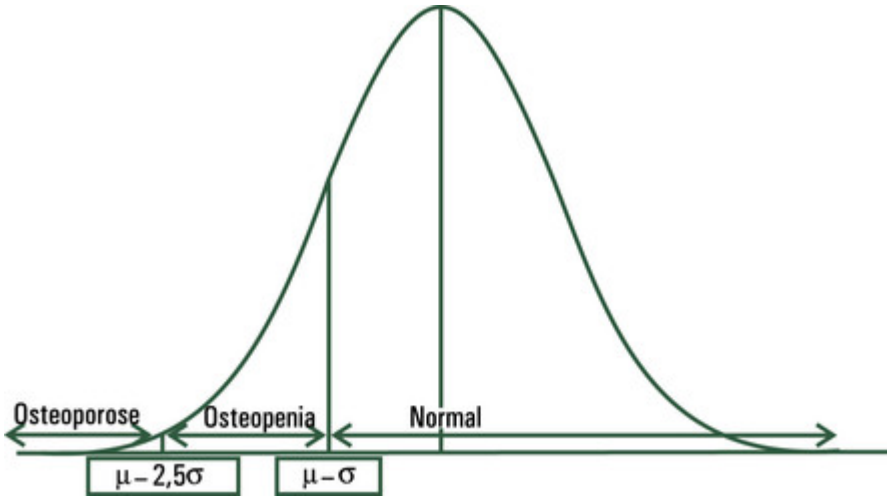


FIGURA 8.13 Distribuição de BMD

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.

8.8 Exercícios resolvidos

8.8.1. Em uma distribuição normal, qual proporção de casos cai: a) fora dos limites $X = \mu + \sigma$ e $X = \mu - \sigma$? b) fora dos limites $X = \mu + 2\sigma$ e $X = \mu - 2\sigma$?

a) Usando a regra prática: 68% (pouco mais de ⅔) dos dados estarão a menos de um desvio padrão de distância da média μ. A área sob a curva vale 100% e a curva é simétrica em torno da média. Então, 100% – 68% = 32% de casos estão fora dos limites $X = \mu \pm \sigma$. Logo, 16% dos casos estarão acima de $\mu + \sigma$ e 16% dos casos estarão abaixo de $X = \mu - \sigma$

b) Usando a regra prática: 95% dos dados estarão a menos de dois desvios padrões de distância da média μ. A área sob a curva vale 100% e a curva é simétrica em torno da média. Então 100% – 95% = 5% de casos estão fora dos limites $X = \mu \pm \sigma$. Logo, 2,5% dos casos estarão acima de $\mu + 2\sigma$ e 2,5% dos casos estarão abaixo de $X = \mu - 2\sigma$.

*8.8.2. Em homens adultos, 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 escurecida na Figura 8.14). Essa probabilidade é 0,4772, encontrada na tabela de distribuição normal reduzida. 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%.

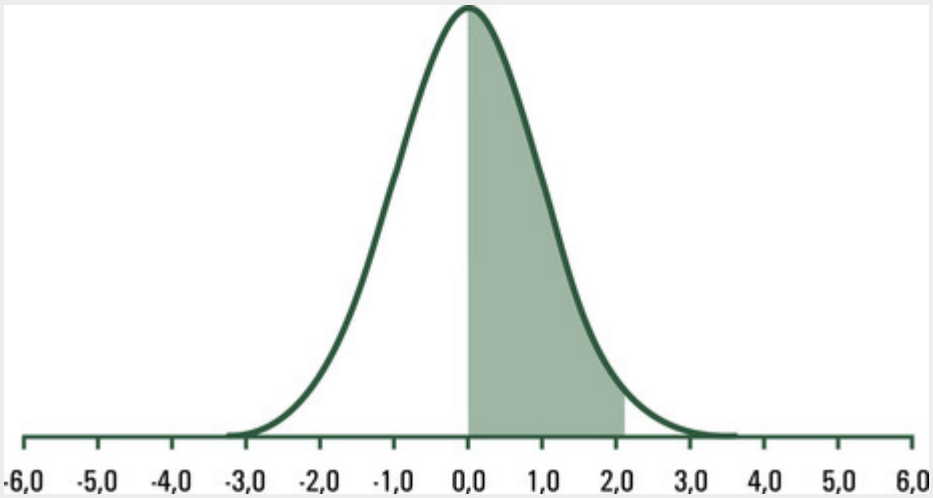


FIGURA 8.14 Probabilidade de um homem ter taxa de hemoglobina entre 16 a 18 g/dL de sangue

*8.8.3. Qual é a probabilidade de um homem ter taxa de hemoglobina maior do que 18 g/dL de sangue?

Para $x = 18$, $z = 2$; a probabilidade de Z assumir valor entre a média zero e o valor $z = 2$ é 0,4772, visto no Exercício 8.8.2. Então, a probabilidade de Z assumir valor maior que 2 é:

$$0,5 - 0,4772 = 0,0228 \text{ ou } 2,28\,\%$$

*8.8.4. 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 teste quiser que apenas 90% dos candidatos terminem o teste, quanto tempo deve dar aos candidatos para que o entreguem?

Para resolver o problema, primeiro observe a Figura 8.15. Lembre-se de 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

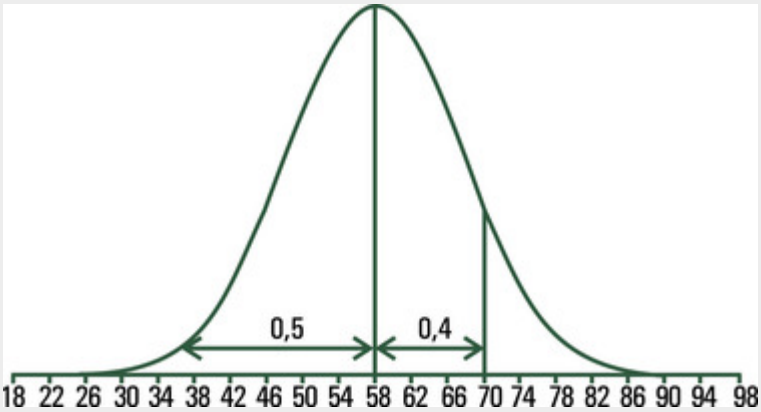


FIGURA 8.15 Distribuição do tempo despendido para completar o teste

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

$$x = \mu + z\sigma = 58 + 1,28 \times 9,5 = 70,16$$

ou seja, devem ser fixados 70,16 minutos para terminar o teste.

*8.8.5. Qual é o desvio padrão da variável aleatória X, que tem distribuição normal de média μ = 150 e 97,5% dos valores menores que 210?

A média delimita 0,5 da distribuição. Observe a Figura 8.16: é preciso encontrar 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

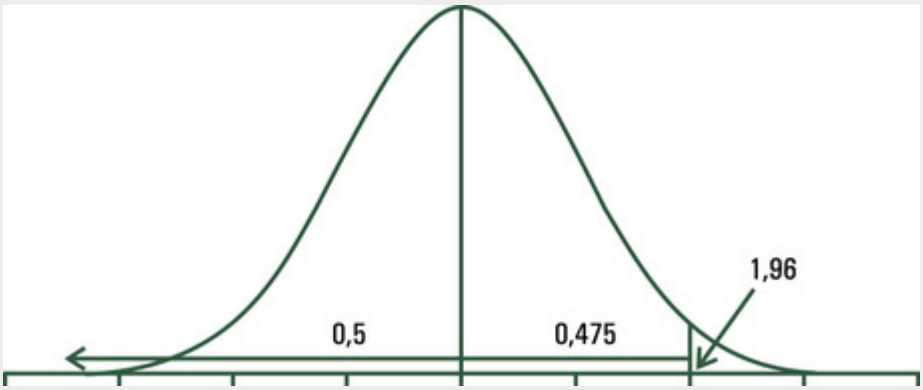


FIGURA 8.16 Distribuição da variável X

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

$$\sigma = \frac{x - \mu}{z} = \frac{210 - 150}{1,96} = 30,61$$

8.9 Exercícios propostos

8.9.1. O quociente de inteligência é uma variável aleatória com distribuição aproximadamente normal de média 100 e desvio padrão 15. Usando a regra empírica, qual é a proporção de pessoas com quociente de inteligência acima de 130?

8.9.2. 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 a concentração de sódio no plasma de uma pessoa está além do limite de normalidade?

*8.9.3. Em uma distribuição normal reduzida, quais 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?

*8.9.4. Suponha que a pressão sanguínea sistólica em indivíduos com idade entre 15 e 25 anos seja uma variável aleatória com distribuição aproximadamente normal de média $\mu = 120$ mmHg e desvio padrão $\sigma = 8$ mmHg. Nessas condições, calcule a probabilidade de um indivíduo dessa faixa etária apresentar pressão: a) entre 110 e 130 mmHg; b) maior do que 130 mmHg.

*8.9.5. A taxa de glicose no sangue humano é uma variável aleatória com distribuição aproximadamente normal de média $\mu = 100$ mg por 100 mL de sangue e desvio padrão $\sigma = 6$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 100mg por 100 mL de sangue.

*8.9.6. Em um hospital psiquiátrico, os pacientes permanecem internados, em média, cinquenta dias, com um desvio padrão de dez 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) por mais de trinta dias? b) por menos de trinta dias?

*8.9.7. A estatura de recém-nascidos do sexo masculino é uma variável aleatória com distribuição aproximadamente normal de média $\mu = 50$ cm e desvio padrão $\sigma = 2{,}50$ cm. Calcule a

probabilidade de um recém-nascido do sexo masculino ter estatura: a) inferior a 48 cm; b) superior a 52 cm.

*8.9.8. 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?

*8.9.9. Na distribuição normal reduzida, a média é sempre zero. Isso sugere que metade dos escores é positiva e metade é negativa? Explique sua resposta.

*8.9.10. 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?

---

[1]"O acaso é conceito mais fundamental que causalidade." Max Born, apud Mlodinow, L. *O andar do bêbado*. Rio de Janeiro: Zahar, 2008, p. 207.

[2]Adolphe Quetelet, 1796-1874.

[3]Os homens eram, em média, menores do que são hoje.

[4]DALY, F.; HAND, D; JONES, C; LUNN, AD. *Elements of Statistics*. Addison Wesley, 1995.

[5]Nos Capítulos 3 e 4, representamos média e desvio padrão por letras do nosso alfabeto porque estávamos nos referindo a amostras. Aqui, usamos letras gregas porque estamos nos referindo à população.

[7]Mlodinow, L. *O andar do bêbado*. Rio de Janeiro: Zahar, 2009, p. 153.

CAPÍTULO 9

Intervalo de Confiança

Muitas pesquisas são realizadas com o objetivo de estimar parâmetros. E, para estimar parâmetros, são necessários dados. Para obter dados, os pesquisadores retiram amostras da população que pretendem conhecer. Mas será que os pesquisadores podem *generalizar* a informação obtida de uma amostra (*algumas pessoas*) para a população (*todas as pessoas*)? É o que chamamos de *inferência*.

A inferência usa a informação obtida de uma amostra para estabelecer conclusões (inferência) sobre a população da qual a amostra foi retirada.

Exemplo 9.1 Inferência

Um professor de Fisioterapia obteve dados biométricos dos alunos que ingressaram na universidade. A média de altura de cem alunos do sexo masculino com 18 anos foi de 175 cm. O professor se pergunta: será que posso dizer que alunos com as características dos amostrados têm, em média, 175 cm de altura? Veja a Figura 9.1.

FIGURA 9.1 Representação da estimativa da média por ponto

A média $\bar{x}$ dos dados de uma amostra constitui *estimativa* da média μ da população (o parâmetro) da qual essa amostra foi retirada. Será que é razoável *generalizar o resultado dessa amostra* para toda a população da qual a amostra proveio? Precisamos ter uma *medida da*

incerteza associada à média da amostra. Temos apenas *uma estimativa* então precisamos conhecer as *margens de erro dessa estimativa*.

Veja o Exemplo 9.1: o professor calculou a média da amostra, mas não deu qualquer medida para informar se a média da amostra *está ou não* perto da média da população. Forneceu um só valor para descrever a amostra, ou seja, fez o que os estatísticos chamam de *estimativa por ponto*. No entanto, é possível calcular, com base em dados de amostras, *intervalos de confiança* que contêm, com certa probabilidade, a média μ da população. E como se calculam esses intervalos? Precisamos, em primeiro lugar, estimar a *variabilidade das médias das amostras*.

9.1 Erro padrão da média

Para entender a variabilidade das médias das amostras,[1] imagine uma população constituída por ⅓ de valores 4, ⅓ de valores 10 e ⅓ de valores 16, mas tão grande que, para finalidade estatística, possa ser considerada infinita. Veja:

4; 4; 4; 4; 4; .....4, 10; 10; 10; 10; 10;...10; 16; 16; 16; 16; 16;...16.

A média da população é:

$$\mu = \frac{4 \times \frac{1}{3} + 10 \times \frac{1}{3} + 16 \times \frac{1}{3}}{\frac{1}{3} + \frac{1}{3} + \frac{1}{3}} = \frac{\frac{30}{3}}{1} = 10$$

Considere, agora, as amostras de dois elementos que podem ser retiradas dessa população. O primeiro número retirado pode ser 4, ou 10 ou 16. O segundo número retirado também pode ser 4, ou 10 ou 16. As amostras possíveis (levando em conta os diferentes arranjos de dados) estão apresentadas na Tabela 9.1, com as respectivas médias e variâncias. Veja que:

Tabela 9.1

Médias das amostras de dois elementos que podem ser obtidas da população constituída por números 4, 10 e 16

	Amostras possíveis	Média
1° retirado	4	4
2° retirado	4	10
Média	4	7
Variância	0	18

- as médias 4 e 16 ocorrem com probabilidade 1/9;
- as médias 7 e 13 ocorrem com probabilidade 2/9;
- a média 10 ocorre com probabilidade 3/9;
- a média das médias é 10 e a média das variâncias é 24 Tabela 9.1.

As médias das amostras apresentadas na Tabela 9.1 estão dispersas em torno da média μ = 10 da população. Será que é possível medir *o grau de dispersão das médias das amostras*, que você vê na Figura 9.2, em torno da média da população?

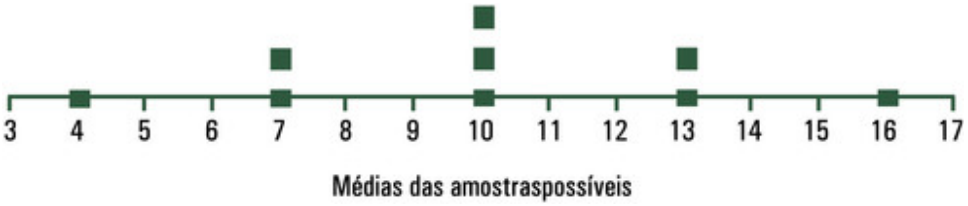


FIGURA 9.2 *Distribuição das médias das amostras* de dois elementos obtidos da população constituída por 4, 10 e 16

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 seguinte fórmula:

$$\sigma_{\bar{x}}^2 = \frac{\sum_{i=1}(\bar{x}_i - \mu)^2 f_i}{r \sum f_i}$$

em que x_i é a média da *i*-ésima amostra e *r* é o número das diferentes amostras de mesmo tamanho que podem ser obtidas da população.

Para as médias apresentadas na Tabela 9.1, a variância da média é:

$$\sigma_{\bar{x}}^2 = \frac{(4-10)^2 \times \frac{1}{9} + (7-10)^2 \times \frac{2}{9} + (10-10)^2 \times \frac{3}{9} + (13-10)^2 \times \frac{2}{9} + (16-10)^2 \times \frac{1}{9}}{9 \times \left(\frac{1}{9} + \frac{2}{9} + \frac{3}{9} + \frac{2}{9} + \frac{1}{9}\right)} = \frac{12}{1} = 12$$

Na prática, é impossível calcular a variância da média pela fórmula apresentada: o pesquisador dispõe de uma *única* amostra — e *não* de todas as amostras possíveis. Existe, porém, uma solução: *já se*

demonstrou que a estimativa da variância da média[2] é dada pela seguinte fórmula:

$$s_{\bar{x}}^2 = \frac{s^2}{n}$$

em que s^2 é a variância e n é o tamanho da amostra.

As médias, as variâncias e as variâncias das médias das amostras dadas na Tabela 9.1 estão apresentadas na Tabela 9.2. Veja que:

Tabela 9.2

Médias, variâncias e variâncias das médias das amostras apresentadas na Tabela 9.1

Estatística	Amostras possíveis									População
	1	2	3	4	5	6	7	8	9	
Média	4	7	10	7	10	13	10	13	16	$\mu = 10$
Variância	0	18	72	18	0	18	72	18	0	
Variância da média	0	9	36	9	0	9	36	9	0	

A *média das médias das amostras* é a média μ = 10 da população. A *média das variâncias das médias das amostras* é a variância das médias $\sigma_x^2 = 12$ da população.

Dizemos, então, que a média de uma amostra é uma estimativa *não tendenciosa* da média da população (todas as amostras possíveis de mesmo tamanho retiradas da mesma população dão a média da população). Da mesma forma, a variância de uma amostra é uma estimativa *não tendenciosa* da variância da população.

Uma amostra permite, ainda, *estimar a variância da média*, que, como vimos, é uma estimativa da variabilidade das médias que seriam obtidas, caso o pesquisador tivesse tomado, nas mesmas condições, todas as amostras possíveis. Podemos calcular o desvio padrão da média, mais conhecido como *erro padrão da média*, que se indica por $s_{\bar{x}}$ e é dado por:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Erro padrão da média é a raiz quadrada com sinal positivo da variância da média.

Exemplo 9.2 Estimando o erro padrão da média

Reveja o Exemplo 9.1: o pesquisador coletou uma amostra de cem alunos e calculou a média das alturas, que resultou em 175 cm. Com os dados em mãos, calculou também o desvio padrão, que resultou em s = 10 cm. A variabilidade das médias (que poderiam ser obtidas caso o pesquisador tivesse tomado todas as amostras possíveis de mesmo tamanho da população) é dada pelo erro padrão da média:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{10{,}0}{\sqrt{100}} = 1$$

O fato de a média de todas as médias das amostras possíveis terem a média μ da população é intuitivo. Mas também é fácil entender que as médias das amostras têm variabilidade *menor* do que os dados. A amostra que tiver um valor muito alto, discrepante dos demais, provavelmente terá valores menores, que farão certa compensação. Isso significa que médias de amostras de n dados têm dispersão menor do que os dados que as compõem.

9.2 Distribuição das médias das amostras

Se a variável *X* em estudo apresentar distribuição normal, as médias de amostras de *qualquer* tamanho tomadas ao acaso da população têm distribuição normal. Se a variável *X* em estudo tiver distribuição aproximadamente normal, amostras de *n =10 unidades* tomadas ao acaso da população são, em geral, suficientemente *grandes* para que as médias tenham distribuição normal.[3] No caso das variáveis biológicas como peso ao nascer, ingestão alimentar, peso corporal, ingestão calórica, taxa de colesterol, pressão arterial, para que as médias tenham distribuição aproximadamente normal, é necessário tomar amostras casuais da população com tamanho *n* variando entre *30 a 100 unidades*.

Veja bem: as médias das amostras têm distribuição normal se a variável em estudo tiver distribuição normal ou aproximadamente normal (pelo menos, não seja assimétrica) ou se as amostras forem suficientemente grandes. Entender o comportamento das médias de dados observados é, portanto, um dos pontos cruciais para quem estuda Estatística.

Quando as médias de amostras de tamanho *n* tomadas ao acaso da população têm distribuição normal com média μ e erro padrão da média $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, vale a regra apresentada em seguida, também mostrada na Figura 9.3:

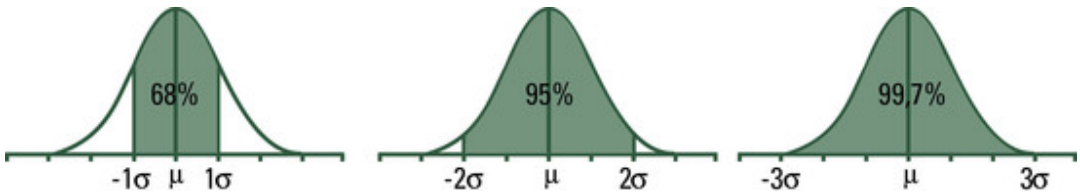


FIGURA 9.3 Probabilidades associadas à distribuição das médias

- cerca de 68% (pouco mais de ⅔) das médias de amostras de tamanho *n* tomadas ao acaso da população estarão a menos de um erro padrão de distância da média da população;

- cerca de 95% das médias de amostras de tamanho n tomadas ao acaso da população estarão a menos de dois erros padrões de distância da média da população;
- 99,7% das médias de amostras de tamanho n tomadas ao acaso da população estarão a menos de três erros padrões de distância da média da população.

Exemplo 9.3 Distribuição das médias

Reveja o Exemplo 8.2, apresentado no Capítulo 8: de acordo com o teste de inteligência de Weschler, o quociente de inteligência tem distribuição normal de média $\mu = 100$ e desvio padrão $\sigma = 15$. Então, médias de amostras de nove pessoas terão distribuição normal de média $\mu = 100$ e erro padrão da média

$$\sigma_{\bar{x}} = \frac{15}{\sqrt{9}} = 5$$

Dadas as características da distribuição normal, cerca de 95% (mais exatamente, 0,9545) das amostras de nove pessoas tomadas ao acaso da população terá média de quociente de inteligência, medida pelo teste de Weschler, no intervalo 100 ± 2 x 5, ou seja, entre 90 e 110.Veja a Figura 9.4.

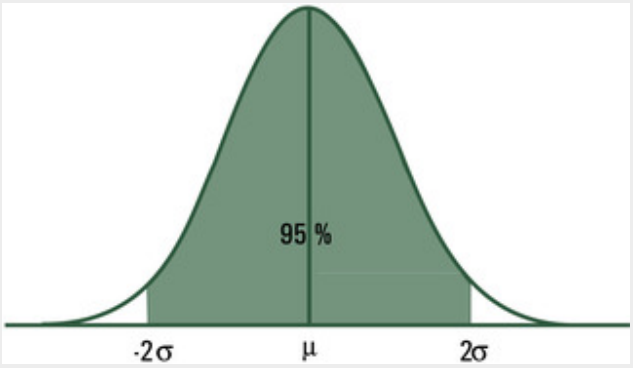


FIGURA 9.4 Distribuição das médias de quociente de inteligência em amostras de nove pessoas

Estamos considerando, neste momento, médias de amostras obtidas ao acaso de populações de variáveis que têm distribuição normal ou aproximadamente normal, como, por exemplo, peso ao nascer de filhos de mães adolescentes, ou taxa de hemoglobina no sangue, ou perda de peso no primeiro mês de uma dieta. Se X tem distribuição normal ou aproximadamente normal, mas com média e desvio padrão desconhecidos, amostras casuais de tamanho n fornecem estimativas da média, do desvio padrão e do erro padrão da média, que permitem calcular intervalos de confiança.

9.3 Cálculo do intervalo de confiança para uma média

No Capítulo 8, vimos uma amostra composta por n = 5732 soldados escoceses. Com uma amostra tão grande, pareceu razoável tomar a média e o desvio padrão calculados como μ e σ. Imagine agora que você tenha tomado uma amostra aleatória de n = 15 soldados escoceses para obter medidas de perímetro torácico. Pode, então, estimar a média, o desvio padrão e o erro padrão da média da variável estudada. Mas, com base em uma amostra pequena, é razoável considerar que $\bar{x}$ é boa estimativa de μ?

Os pesquisadores tomam *uma única amostra* e, em geral, essas amostras são *pequenas*. É, portanto, legítimo que o leitor de uma pesquisa se pergunte: posso ter *confiança* nos resultados que foram obtidos com base em uma amostra de, por exemplo, quinze ou trinta pessoas? Para dar essa confiança ao leitor, as pesquisas que fornecem médias de dados coletados por amostragem devem fornecer, também, as *margens de erro* que delimitam um *intervalo* com probabilidade (estabelecida pelo pesquisador) de conter a média μ da população. Temos, então, o que chamamos *intervalo de confiança*.

No Exemplo 9.1, o pesquisador deve relatar um *intervalo de confiança* para a média de altura de alunos do sexo masculino com 18 anos que tenham ingressado recentemente na universidade. Vamos ver então como se acha esse intervalo. Você pode calcular as *margens de erro* que dão, por exemplo, 95% de confiança de conter a verdadeira média da população por meio da seguinte expressão:

$$\bar{x} \pm t_{\alpha} \frac{s}{\sqrt{n}}$$

É bem conhecida a expressão *margens de erro*. Elas delimitam o erro da estimação. Mas, antes de entender o procedimento de cálculo, lembre-se de que n é o tamanho da amostra, $\bar{x}$ é a média e s é o desvio

padrão. O valor de t é encontrado na *Tabela de distribuição de t*, trazida neste livro nos Anexos. Vamos, então, encontrar o valor de t.

Veja a Tabela 9.3, que é uma reprodução parcial da Tabela 6 apresentada em Apêndice. Na coluna, estão os graus de liberdade, que se abrevia por gl. Para uma amostra de tamanho n, os graus de liberdade são $gl = n - 1$. Esses graus de liberdade se referem, portanto, à estimativa do desvio padrão. Se você tomou uma amostra de $n = 15$ pessoas, estimou o desvio padrão com $n - 1 = 14$ *graus de liberdade*. Procure, então, o valor 14 na primeira coluna.

Em seguida, procure na primeira linha da Tabela 9.3 o *nível de significância* indicado por α, que será definido no Capítulo 10. De qualquer forma, você já precisa saber que o nível de confiança do intervalo é dado por:

Tabela 9.3

Valores de *t* segundo os graus de liberdade e o nível de significância

Graus de liberdade	α		
	0,01	0,05	0,10
11	3,11	2,20	1,80
12	3,06	2,189	1,78
13	3,01	2,16	1,77
14	2,98	2,14	1,76
15	2,95	2,13	1,75
16	2,92	2,12	1,75

$$nível\ de\ confiança = 1 - nível\ de\ significância = 1 - \alpha$$

Em geral, os pesquisadores calculam intervalos com nível de confiança de 90, 95 ou 99%. Se você quiser um nível de 95% de confiança, como é mais usual, procure na primeira linha o valor α =

0,05 (porque 1 – 0,05 = 0,95). No cruzamento da linha que exibe 14 graus de liberdade e da coluna que exibe 0,05, você encontra $t = 2{,}14$.

Exemplo 9.4 Obtendo as margens de erro do intervalo de confiança

No Exemplo 9.1, o professor de Fisioterapia obteve a média, o desvio padrão e o erro padrão da média (Exemplo 9.2) de altura de cem alunos do sexo masculino com 18 anos que ingressaram recentemente na universidade. Para obter as margens de erro do intervalo de 95% de confiança, é preciso calcular:

$$\bar{x} \pm t_{0,05}\frac{s}{\sqrt{n}}$$

Você já tem $\bar{x} = 175, \frac{s}{\sqrt{n}} = 1$. O valor de t com $n - 1 = 99$ graus de liberdade (porque a amostra é de tamanho 100) e com o nível de confiança de 0,95 ($\alpha = 0{,}05$) é, na Tabela 6 dos Anexos, um valor entre 2,00 e 1,98. A tabela não dá o valor de t para 99 graus de liberdade. Vamos, então, tomar $t = 2{,}00$. Logo:

$$\bar{x} - t \times \frac{s}{\sqrt{n}} = 175 - 2{,}00 \times 1 = 173$$

$$\bar{x} + t \times \frac{s}{\sqrt{n}} = 175 + 2{,}00 \times 1 = 177$$

A média é 175 cm, com margens de erro de 173 e 177 cm. Veja a Figura 9.5. Escrevemos:

FIGURA 9.5 Representação da estimativa da média por intervalo

$$173 < \mu < 177$$

O intervalo de confiança fornece a amplitude dos valores que muito provavelmente incluem o verdadeiro valor do parâmetro (neste capítulo, a média μ da população). Temos, então, uma *estimativa da média por intervalo* (Fig. 9.5) que traz mais informação do que a *estimativa da média por ponto* (Fig. 9.1). Isso porque a amplitude do intervalo de confiança dá ideia de quanto de incerteza devemos associar à estimativa do parâmetro.

É importante entender o significado do intervalo de confiança para a média, que dá uma *estimativa da média por intervalo*. Em teoria, se forem tomadas sucessivas amostras e forem calculados os respectivos intervalos de 95% de confiança, 95% dos intervalos devem conter a média μ da população.

Exemplo 9.5 Cálculo do intervalo de confiança para a média

Uma amostra de trinta homens sadios com idade entre 30 e 48 anos, não fumantes e que tinham atividade física regular forneceu, em repouso, dados de pressão diastólica.[4] A média foi de 80 mm Hg,

com desvio padrão 7,1 mm Hg. Para calcular o intervalo de 95% de confiança para a média, é preciso obter

$$\overline{x} \pm t_{\alpha} \frac{s}{\sqrt{n}}$$

Dados o tamanho da amostra, a média e o desvio padrão, falta apenas o valor de $t_{0,05}$. É preciso procurar, na Tabela 6 dos Anexos, o valor de t para $n - 1 = 30 - 1 = 29$ graus de liberdade e nível de confiança de 95% ($\alpha = 0,05$). Você encontra, na mesma Tabela 6, $t = 2,04$. Então:

$$80 - 2,04 \times \frac{7,1}{\sqrt{30}} = 77,3$$

$$80 + 2,04 \times \frac{7,1}{\sqrt{30}} = 82,7$$

Podemos agora escrever o intervalo

$$77,3 \leq \mu \leq 82,7$$

[4]Com base em Brett, S. E. *et al*. Diastolic blood pressure change during exercise positively correlated with serum cholesterol and insulin resistance. *Circulation* 2000; 101:611-615.

A expressão calculada no Exemplo 9.5 aponta que, se os médicos repetirem o trabalho muitas e muitas vezes, 95 de cada cem amostras de trinta homens sadios com idade entre 30 e 48 anos não fumantes e com atividade física regular deverão conter a média de pressão diastólica da população com as características estudadas.[5]

9.4 Outras maneiras de estabelecer intervalos

Algumas revistas não aceitam resultados escritos, como, por exemplo, 19,3 ± 2,1, porque essa expressão não informa se 2,1 é o desvio padrão ou o erro padrão da média. É importante indicar como foram obtidos os limites relatados. Então, pode estar escrito, por exemplo:

$$\bar{x} \pm s = 19{,}3 \pm 2{,}1$$

Esse intervalo refere-se aos *dados – porque, na fórmula, está o desvio padrão, que mede a variabilidade dos dados*, mas não é um intervalo de confiança. Se a amostra for suficientemente grande para que se possa admitir que a média e o desvio padrão da amostra sejam boas estimativas dos parâmetros μ e σ, é razoável considerar, como vimos no Capítulo 8, que ⅔ dos dados estão no intervalo calculado.

Além disso, é comum apresentar o resultado do trabalho na forma:

$$\bar{x} \pm 2 \times s_{\bar{x}}.$$

Desde que a amostra seja suficientemente grande – mais de cem –, essa expressão pode ser vista como um intervalo de 95% de confiança para o parâmetro μ – a *média da população*, porque você está usando a fórmula do *erro padrão da média* e 2 é o valor (aproximado) de t para grandes amostras. Mas isso não é verdade no caso das pequenas amostras – de tamanho seis ou dez unidades.

9.5 Cuidados na interpretação dos intervalos de confiança

A interpretação do intervalo de confiança exige cuidado. Na prática, o pesquisador dispõe de uma única amostra que fornece uma só estimativa de determinado parâmetro. Calcula, então, um intervalo de 95% de confiança, mas *não sabe* se o parâmetro está, ou não, contido no intervalo que calculou. Sabe-se apenas que intervalos de confiança calculados da mesma forma têm 95% de probabilidade de conter 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, mas o fato de o intervalo de confiança ficar menor não significa que *contenha o parâmetro*. Conter o parâmetro é apenas uma probabilidade.

9.6 Exercícios resolvidos

9.6.1. Foram obtidos dados sobre o nível de colesterol total em jejum de 25 universitários saudáveis. A média e o desvio padrão, medidos em mg/dL, foram de 200 e 20, respectivamente. Encontre o intervalo de 90% de confiança.

Para um nível de 90% de confiança, $\alpha = 10\%$; $n - 1 = 25 - 1 = 24$. Então, o valor de t, na Tabela 6 dos Anexos, é 1,71. A expressão do intervalo de confiança fica, então, como segue:

$$\bar{x} \pm t_{\alpha}\frac{s}{\sqrt{n}} = 200 \pm 1{,}71 \times \frac{20}{\sqrt{25}}$$

$$193{,}16 \leq \mu \leq 206{,}84$$

9.6.2. Um professor obteve dados de idade de uma amostra de 61 alunos matriculados na universidade. A média de idade foi de 23,5 anos e o desvio padrão foi 3,0. Calcule o intervalo de 99% de confiança para a média.

Sabemos que as margens de erro do intervalo de confiança são dadas por

$$\bar{x} \pm t_{\alpha}\frac{s}{\sqrt{n}}.$$

Temos média de 23,5, desvio padrão 3,0, tamanho da amostra 61 e nível de confiança pedido de 99%. Para calcular o valor de t, é

preciso procurar na mesma Tabela 6 o valor que corresponde a $n - 1 = 61 - 1 = 60$ graus de liberdade e $\alpha = 100\% - 99\% = 1\%$. Você acha $t = 2{,}66$. Então:

$$\bar{x} \pm 2{,}66 \times \frac{s}{\sqrt{n}} = 23{,}5 \pm 2{,}66 \times \frac{3{,}0}{\sqrt{61}} = 23{,}5 \pm 0{,}131$$

$$23{,}5 - 0{,}131 = 23{,}369$$

$$23{,}5 + 0{,}131 = 23{,}631.$$

O intervalo de 99% de confiança para a média de idade dos alunos apresenta margens de erro 23,369 e 23,631 anos.

9.6.3. O limite inferior de um intervalo de confiança para a média para peso ao nascer pode ser negativo? Pode ser igual a zero?

Se a amostra for pequena e a variabilidade for alta, pode acontecer de o limite inferior ser zero ou até mesmo negativo, o que não tem sentido biológico. O problema é que, no cálculo do intervalo de confiança, não se leva em conta qualquer informação sobre a média da população, mas apenas os dados da amostra.

9.6.4. A pressão sanguínea sistólica medida em uma amostra de cem militares apresentou média igual a 125 mm Hg e desvio padrão igual a 9 mmHg. Calcule o erro padrão da média e ache o intervalo de 95% para a média populacional.

$$s_{\bar{x}} = \frac{9}{\sqrt{100}} = 0{,}90$$

Como no Exemplo 9.4, vamos tomar t =2,00. Então:

$$\bar{x} \pm 2 \times s_{\bar{x}} = 125 \pm 2{,}00 \times 0{,}90 = 125 \pm 1{,}80$$

O intervalo de 95% tem limites 123,20 mm Hg e 126,80 mm Hg.

9.6.5. A pressão sanguínea sistólica medida em uma amostra de nove militares apresentou média igual a 125 mm Hg e desvio padrão de 9 mmHg. Calcule o erro padrão da média e ache o intervalo de 95% para a média populacional.

$$s_{\bar{x}} = \frac{9}{\sqrt{9}} = 3{,}00$$

No nível de confiança de 95%, com n = 9-1 = 8, temos t = 2,31. Então:

$$\bar{x} \pm 2{,}31 \times s_{\bar{x}} = 125 \pm 2{,}31 \times 3{,}00 = 125 \pm 6{,}93$$

O intervalo de 95% para a variável em estudo tem limites 111,07 mm Hg e 131,93 mm Hg.

9.6.6. Compare os intervalos de confiança obtidos nos exercícios 9.6.4 e 10.6.5.

A amplitude do intervalo de confiança dá ideia de quão incertos estamos acerca do valor do parâmetro que desconhecemos. Amplitude grande pode estar indicando que a amostra deveria ser maior. Não existe efeito do tamanho da amostra sobre o

valor numérico do desvio padrão calculado. No entanto, o erro padrão da média tende a diminuir porque o valor da média da amostra tende a se aproximar do valor da média verdadeira (veja que você divide o desvio padrão por $\sqrt{n}$). O valor de t é maior quando a amostra é pequena.

9.7 Exercícios propostos

9.7.1. Um intervalo de 95% de confiança para a média tem a seguinte interpretação:
   a. se forem tomadas repetidamente muitas amostras e calculados seus intervalos de confiança, 95% devem conter a média;
   b. 95% da população está contida no intervalo de 95% de confiança.

9.7.2. Responda se a afirmativa "Intervalos de confiança só podem ser calculados para a média" é
   a. verdadeira
   b. falsa

9.7.3. Seja X a variável aleatória que representa a pressão sanguínea sistólica de indivíduos com idade entre 20 e 25 anos. Essa variável apresenta distribuição aproximadamente normal. Suponha que, com base em uma amostra de cem indivíduos, tenham sido obtidos a média de 123 mL de mercúrio e o desvio padrão de 8 mL de mercúrio. Determine o intervalo de 90% de confiança para a média.

9.7.4. Seja X a variável aleatória que representa a quantidade de hemoglobina, em gramas, encontrada em um decilitro (100 mL) de sangue total. Com base em uma amostra aleatória de duzentas mulheres adultas sadias, obteve-se a média de 14g/dL e erro padrão da média de 1,1g/dL. Determine o intervalo de 95% de confiança para μ, supondo que X seja uma variável com distribuição aproximadamente normal.

9.7.5. Seja X a variável aleatória que representa o comprimento ao nascer de filhos do sexo masculino, de mães sadias com período completo de gestação. Com base em 28 recém-nascidos masculinos, uma enfermeira calculou a média e o desvio padrão, que resultaram em 50 cm e 2,5 cm, respectivamente. Calcule o intervalo de 90% de confiança para μ, pressupondo distribuição aproximadamente normal.

9.7.6. 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 tenha

fornecido média $\bar{x}$ = 95,0 mg de glicose por 100 mL de sangue e o desvio padrão s = 23,5 mg de glicose por 100 mL de sangue. Suponha que X tenha distribuição aproximadamente normal.

9.7.7. Uma amostra de trinta homens sadios com idade entre 30 e 48 anos, não fumantes e que tinham atividade física regular, forneceu, em repouso, dados de frequência cardíaca.[6] A média foi de 63,9 bpm (batimentos por minuto) com erro padrão da média de 1,3 bpm. Calcule o intervalo de 95% de confiança para a média.

9.7.8. Num estudo sobre qualidades nutricionais[7] de lanches rápidos, mediu-se a quantidade de gordura em cem hambúrgueres de determinada cadeia de restaurantes. Foram obtidos a média de 30,2 gramas e o desvio padrão de 3,8 gramas. Construa um intervalo de 95% de confiança para a quantidade média de gordura nos hambúrgueres servidos nesses restaurantes.

9.7.9. No mesmo estudo citado no Exercício 9.7.8, foi medida a quantidade de sal e se obtiveram a média de 658mg e o desvio padrão de 47mg. Ache o intervalo de 90% de confiança.

9.7.10. Uma enfermeira mediu o comprimento de 105 bebês do sexo masculino e obteve o intervalo de 90% de confiança para a média, em centímetros: (45,3; 53,2). Responda brevemente às questões feitas em seguida:

a. A média da população está no intervalo (45,3; 53,2)?
b. A média da amostra está no intervalo (45,3; 53,2)?
c. Novas amostras de 105 bebês do sexo masculino darão médias no intervalo (45,3; 53,2)?
d. Um intervalo de 99% de confiança seria mais estreito?

---

[6]Com base em Brett, S. E. *et al*. Diastolic blood pressure change during exercise positively correlated with serum cholesterol and insulin resistance. Circulation 2000 (101): 611-615.

[7]Johnson, R. e Tsui, K. W. *Statistical reasoning and methods*. Nova York: Wiley,1998, p. 338.

[1]The Behavior of the Sample Mean. Disponível em www.jerrydallal.com.1hsp.meandist.htm. Acesso em: 20 nov. 2014.

[2]Note que, para isso ser verdade, é preciso que as variâncias das amostras tenham sido estimadas usando os graus de liberdade como divisores.

[3]Esse comportamento é descrito pelo *Teorema do Limite Central*, que diz, mais ou menos, o seguinte: a distribuição da soma de variáveis aleatórias independentes é normal, desde que a amostra seja suficientemente grande. Esse teorema é assim chamado não por fornecer um "limite central", mas por ser um teorema do limite que é *central* para a prática da Estatística, descrevendo o comportamento da média da amostra à medida que o tamanho da amostra vai aumentando.

[5]É errado dizer que um intervalo de confiança, com valores calculados com base em uma amostra, tem 95% de probabilidade de conter μ. O intervalo ou contém ou não contém μ. Sabemos apenas que temos probabilidade 95% de os intervalos calculados da mesma forma conterem μ.

CAPÍTULO 10

Teste *t* para uma Amostra

Muitas vezes, é preciso verificar se certas diretrizes ou determinações estão sendo acatadas. Neste capítulo, veremos como se faz um *teste estatístico* para informar, com certo nível de confiança e a partir dos dados de uma amostra, que as medidas tomadas em determinada população têm, em média, o valor especificado por uma instituição ou uma empresa. O teste é necessário porque se faz uma *inferência*, ou seja, usamos dados de uma amostra para informar a média da população. Toda inferência está sujeita a erro, mas o teste estatístico garante certo grau de confiança nas afirmativas.

Exemplo 10.1 Teste de uma taxa

A Organização Mundial da Saúde (OMS)[1] preconiza 15% para a taxa[2] de parto cesáreo no mundo, mas no Brasil essa taxa é muito maior. Imagine que a maior maternidade de uma metrópole brasileira informe que, nos últimos anos, tem mantido a taxa de parto cesáreo com valor próximo ao recomendado pela OMS. Para confirmar essa informação, um pesquisador precisa comparar a taxa de parto cesáreo obtida em uma amostra aleatória de prontuários dessa maternidade com a taxa de 15%, recomendada pela OMS, usando um *teste estatístico*.

[1] Disponível em http://bvsms.saude.gov.br/bvs/publicacoes/qualificacao_saude_sup/pdf/Atenc_saude2fase.pdf. Acesso em: 5 fev. 2015.

[2] Taxa de parto cesáreo é a relação entre o número total de partos cesáreos e o total de partos (normais e cesáreos) realizados por uma operadora no ano considerado.

Exemplo 10.2 Teste de uma média

Para verificar se a quantidade de flúor em dentifrícios de determinada marca comercial corresponde à quantidade especificada nas embalagens dessa marca vendidas no mercado, um químico pode tomar uma amostra de vários tubos de dentifrício da marca em questão, analisar a quantidade de flúor em cada tubo e comparar a média calculada com o valor informado nas embalagens, por meio de um teste estatístico.[3]

---

[3]Ver Vieira, S. *Estatística para a qualidade*. 3 ed. Rio de Janeiro: Elsevier, 2014.

10.1 Tomada de decisão em condições de incerteza

Imagine uma situação em que é preciso tomar uma decisão: por exemplo, você comprou um carro e precisa decidir se faz ou não o seguro contra roubo. Você pensa: se o carro for roubado e estiver segurado, recebe outro carro. Você teria, então, tomado a decisão certa. Mas, se seu carro não for roubado, você talvez até lamente ter pagado o seguro, porque não precisou dele. E se não fizer o seguro? Seu carro também pode ser ou não roubado e você irá se lamentar (se tiver perdido o carro) ou se congratular (se não tiver despendido dinheiro com seguro). Veja a Figura 10.1.

FIGURA 10.1 Decidindo certo ou errado?

Ao tomar uma decisão, *pensamos estar tomando a decisão correta*, mas podemos estar errados. Por essa razão, nas decisões que você toma na sua vida pessoal, leva em conta a própria experiência, sua intuição, os conselhos de terceiros para estimar probabilidades etc. Mas o pesquisador precisa tomar *decisões objetivas* com base em dados e dar conta a seus leitores das probabilidades de erro envolvidas em suas decisões. Deve, então, recorrer a um *teste estatístico*. É o que vamos ver neste capítulo.

10.2 Teste estatístico

Para apresentar uma pesquisa, o pesquisador precisa de dados – coletados, organizados, analisados e interpretados. Se os dados provêm de uma amostra retirada da população, o pesquisador pode apenas descrever essa amostra ou pode usá-la como base para *generalização*. A generalização passa, necessariamente, por análise estatística. Este capítulo apresenta um teste estatístico antigo, mas muito usado hoje em dia para comparar a média de uma população, estimada por meio de uma amostra, com um *valor especificado*.

Exemplo 10.3 Teste de uma média

Uma análise de dados da literatura indicou que o peso de um menino de 7 anos morador do sul do Brasil deve ser 25 kg. Um professor de Educação Física considera que esse parâmetro deve ter mudado. Pesou, então, cem meninos de 7 anos e calculou a média. Olhando essa média, o professor pode dizer se, em média, os meninos de *sua amostra* têm ou não 25 kg. Mas também pode *generalizar* seu resultado e, eventualmente, refutar a informação da literatura. Mas – para essa refutação – precisa de um *teste estatístico*.

O pesquisador tem *apenas uma amostra* e quer generalizar seus achados para toda a população. Aplica, então, um teste estatístico. O teste estatístico não impede o erro, mas calcula a probabilidade de esse erro ocorrer nesse tipo de pesquisa. Vamos ver isso devagar. Para fazer o teste, siga os passos, explicados em seguida:

1. construa as hipóteses;
2. especifique o nível de significância;
3. calcule o valor do teste;
4. interprete o resultado.

10.2.1 Construindo as hipóteses

O pesquisador coleta dados com um objetivo em mente. No Exemplo 10.3, o objetivo era verificar se o parâmetro citado na

literatura – peso de um menino de 7 anos – mudou no tempo ou em determinada população. São possíveis duas *hipóteses*: a primeira é a de que, nessa população, o peso médio de um menino de 7 anos seja de 25 kg e a segunda é a de que, nessa população, o peso médio de um menino de 7 anos *não* seja de 25 kg. Com base nos dados coletados e no resultado de um teste estatístico, o pesquisador deve decidir por uma dessas duas hipóteses, lembrando sempre que está sujeito a erro.

A primeira *hipótese* é chamada de *hipótese da nulidade* e é indicada por H_0 (lê-se agá-zero). No exemplo que estamos discutindo, a *hipótese da nulidade* afirma que a média μ dos pesos de meninos de 7 anos na população de onde o pesquisador retirou a amostra é igual a 25 kg. A segunda hipótese contradiz a primeira e, por isso, é chamada de *hipótese alternativa*. Indica-se por H_1 (lê-se agá-um). No exemplo, a *hipótese alternativa* diz que a média dos pesos de meninos de 7 anos na população de onde a amostra proveio é diferente de 25 kg.

É importante deixar claro: *as hipóteses são feitas sobre os parâmetros* – nunca sobre as estimativas. No Exemplo 10.3, o pesquisador não se perguntou se a média da amostra que obteve correspondia à média informada na literatura – era fácil ver isso. O objetivo da pesquisa era estabelecer *se o que foi observado na amostra poderia ser estendido para toda a população* de onde a amostra foi retirada.

10.2.2 Testes unilaterais e testes bilaterais

A hipótese da *nulidade* afirma: "não há diferença..." ou, então, "a diferença é *n*ula". No exemplo que acabamos de ver:

$$H_0 : \mu = 25\text{kg}$$

A hipótese *alternativa* afirma: "na população estudada, a média é *diferente*...". Dizemos, então, que o teste é *bilateral* porque, na população estudada, a média tanto pode ser *maior* como *men*or que o

parâmetro estabelecido na literatura. Pode acontecer, porém, de o pesquisador especificar o sinal da diferença (maior ou menor). Dizemos, então, que o teste é *unilateral*. É sempre mais seguro proceder a um teste bilateral. Isso porque – qualquer que seja a área de conhecimentos – alguns tratamentos têm, eventualmente, efeito contrário ao esperado.

Exemplo 10.4 Teste bilateral

Em média, comprimidos para cefaleia (dor de cabeça) aliviam a dor por 100 minutos. Para saber se uma nova formulação tem o mesmo efeito, dez voluntários usaram a nova formulação em situação de dor. A hipótese da nulidade (H_0) é a de que, em média, o tempo de alívio de dor é 100 minutos, como acontece com as outras formulações. A hipótese alternativa (H_1) é a de que o tempo médio para alívio de dor é diferente de 100 minutos.

$$H_0 : \mu = 100 \text{ minutos}$$
$$H_1 : \mu \neq 100 \text{ minutos}$$

Exemplo 10.5 Teste unilateral

A Organização Mundial de Saúde (OMS) informa que o peso médio ao nascer de nascidos a termo em países desenvolvidos no ano de 2000 era de 3,4 kg (7,5 lb). Duas médicas australianas[4] se perguntaram se o peso ao nascer de filhos de mães que fizeram uso continuado de drogas ilícitas durante a gestação não seria menor do que o informado pela OMS. Levantaram então, por volta de 2001, dados de peso ao nascer de filhos de 62 mulheres que usaram maconha durante todo o período de gestação. Obtiveram, para a idade gestacional média de 38 semanas, peso

médio ao nascer de 3,068 kg e erro padrão da média de 0,096 kg. Veja as hipóteses colocadas em teste:

- hipótese da nulidade: não há diferença entre o peso médio ao nascer de nascidos a termo de mães que fizeram uso continuado de drogas ilícitas durante a gestação e o peso médio ao nascer de nascidos a termo em países desenvolvidos informado pela OMS (3,4 kg ou 7,5 lb);
- hipótese alternativa: o peso médio ao nascer de nascidos a termo de mães que fizeram uso continuado de drogas ilícitas durante a gestação é *menor* que o peso médio ao nascer de nascidos a termo em países desenvolvidos informado pela OMS (3,4 kg ou 7,5 lb).

[4]Quilivan, JA; Evans, SF. The impact of continuing illegal drug use on teenage pregnancy outcomes. Australia: BJOG*: An International Journal of Obstetrics & Gynaecology*:109 (10):1.148-53, 2002.

10.2.3 Definindo os erros

Para quem busca informação científica, não há interesse em saber – lembrando o Exemplo 10.5 – que algumas mulheres australianas (a amostra), usuárias de maconha durante a gestação, tiveram ou não filhos com peso ao nascer mais baixo do que o esperado; o que interessa é saber se o uso de maconha na gestação é ou não fator de risco para baixo peso ao nascer (toda a população). Mas não há como estudar toda a população. Então, os pesquisadores levantam dados de *amostras* e fazem *inferência estatística* para a *população*. Veja a Figura 10.2: a *inferência estatística*, como toda inferência, está sujeita a *e*rro:

FIGURA 10.2 Erro tipo I e erro tipo II

- 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 10.6 Definindo os erros

Reveja o Exemplo 10.5. Feitas as hipóteses, quais são os erros possíveis?

Erro tipo I: rejeitar H_0 quando H_0 é verdadeira:

Dizer que o uso de maconha durante a gestação faz diminuir o peso ao nascer dos bebês, *se isso não for verdade*.

Erro tipo II: não rejeitar H_0 quando H_0 é falsa:

Dizer que o uso de maconha durante a gestação não faz diminuir o peso ao nascer dos bebês, *se isso não for verdade*.

É importante saber que a pesquisa científica deve responder a uma pergunta. O profissional de Estatística transforma a pergunta do pesquisador em duas hipóteses que se contradizem: uma negativa, outra positiva. Apenas uma das hipóteses pode ser verdadeira. Um teste estatístico conduz a decisão por uma das hipóteses. Veja a Figura 10.3.

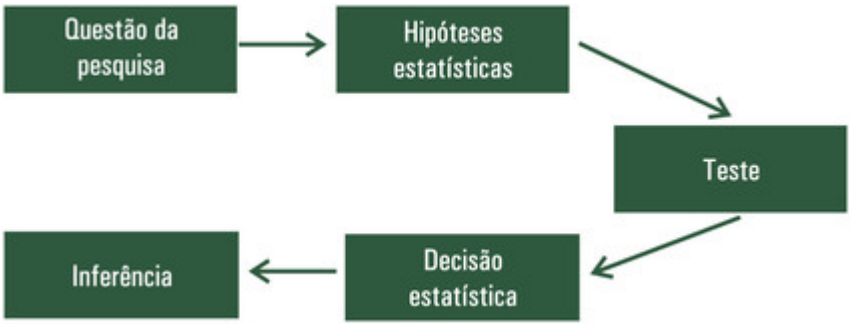


FIGURA 10.3 Decisão

Sempre é possível tomar uma decisão errada, mas os pesquisadores preferem *diminuir a probabilidade de cometer erro tipo I*. Por quê? Porque cometer erro tipo I significa dizer que uma intervenção *tem efeito* quando, na verdade, *essa intervenção não tem efeito*. O erro no resultado da pesquisa pode determinar mudanças de tratamento de pacientes, investimentos, mudanças de hábitos, sem necessidade. Veja o Exemplo 10.7.

Exemplo 10.7 Erros tipo I

- O pesquisador sugere mudança de tratamento quando conclui: "A velocidade de ação da nova droga é maior que a da droga convencional na redução da pressão sistólica".

Se *não for verdade* que a velocidade de ação da nova droga é maior que a da droga convencional, o pesquisador terá cometido erro tipo I. Evidentemente, o pesquisador não sabe *disso* quando conclui. Foi levado à conclusão errada porque errou na amostragem, ou na coleta de dados, ou no delineamento do ensaio, ou foi simples azar.

- O pesquisador sugere mudança de hábito quando conclui: "Exercício físico melhora o aproveitamento da glicose pelos músculos".

Se não for verdade que exercício físico melhora o aproveitamento da glicose pelos músculos, o pesquisador terá cometido erro tipo I.

- O pesquisador conclui: "O novo modelo de aparelho de raios X *não* é mais seguro que o antigo".

O pesquisador *n*ão estará cometendo erro tipo I, porque erro tipo I seria concluir que *o novo modelo de aparelho de raios X (investimento) é mais seguro* que o antigo. Não foi essa a conclusão.

Nível de significância é a probabilidade de se cometer erro tipo I – rejeitar H_0 quando H_0 é verdadeira. Indica-se pela letra grega α (lê-se alfa).

Nível de significância = 1 – nível de confiança

Os pesquisadores se sentem seguros para rejeitar a hipótese da nulidade (concluir que a diferença existe) quando a probabilidade de errar nessa decisão é pequena. Por essa razão, na pesquisa científica, é comum usar nível de significância de 10%, 5% ou 1%.

Se o pesquisador rejeita a hipótese da nulidade no nível de significância α = 0,05, diz que o resultado é *significante* – embora fosse melhor especificar *significante no nível de 5%*.

Se o pesquisador rejeita a hipótese da nulidade no nível de significância {Iα}{/I} = 0,01, diz que o resultado é *altamente significa*nte, embora fosse melhor especificar *significante no nível de 1%*.

Exemplo 10.8 Nível de significância

Reveja o Exemplo 10.5. Feitas as hipóteses, estabeleceu-se o nível de significância de 5% e, então, aplicou-se o teste *t*. O resultado foi significante no nível de 5%. A conclusão da pesquisa foi a de que o uso continuado de maconha durante a gestação faz diminuir o peso ao nascer dos bebês.

10.2.4 Aplicando o teste *t*

O teste *t* para uma amostra (*one sample t-test*) permite estabelecer se a média da população de onde essa amostra foi retirada tem um valor especificado. Para aplicar o teste, o pesquisador precisa ter coletado a amostra, que fornece média e erro padrão da média. O pesquisador,

então, constrói as hipóteses, estabelece o nível de significância e calcula o valor de *t* por meio da seguinte fórmula:

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}}$$

em que x/ é a média da amostra, μ é a especificação e $s_{\bar{x}} = \frac{s}{\sqrt{n}}$ é o erro padrão da média.

Exemplo 10.9 Aplicando o teste

A média de tempo de sono dos idosos internados numa instituição é de 6 horas e 8 minutos. Uma enfermeira quer saber se os idosos que residem no pavilhão em que trabalha têm ou não o mesmo tempo de sono dos demais. Uma amostra de quatro pessoas forneceu os seguintes tempos de sono, medidos em horas: 5; 4; 6; 5. O nível de significância estabelecido pela pesquisadora é de 10%. Aplique o teste *t*.

O *valor especificado* é de 6 horas e 8 minutos. As hipóteses são:

$$H_0 : \mu = 6\text{h}08\text{min}$$
$$H_1 : \mu \neq 6\text{h}08\text{min}$$

A *média* da amostra é:

$$\bar{x} = \frac{\Sigma x}{n} = \frac{5+4+6+5}{4} = \frac{20}{4} = 5$$

Para calcular o *erro padrão da média*, é preciso obter a variância. Veja os cálculos intermediários na Tabela 10.1.

Tabela 10.1

Cálculos intermediários para o cálculo da variância

x	x^2
5	25
4	16
6	36
5	25
20	102

$$s^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1} = \frac{102 - \frac{400}{4}}{4-1} = \frac{102-100}{3} = 0,667$$

$$s = \sqrt{0,667} = 0,816$$

O valor especificado para a média de tempo de sono dos idosos internados na instituição é de 6 horas e 8 minutos. Transformando 8 minutos em decimais, tem-se 6,13 h.

O valor de t é:

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}} = \frac{\bar{x} - \mu}{\sqrt{\frac{6,67}{4}}} = \frac{5 - 6,13}{0,408} = -2,77$$

Feitos os cálculos, o pesquisador deve comparar o *valor absoluto* do *t calculado* com o *valor crítico* dado em tabela de valores de *t*, com os graus de liberdade da amostra e no nível estabelecido de significância. Toda vez que o *valor absoluto* do *t* calculado for igual ou maior que o valor crítico dado na tabela, o pesquisador deve rejeitar a hipótese de nulidade, no nível estabelecido de significância.

Para entender como se encontra o valor crítico de *t*, veja a Tabela 10.2, que reproduz parte da tabela de valores de *t*, trazida, neste livro, nos Anexos. O valor crítico de *t* para um teste bilateral com, por exemplo, 4 graus de liberdade e 0,05 de significância está no cruzamento da linha 4 com a coluna 0,05. É 2,776, em negrito na Tabela 10.2.

Tabela 10.2

Tabela (parcial) de valores de *t*

Unilateral	*GL*	0,10	0,05	0,025	0,01
Bilateral	*GL*	0,20	0,10	0,05	0,02
	1	3,078	6,314	12,710	31,821
	2	1,886	2,920	4,303	6,965
	3	1,638	2,353	3,182	4,541
	4	1,533	2,132	**2,776**	3,747
	5	1,476	2,015	2,571	3,365
	6	1,440	1,943	2,447	3,143
	etc,				

Exemplo 10.10 Interpretando o resultado do teste

Reveja o Exemplo 10.9, sobre a média de tempo de sono dos idosos internados numa instituição. Estabeleceu-se nível de significância de 10% para o teste bilateral. O valor de t calculado foi -2,77. O valor crítico de t dado na tabela para 3 graus de liberdade e 10% de significância é 2,353. O *valor absoluto* do t calculado é maior que o valor crítico dado na tabela. Logo, a pesquisadora deve rejeitar a hipótese de nulidade, ou seja, deve dizer que a média de tempo de sono dos idosos sob sua responsabilidade é diferente da especificada, de 6 horas e 8 minutos ($\alpha = 10\%$).

Quem rejeita a hipótese da nulidade não tem certeza, total e absoluta, de que a decisão tomada está correta (não tem 100% de confiança). O teste estatístico *fixa o valor da probabilidade de cometer erro tipo I*, mas *não* elimina a probabilidade desse erro. De qualquer modo, é o teste estatístico que deixa claro para o pesquisador a possibilidade de estar errado em sua afirmativa – está escrito na conclusão – e ainda esclarece a probabilidade de erro nesse tipo de pesquisa.

10.2.5 Calculando o *p*-valor

Os estatísticos usam computador para fazerem testes. E – para fazerem testes estatísticos usando um programa – não se estabelece o nível de significância *a priori*, porque esses programas fornecem o *p-valor*. Calcular o *p*-valor é extremamente difícil e isso só é feito, hoje em dia, usando computador. Mas o que significa *p*-valor?

O *p*-valor diz quão provável seria obter uma amostra tal qual a que foi obtida quando a hipótese da nulidade for verdadeira.

Exemplo 10.11 Interpretando o *p*-valor

Reveja o Exemplo 10.9, sobre a média de tempo de sono dos idosos internados numa instituição. Usando o Minitab, você obtém:

One-Sample t: Tempo de sono

Test of μ = 6,13 vs ≠ 6,13

Variable	N	Mean	StDev	SE Mean	95% CI	T	P
Tempo de sono	4	5,000	0,816	0,408	(3,701; 6,299)	-2,77	0,070

Veja: testa-se a hipótese de que $\mu = 6{,}13$ contra a hipótese de que $\mu \neq 6{,}13$. Você tem $n = 4$, que é o tamanho da amostra; média igual a 5,0; desvio padrão igual a 0,816; erro padrão da média igual 0,408; intervalo de 95%; confiança para a média de 3,701 a 6,299; valor de t igual a -2,77 e p-valor igual a 0,070.

O que significa p-valor igual a 0,070? Quando a hipótese de nulidade é verdadeira, a probabilidade de se obter uma amostra tal qual a que foi obtida é 0,070, ou 7%. Como esse valor é menor que os 10% admitidos de erro, rejeita-se a hipótese de nulidade no nível de 10% de significância.

O p-valor (valor de probabilidade) permite decidir se existe *evidência suficiente* para rejeitar a hipótese de nulidade, embora o teste de hipóteses não elimine a probabilidade de erro. De qualquer modo, os pesquisadores se sentem seguros para rejeitar a hipótese de nulidade (assumir que existe a diferença procurada) quando o p-valor é pequeno.[5] Quando $p < 0{,}05$, dizemos que os resultados são significantes e, quando $p < 0{,}01$, dizemos que os resultados são altamente significantes. Isso porque seria *muito pouco provável* chegar ao resultado obtido se a diferença entre médias não existisse.

10.3 Exercícios resolvidos

10.3.1. Um réu está sendo julgado. Quais são as hipóteses possíveis? Quais são as decisões possíveis? Quais são os erros associados às decisões possíveis?

Hipóteses:

- o réu é inocente do ato de cuja prática o acusam;
- o réu é culpado do ato de cuja prática o acusam.

Decisões possíveis:

- considerar o réu culpado;
- considerar o réu inocente.

Erros possíveis:

- dizer que o réu é culpado quando é inocente;
- dizer que o réu é inocente quando é culpado.

10.3.2. Uma pessoa garante que um cão pode ser treinado para alertar seus donos no caso de o telefone tocar. Quais são as hipóteses possíveis? Quais são as decisões possíveis? Quais são os erros associados às decisões possíveis?

Hipóteses:

- não se consegue dar esse tipo de treinamento;
- consegue-se dar esse tipo de treinamento.

Decisões possíveis:

- considerar que se conseguiu o resultado com treinamento;
- considerar que não se conseguiu o resultado com treinamento.

Erros possíveis:

- dizer que se conseguiu resultado com o treinamento quando não se conseguiu;
- dizer que não se conseguiu resultado com o treinamento quando se conseguiu.

10.3.3. Um pesquisador requisitou ao biotério da universidade em que trabalha oito ratos machos da raça Wistar com 30 dias, pesando 80 gramas. Recebe, então, ratos machos da raça indicada com os seguintes pesos em gramas: 76; 81; 50; 47; 63; 65; 63; 64. Por simples inspeção, o pesquisador, acostumado a treinar ratos de laboratório, suspeita que os ratos que recebeu tenham peso menor do que o pedido. Aplicando um teste

estatístico, você diria que o peso médio dos ratos que o pesquisador recebeu corresponde ao especificado na requisição ou é menor que esse valor, no nível de significância α = 5%?

$$H_0 : \mu = 70\,\text{g}$$
$$H_1 : \mu < 70\,\text{g}$$

Para obter a média aritmética, calcule:

$$\overline{x} = \frac{\sum x}{n} = \frac{509}{8} = 63.$$

Para obter o desvio padrão, primeiro calcule a variância:

$$s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1} = \frac{33.305 - \frac{259.081}{8}}{8-1} = 131,41$$

O desvio padrão é:

$$s = \sqrt{131,41} = 11,5$$

O valor de *t* é:

$$t = \frac{\mu - 70}{\frac{s}{\sqrt{n}}} = \frac{63,6 - 70}{\frac{11,5}{\sqrt{8}}} = -1,57$$

Como a hipótese de nulidade será rejeitada apenas em uma direção (se o peso dos ratos do biotério for significantemente menor do que o valor especificado), esse é um teste unilateral. Com $n - 1 = 8 - 1 = 7$ graus de liberdade, o valor crítico na tabela de t no nível de 5% é 1,895. Não se rejeita a hipótese de nulidade, ou seja, não se pode afirmar que os pesos de ratos do biotério sejam significantemente menores do que o valor especificado.

10.3.4. Uma análise de dados da literatura indicou que a escovação de dentes com dentifrício fluoretado reduz a incidência de cárie em 30% quando comparada com o dentifrício sem flúor, considerando um acompanhamento de três anos.[6] Um cirurgião-dentista considerou esse valor muito alto. Resolveu, então, fazer uma pesquisa. Durante três anos, examinou periodicamente cem crianças de 10 a 12 anos, metade das quais usou dentifrício fluoretado, enquanto a outra metade usou dentifrício sem flúor. O cirurgião-dentista calculou as médias de incidência de cáries no grupo que usou flúor e naquele que não usou flúor. Em seguida, calculou a redução de incidência de cárie na amostra. Quais são as hipóteses em tese para um teste bilateral? E para um teste unilateral?

As hipóteses em teste são:

Para um teste bilateral:

H_0: a redução de cárie com bochechos de solução fluoretada é *igual* a 30%;

H_1: a redução de cárie com bochechos de solução fluoretada é *diferente* de 30%.

Para um teste unilateral:

H_0: a redução de cárie com bochechos de solução fluoretada é *igual* a 30%;

H_1: a redução de cárie com bochechos de solução fluoretada é *menor* de 30%.

---

[6]Chaves, SCL e Silva, LMV. A efetividade do dentifrício fluoretado no controle da cárie dental: uma meta- análise. *Rev. Saúde Pública*, v. 36 (5). São Paulo, out. de 2002.

10.4 Exercícios propostos

10.4.1. Você vai sair de casa e o céu está nublado, prenunciando chuva. Quais hipóteses você pode pôr em teste? Quais são as decisões possíveis considerando que você tem um guarda-chuva? Quais são os erros associados às decisões possíveis?

10.4.2. Um dos melhores indicadores da saúde do bebê é seu peso ao nascer.[7] Mas o peso ao nascer sofre o efeito de diversos fatores, particularmente da privação de alimentos, que pode ocorrer durante a gestação. Embora o peso médio ao nascer nos Estados Unidos seja 3.300 g, a média de peso ao nascer para filhos de mulheres que vivem em extrema pobreza é de 2.800 g. Um hospital introduziu um novo programa de cuidado pré-natal para diminuir o número de bebês com baixo peso ao nascer. No primeiro ano, 25 gestantes que viviam em extrema pobreza participaram do programa. Dados do hospital revelam que os bebês nascidos dessas mães tiveram peso médio ao nascer de 3.075 g e desvio padrão 500 g. O programa é efetivo para gestantes que vivem em extrema pobreza?

10.4.3. Um professor de Estatística quer saber se os alunos que entram na universidade têm conhecimento de Matemática suficiente para enfrentar os cursos básicos de Estatística. Ele considera que, se os alunos não conseguirem, em média, pelo menos 7 em determinada prova, devem estudar Matemática antes de iniciar o curso. Seis alunos são escolhidos ao acaso para fazer a prova. As notas deles foram: 6,2; 9,2; 7,5; 6,8; 8,3; 9,5. O professor pode ter 90% de confiança de que a nota média dos alunos está acima de 7?

10.4.4. As notas finais de estudantes de certo curso podem variar entre 1 (pior nota) e 6 (excelente). Nos últimos cinco anos, a média foi 4,7. A média e o desvio padrão de uma amostra aleatória de 22 estudantes do ano em curso foram 5,0 e 0,452, respectivamente. Há razão para suspeitar de que os novos alunos tenham notas melhores que os alunos de anos anteriores, em um nível de significância de 5%?

10.4.5. Crianças com baixa estima têm mais depressão do que crianças em geral? O escore para depressão, na população em questão, é sabidamente 90.[8] Você estuda uma amostra de cem crianças com baixa estima e encontra um escore médio para depressão de 92, com desvio padrão de 14. Qual é sua conclusão?

10.4.6. Imagine que você esteja conduzindo um ensaio para saber se determinada terapia reduz a ansiedade em alunos do curso fundamental. O valor teoricamente estabelecido para o teste de ansiedade que você vai fazer é 20. Com uma amostra casual simples de 81 alunos, você encontrou média 18 e desvio padrão 9. Qual seria sua conclusão?

10.4.7. Uma amostra aleatória dos escores da avaliação do desempenho de funcionários de uma faculdade será comparada com a média dos escores de toda a universidade nos últimos cinco anos, que foi 5,0. Os escores de avaliação do desempenho variavam de zero a 10. Qual seria sua avaliação?

Nº do funcionário	Escore	Nº do funcionário	Escore
1	5	12	4,5
2	5,5	13	4,5
3	4,5	14	5,5
4	5	15	4
5	5	16	5
6	6	17	5
7	5	18	5,5
8	5	19	4,5
9	4,5	20	5,5
10	5	21	5
11	5	22	5,5

10.4.8. A frase que segue está certa ou está errada? “O teste *t* para uma amostra é usado para verificar se a média de uma amostra é significantemente diferente de um valor especificado.”

10.4.9. Aprenda a usar um programa de computador para fazer o teste *t* para uma amostra (*one-sample t-test*). Em seguida, use o programa para refazer o Exercício 10.4.3. Encontre o *p*-valor.

10.4.10. Ache o *p*-valor para o Exercício 10.4.7. Interprete o resultado.

10.4.11. Comprimidos para cefaleia (dor de cabeça) aliviam a dor por 100 minutos, em média. Para saber se uma nova formulação tem o mesmo efeito, dez voluntários usaram a nova formulação em ocasião de dor. O tempo de alívio de dor, registrado por esses voluntários, foi de: 90; 93; 93; 99; 98; 100; 103; 104; 99; 102. Aplique o teste.

---

[7]Quantitative Methods in Social Research. Disponível em http://ccnmtl.columbia.edu/projects/qmss/ Acesso em 10 de fevereiro de 2015.

[8]http://pt.slideshare.net/shoffma5/one-sample-t-test

---

[5]Quando 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%.

CAPÍTULO 11

Teste *t* para a Comparação de Médias

Os pesquisadores trabalham com *amostras,* mas, por meio de testes estatísticos, fazem *inferência*, ou seja, generalizam suas conclusões para *as populações das quais as amostras foram retiradas*. São sempre duas as hipóteses em teste: a *hipótese da nulidade*, que, na grande maioria das vezes, afirma não existir diferença entre as duas populações em comparação, e a *hipótese alternativa*, que contradiz a primeira.

Os testes estatísticos fornecem o *p*-valor (valor de probabilidade), que permite decidir se há evidência suficiente para rejeitar a hipótese da nulidade. Em geral e por tradição, se o *p*-valor for menor do que 0,05 ($p < 0,05$), a hipótese da nulidade é rejeitada.[1] Em outras palavras, se $p < 0,05$, os resultados são *estatisticamente significantes*.

Neste capítulo, veremos como aplicar um teste estatístico para comparar *duas médias*[2] *da mesma variável quantitativa*.

Exemplo 11.1 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 da qual 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 da qual 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

Para comparar duas médias, aplica-se o teste *t* de Student, desde que seja razoável pressupor que a variável em análise tem distribuição normal ou aproximadamente normal. Vamos ver como se faz esse teste em duas situações diferentes:

1. quando os dados são pareados;
2. quando as amostras são independentes.

11.1 Teste *t* nos estudos com dados pareados

Dizemos que os dados são pareados se o pesquisador adotar um dos seguintes métodos para seu trabalho:

- medir a mesma variável nas mesmas unidades, antes e depois de uma intervenção;
- recrutar participantes da pesquisa aos pares, ou parear os participantes por idade, sexo, estágio da doença. Depois, administrar o tratamento em teste a um dos participantes de cada par, escolhido ao acaso, e ao outro, o tratamento convencional;
- medir a mesma variável em gêmeos ou outro tipo de par, como mãe e filho.

Exemplo 11.2 Ensaio com dados pareados: duas medidas obtidas em cada 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 consecutivas. A proposta consiste em comparar as médias de tempo de sono obtidas com cada antitussígeno.

Exemplo 11.3 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. Em seguida, 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), enquanto o outro foi mantido como controle. A ideia é comparar as médias dos tamanhos de tumores de ratos tratados e de ratos controles.

Quando temos dados pareados, aplicamos o teste *t*. Mas o pareamento deve ter algum tipo de lógica; não basta ter duas amostras com o mesmo número de dados. Para fazer o teste *t*,

1. estabeleça as hipóteses;
2. escolha o nível de significância;
3. calcule as diferenças entre todas as observações pareadas:

$$d = x_2 - x_1$$

4. calcule a média dessas diferenças:

$$\bar{d} = \frac{\Sigma d}{n}$$

5. calcule a variância dessas diferenças:

$$s^2 = \frac{\Sigma d^2 - \frac{(\Sigma d)^2}{n}}{n-1}$$

6. calcule o valor de *t*, que está associado a (n – 1) graus de liberdade, pela seguinte fórmula:

$$t = \frac{\bar{d}}{\sqrt{\frac{s^2}{n}}}$$

7. compare o *valor absoluto* do t calculado com o valor crítico dado em 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.

Exemplo 11.4 Aplicando o teste *t* em ensaio com dados pareados

Lembre o Exemplo 11.2: realizou-se um ensaio para verificar se pessoas submetidas a antitussígenos diferentes em duas noites consecutivas têm, em média, o mesmo tempo de sono nas duas noites. Na Tabela 11.1, estão registrados os tempos de sono de nove voluntários com cada droga. As hipóteses em teste são:

Tabela 11.1

Tempos de sono dos voluntários, em horas, segundo a droga

Voluntário	Droga	
	A	B
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

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.

Para fazer o teste:

a. calcule as diferenças entre os tempos de sono observados para cada voluntário, quando tomaram drogas diferentes, conforme apresentado na Tabela 11.2;

Tabela 11.2

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. Se você fizer os cálculos em computador,[3] vai obter o p-valor 0,0171. A conclusão é a mesma.

[3] É muito complicado calcular o p-valor, razão pela qual não se fornece, aqui, nenhuma fórmula de cálculo.

Dados pareados podem ser submetidos a testes unilaterais, desde que a pesquisa assim o exija. Veja o Exemplo 11.5.

Exemplo 11.5 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 testar se as duas drogas dão o mesmo tempo de alívio da dor, realizou-se um ensaio com sete voluntários.[4] Todos os voluntários usaram, em períodos distintos, tanto a droga tradicional como a genérica. Os tempos de alívio da dor registrados pelos voluntários com cada droga estão na Tabela 11.3.

Tabela 11.3

Tempos de alívio da dor, em horas, segundo a droga

Voluntário	Droga	
	Tradicional	Genérica
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

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%.

Para fazer o teste:

a. calcule as diferenças entre antes e depois, conforme apresentado na Tabela 11.4;

Tabela 11.4

Tempos de alívio da dor, em horas, segundo a droga, e as respectivas diferenças

Voluntário	Droga		Diferença
	Tradicional	Genérica	
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 obtém 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 *t*, é 1,94 (leia na coluna de 10%). Como o valor absoluto do *t* calculado é menor que

o valor crítico (1,871 < 1,94), não rejeite a hipótese de que o tempo de alívio da dor é, em média, o mesmo para droga tradicional e genérica, no nível de significância de 5%. Em termos do pesquisador, não há evidência estatística de que o tempo de alívio da dor seja menor quando se usa a droga genérica (*p*-valor = 0,0553 > 0,05).

---

[4]Esste tipo de teste é conhecido como de não inferioridade. O número de voluntários deve estar em torno de 25.

11.2 Teste *t* na comparação de grupos independentes

Muitas vezes, o pesquisador retira amostras de *populações independentes*. Por exemplo, pode comparar o nível de ansiedade de meninos e meninas no primeiro dia de aula. Também pode comparar dois grupos de pessoas, um grupo submetido a um novo tratamento enquanto o outro grupo é submetido a tratamento convencional.[5]

Exemplo 11.6 Ensaio para comparação de grupos independentes

Para saber se determinado produto faz nascer cabelos em pessoas calvas, um dermatologista 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.

11.2.1 Comparação das variâncias dos grupos

O teste *t* para grupos independentes compara as médias de medidas da mesma variável contínua, obtidas de forma independente em cada um de dois grupos. Antes, porém, de proceder ao teste *t*, é preciso verificar se as variâncias dos grupos são ou não desiguais.[6] Para testar a hipótese de que as variâncias das duas populações são iguais,[7] aplica-se o teste *F*, como segue:

1. estabeleça as hipóteses

   H_0: as variâncias na população são iguais

   H_1: as variâncias são diferentes ponto final

2. Escolha o nível de significância α ponto final
3. Siga os seguintes 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. 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) compare o valor calculado de *F* com o valor dado na tabela de valores *F*, com o nível de significância igual à metade do nível estabelecido e com $(n_1 - 1)$ e $(n_2 - 1)$ graus de liberdade. Rejeite a hipótese de que as variâncias das duas populações são iguais[8] no nível de significância α toda vez que o valor calculado de *F* for igual ou maior do que o valor da tabela de valores *F, no nível de significância* $\alpha/2$.

Para entender como se obtém o valor de *F* na tabela, observe a Tabela 11.5, que reproduz parte da tabela apresentada, neste livro, nos Anexos. Foi colocado em negrito o valor de *F* que deve ser utilizado para um teste bilateral com nível de significância $\alpha = 5\%$, $n_1 = 7$ graus de liberdade no numerador e $n_2 = 8$ graus de liberdade no denominador, na forma descrita aqui. O nível de significância que deve ser procurado na tabela é $\alpha/2 = 2{,}5\%$, com 7 e 8 graus de liberdade.

Tabela 11.5

Tabela (parcial) de valores de F para α = 2,5%

Número de graus de liberdade do denominador	Número de graus de liberdade do numerador								
	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

Exemplo 11.7 Teste *F* para comparar variâncias

Para verificar se a quantidade de sódio em duas marcas comerciais de sopas industrializadas tem a mesma variabilidade, um nutricionista tomou uma amostra de dez unidades de cada marca em supermercados e mediu a quantidade de sódio em cada unidade.[9] Os valores são apresentados na Tabela 11.6.

Tabela 11.6

Quantidade de sódio em miligramas por 100 mL de sopa, medida em dez unidades de cada uma das duas marcas comerciais do produto

Marca	
A	B
860	540
850	640
750	600
870	640
940	300
410	610
410	430
820	280
890	300
890	610

Para proceder ao teste, é preciso estabelecer as hipóteses e o nível de significância. Seja H_0: $\sigma_1^2 = \sigma_2^2$ contra H_1: $\sigma_1^2 \neq \sigma_2^2$; $\alpha = 5\%$. Em seguida, é preciso calcular:

a. a variância de cada grupo

Para a marca A, a variância é

$$s_A^2 = \frac{6257900 - \frac{(7690)^2}{10}}{10 - 1} = 38254,44$$

Para a marca B, a variância é

$$s_B^2 = \frac{2658300 - \frac{(4950)^2}{10}}{10 - 1} = 23116,67$$

b. o valor de *F:*

$$F = \frac{s_A^2}{s_B^2} = \frac{38254,44}{23116,67} = 1,65$$

O valor calculado de *F* está associado a 9 graus de liberdade no numerador e 9 graus de liberdade no denominador. A Tabela de valores *F* nos Anexos fornece, para α = 2,5% com 9 e 9 graus de liberdade, o valor *F* = 4,03. Então, não se rejeita a hipótese de que as variâncias sejam iguais ao nível de significância de 5%.

[9]Disponível em www.statisticshowto.com/how-to-conduct-a-statistica... Acesso em: 3 mar. 2015.

11.2.2 Teste *t* para comparar médias quando as variâncias são iguais (homocedásticas)

Quando o teste *F* resulta não significante, podemos considerar que as variâncias *não* são desiguais. Para calcular o valor de *t*, siga estes passos:

1. estabeleça as hipóteses;
2. estabeleça o nível de significância;
3. calcule a média de cada grupo;
4. calcule a variância de cada grupo;
5. calcule a *variância ponderada*, dada pela fórmula:

$$s_P^2 = \frac{(n_1 - 1)s_P^2 + (n_2 - 1)s_P^2}{n_1 + n_2 - 2}$$

6. calcule o valor de t, que está associado a $n_1 + n_2 - 2$ graus de liberdade, pela seguinte fórmula:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)s_p^2}}$$

7. compare o *valor calculado de t* (em valor absoluto) com o *valor crítico de t*, com o nível estabelecido de significância e com os mesmos graus de liberdade. Se o valor absoluto do t calculado for igual ou maior que o da tabela, rejeite a hipótese de que as médias são iguais, com o nível estabelecido de significância.

Exemplo 11.8 Teste *t* para comparar as médias de dois grupos independentes com variâncias iguais

Reveja o Exemplo 11.7: um nutricionista tomou amostras de duas marcas comerciais de sopas industrializadas, A e B, e mediu a quantidade de sódio em cada unidade.[10] Os dados estão apresentados na Tabela 11.6. Para comparar as médias da quantidade de sódio nas duas marcas:

$$H_0 : \mu_A = \mu_B$$
$$H_1 : \mu_A = \mu_B$$

Nível de significância: 0,05.

a. as médias de A e B são, respectivamente:

$$\bar{x}_A = 769$$
$$\bar{x}_B = 495$$

b. as variâncias de grupo são:

$$s_P^2 = 38254,44$$
$$s_P^2 = 23116,67$$

c. a variância ponderada é:

$$s^2 = \frac{(10-1)\times 38254,44 + (10-1)\times 23116,67}{10+10-2} = 30685,56$$

d. o valor de t com $n_1 + n_2 - 2 = 10 + 7 - 2 = 15$ graus de liberdade é:

$$t = \frac{769 - 495}{\sqrt{\left(\frac{1}{10} + \frac{1}{10}\right)30685,56}} = 3,50$$

e. como o valor calculado de *t* (em valor absoluto) é maior que o valor crítico de *t* (3,50 > 2,13) ao nível de 5% de significância, você rejeita a hipótese de que as duas marcas comerciais de sopa, A e B, tenham, em média, a mesma quantidade de sal no mesmo volume de líquido.

Em termos práticos, o nutricionista pode concluir que as quantidades de sal por 125 mL são, em média, significantemente maiores nas sopas da marca A do que nas da marca B. O *p*-valor, neste exemplo, é 0,00257 < 0,05.

[10]Disponível em http://www.statisticshowto.com/how-to-conduct-a-statistical-f-test-to-compare-two-variances/. Acesso em: 3 mar. 2015.

11.2.3 Teste *t* para comparar médias quando as variâncias são desiguais (heterocedásticas)

Quando as variâncias são diferentes, para comparar duas médias, aplica-se o teste *t*, na forma aqui descrita:

1. estabeleça as hipóteses;
2. estabeleça o nível de significância;
3. calcule a média de cada grupo:
   $\bar{x}_1$: média do grupo 1
   $\bar{x}_1$: média do grupo 2
4. calcule a variância de cada grupo:
   s_1^2: variância do grupo 1
   s_2^2: variância do grupo 2
5. calcule o valor de *t*, dado pela seguinte fórmula:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}\right)}}$$

onde n_1 é o número de elementos do grupo 1 e n_2 é o número de elementos do grupo 2.

6. calcule o número de graus de liberdade associado ao valor de t, que é a parte inteira do número g, obtido pela seguinte fórmula:

$$g = \frac{\left(\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}\right)^2}{\frac{\left(\frac{s_p^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_p^2}{n_2}\right)^2}{n_2 - 1}}$$

7. Feitos os cálculos, é preciso procurar o valor de t na tabela de valores de t, com o 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, ao nível estabelecido de significância, as médias não são iguais.

Exemplo 11.9 Teste *t* para comparar as médias de dois grupos independentes com variâncias desiguais

Para verificar se determinada droga tem efeito sobre cefaleia, um médico separou, ao acaso, um conjunto de pacientes em dois grupos: um grupo foi submetido à droga em teste (grupo tratado), enquanto o outro recebeu tratamento padrão (grupo controle). O tempo de alívio da cefaleia, em minutos, para cada participante da pesquisa, está apresentado na Tabela 11.7.

Tabela 11.7

Perdas de peso em quilogramas de pacientes segundo o grupo

Grupo	
Tratado	Controle
80	100
93	103
83	104
89	99
98	102

Para proceder ao teste *t*, é preciso estabelecer se as variâncias são ou não iguais.

Então:

1. estabeleça as hipóteses
   H_0: as variâncias na população são iguais;
   H_1: as variâncias são diferentes;
2. escolha o nível de significância α;
3. siga os passos:
   a. calcule a variância de cada grupo:
      s_1^2: a variância do grupo tratado é 53,3
      s_2^2: a variância do grupo controle é 4,3.
   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

$$g = \frac{(11,52)^2}{28,4089+,1849} = \frac{132,7104}{28,5938} = 4,64 \approx 5$$

O valor calculado de *F* está associado a 4 (numerador) e 4 (denominador) graus de liberdade. A Tabela de valores *F* nos Anexos fornece para α = 2,5% com 4 e 4 graus de

liberdade o valor $F = 9{,}60$. Então, rejeita-se a hipótese de que as variâncias são iguais com o nível de significância de 5%. Em termos práticos, a variabilidade das respostas com a nova droga é muito grande. O resultado parece não ser previsível.

Para aplicar o teste *t:*

H0: $\mu A = \mu B$

H1: $\mu A \neq \mu B$

Nível de significância: 0,05.

Agora, calcule:

1. as médias de A e B são, respectivamente:

$$\overline{x}_A = 88,60$$
$$\overline{x}_B = 101,60$$

2. as variâncias de grupo são:

$$s_A^2 = 53,30$$
$$s_B^2 = 4,30$$

3. o valor de *t*, no caso de variâncias desiguais, é dado pela seguinte fórmula:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

$$t = \frac{88,60 - 101,60}{\sqrt{\left(\frac{53,30}{5} + \frac{4,30}{5}\right)}} = \frac{-13}{3,3941} = -3,83$$

c. calcule o número de graus de liberdade associados ao valor de *F*:

$$g = \frac{(11,52)^2}{28,4089 + 0,1849} = \frac{132,7104}{28,5938} = 4,64 \approx 5$$

O valor calculado de *t* está associado a aproximadamente 5 graus de liberdade. Como o valor de *t* na tabela de valores *t* nos Anexos, com o nível de significância de 5% e com 5 graus de liberdade, é 2,57, rejeita-se a hipótese de que as médias sejam iguais. Em termos práticos, o tempo de alívio da cefaleia, em minutos, é, em média, significativamente maior no grupo que recebeu tratamento padrão. Se você fizer o teste no programa SAS, vai obter *p*-valor de 0,0141.

11.3 Exercícios resolvidos

11.3.1. Os valores apresentados na Tabela 11.8 permitem testar a hipótese de que recém-nascidos de ambos os sexos têm, em média, a mesma altura, contra a hipótese de que, em meninos, essas medidas são, em média, maiores. Teste essa hipótese, com o nível de significância de 5%.

Tabela 11.8

Tamanho da amostra, média e variância da estatura, em centímetros, de recém-nascidos, segundo o sexo

Sexo	n		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, vamos estabelecer:

H_0: as variâncias são iguais;

H_1: as variâncias são diferentes;

Nível de significância: 0,05.

Agora, calcule:

$$F = \frac{6,30}{5,76} = 1,09$$

que está associado a 1.360 (numerador) e 1.441 (denominador) graus de liberdade. Para o 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 α = 2,5%, com 1.360 e 1.441 graus de liberdade. A tabela não tem esses números de graus de liberdade, que 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, com o nível de significância de 5%, as variâncias são diferentes. A variabilidade de peso ao nascer é maior para o sexo feminino.

Para aplicar o teste *t* – no caso de variâncias desiguais:

H_0: $\mu_A = \mu_B$

H_1: $\mu_A \neq \mu_B$

Nível de significância: 0,05.

Agora, calcule:

$$t = \frac{\overline{X}1 - \overline{X}2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} = \frac{49,29 - 48,54}{\sqrt{\frac{5,76}{1442} + \frac{6,30}{1361}}} = 8,076$$

que está associado aos graus de liberdade

$$g = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1}\frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} = \frac{\left(\frac{5,76}{1442} + \frac{6,30}{1361}\right)^2}{\frac{\left(\frac{5,76}{1442}\right)^2}{1441} + \frac{\left(\frac{6,30}{1361}\right)^2}{1330}} = 2772$$

O valor calculado de *t* é maior do que o valor dado na Tabela de valores *t* nos Anexos. Rejeite, então, ao nível de significância de 5%, a hipótese de que recém-nascidos de ambos os sexos têm, em média, a mesma altura. Em termos práticos, em média, os meninos nascem com estatura maior do que as meninas.

11.3.2. Com base nos dados apresentados na Tabela 11.9, teste, com o 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 11.9

Calibre da veia esplênica em seis cães antes e após a oclusão da veia porta

Número do cão	Oclusão da veia porta	
	Antes	Depois
1	75	85
2	50	75
3	50	70
4	60	65
5	50	60
6	70	90

Note que foram tomadas duas medidas 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 apresentadas na Tabela 11.10.

Tabela 11.10

Diferenças de calibre da veia esplênica antes e após a oclusão da veia porta

Número do cão	Oclusão da veia porta		
	Antes	Depois	Diferença
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 $s^2 = 60{,}00$. Para aplicar o teste,

H_0: o calibre da veia esplênica é o mesmo, antes e após a oclusão da veia porta;
H_1: o calibre da veia esplênica é diferente após a oclusão da veia porta;
Nível de significância: 0,05.
O valor de *t*, associado a 5 graus de liberdade, é:

$$t = \frac{\bar{d}}{\sqrt{\frac{s^2}{n}}} = \frac{15,0}{\sqrt{\frac{60,0}{6}}} = 4,74$$

Para α = 5% e com 5 graus de liberdade, o valor na tabela de *t* é 2,57. Como o valor calculado de *t* é maior que o da tabela, a hipótese de que, em média, o calibre da veia esplênica seja o mesmo, antes e depois da oclusão da veia porta, deve ser rejeitada. Em termos do problema em estudo, a oclusão da veia porta determina aumento significativo do calibre da veia esplênica.

11.3.3. Reveja o Exercício 4.6.11: um professor de Odontologia quer saber se alunos que começam a atender pacientes em disciplinas clínicas têm aumento na frequência do batimento cardíaco. Mediu, então, a frequência dos batimentos cardíacos de cinco alunos de primeiro ano (que não cursam disciplinas clínicas) e de cinco alunos do segundo ano, imediatamente antes do primeiro atendimento de pacientes. Você já calculou as médias e os desvios padrões. Aplique agora um teste *t* unilateral, considerando as variâncias iguais.

Você calculou: 1° ano: média = 100,0; desvio padrão = 15,7; 2° ano: média = 125,0; desvio padrão = 15,2. Você já considerou (no Exercício 4.6.11 do Capítulo 4) que as variabilidades são praticamente iguais. Então, pressupondo variâncias iguais, o teste *t* unilateral fornece *t* = 2,56, com *p*-valor = 0,0169. Com base nesse resultado, é razoável concluir que alunos que começam a atender pacientes em disciplinas clínicas têm

aumento significante no número de batimentos cardíacos por minuto ($p < 0,05$).

11.3.4. Um nutricionista[11] quer saber se existe diferença na firmeza de iogurtes feitos de leite desnatado se, no processo de fabricação, for (ou não) adicionada determinada bactéria ao produto. 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 11.11. Faça o teste.

Tabela 11.11

Firmeza da massa de iogurte, segundo a marca e a presença ou não de bactéria

Marca	Bactéria	
	Sim	Não
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.

Os resultados estão apresentados na Tabela 11.12. O valor para t é significante. Portanto, há evidência de que a bactéria modifica a firmeza do iogurte.

Tabela 11.12

Médias, desvios padrões, valor de t para firmeza da massa de iogurte

Bactéria	Média	Desvio padrão	Valor de *t*	*p*-valor
Presente	65,9	13,7		
Ausente	61,1	12,6		
Diferença	4,71	4,35	2,87	0,0285

11.3.5. Um nutricionista quer comparar o efeito de duas dietas alimentares para perda de peso. Então, seleciona 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 são apresentados na Tabela 11.13. Faça o teste *t*.

Tabela 11.13

Perda de peso, em quilogramas, segundo a dieta

Dieta	
A	B
12	15
8	19
15	15
13	12
10	13
12	16
14	15
11	
12	
13	

Para aplicar o teste *t*:

H_0: as perdas de peso são, em média, as mesmas, para qualquer das duas dietas;

H_1: as dietas determinam as perdas médias de peso diferentes;
Nível de significância: 0,05.
Calcule

a) as médias de grupos:

$$\overline{x_1} = 12$$
$$\overline{x_2} = 15$$

b) as variâncias de grupo:

$$s_1^2 = 4,0$$
$$s_2^2 = 5,0$$

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 de liberdade é:

$$t = \frac{15 - 12}{\sqrt{\left(\frac{1}{10} + \frac{1}{7}\right)4,4}} = 2,902$$

Como o valor calculado de t (em valor absoluto) é maior que o valor crítico de t (2,902 > 2,13) ao 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, significativamente maiores quando os voluntários são submetidos à dieta B. O *p*-valor, neste exemplo, é 0,0109 < 0,05.

---

[11] Johnson, R. e Tsui, K. W. *Statistical reasoning and methods*. Nova York: Wiley, 1998, p. 437.

11.4 Exercícios propostos

11.4.1. Dez ratos machos adultos, criados em laboratório, foram separados aleatoriamente em dois grupos: um grupo foi tratado com a ração normalmente usada no laboratório, enquanto o outro grupo foi submetido a uma nova ração (experimental). Decorrido certo período, pesaram-se os ratos. Os pesos estão apresentados na Tabela 11.14. Teste a hipótese de que o peso médio dos ratos é o mesmo, para ambos os tipos de ração.

Tabela 11.14

Pesos em gramas de ratos adultos, segundo a ração

Ração	
Padrão	Experimental
200	220
180	200
190	210
190	220
180	210

11.4.2. Os quocientes de inteligência (QI) de dez crianças, medidos segundo dois testes de inteligência, A e B, estão apresentados na Tabela 11.15. Os dois testes de inteligência, A e B, fornecem, em média, o mesmo resultado?

Tabela 11.15

Valores de QI em dez crianças, segundo o teste de inteligência aplicado

Teste	
A	B
100	105
105	108
98	102
101	103
100	100
108	110
98	106
100	100
99	103
99	103

11.4.3. A Tabela 11.16 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 11.16

Pressão sanguínea sistólica de mulheres de 30 a 35 anos segundo o uso de anticoncepcionais

Uso de anticoncepcionais	
Sim	Não
111	109
119	113
121	120
113	117
116	108
126	120
128	122
123	124
122	115
121	112

11.4.4. A Tabela 11.17 apresenta o tamanho da amostra, a média e a variância dos pesos ao nascer de nascidos vivos de ambos os sexos. Teste, com o nível de significância de 1%, a hipótese de que os dois sexos têm, em média, o mesmo peso ao nascer.

Tabela 11.17

Tamanho da amostra, média e variância de pesos ao nascer de nascidos vivos, segundo o sexo

Sexo	n		s^2
Masculino	14	3,253	0,261
Feminino	13	3,130	0,265

Fonte: Arena, JFP. Estudo biométrico de recém-nascidos de uma população. *Rev. Paul. Med.*, 89 (3/4): 71-109, 1.076.

11.4.5. Para saber o efeito do frio em humanos,[12] 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, enquanto 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 11.18. O que você conclui?

Tabela 11.18

Pressão sanguínea dos ratos segundo a temperatura à qual foram submetidos

Nº do rato	26° C	Nº do rato	5° C
1	121	7	152
2	142	8	157
3	132	9	179
4	120	10	182
5	134	11	176
6	150	12	149

11.4.6. 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. Determinou-se o tempo que levou até as drogas alcançarem determinado nível no sangue. Com base nas estatísticas apresentadas na Tabela 11.19, faça o teste *t*.

Tabela 11.19

Médias e variâncias do tempo despendido para as drogas alcançarem determinado nível no sangue

Estatísticas	Droga	
	A	B
Nº de pessoas	9	7
Média	27,2	33,5
Variância	16,36	18,92

11.4.7. Para saber se o tempo de alívio da dor no pós-operatório é significativamente maior quando se administra a droga A em

vez da droga B, mais comumente usada, observou-se o tempo de alívio da dor de 25 pessoas que receberam a droga A no pós-operatório e de vinte que receberam a droga B. Com base nas estatísticas apresentadas na Tabela 11.20, faça o teste *t*.

Tabela 11.20

Médias e variâncias do tempo de alívio da dor, segundo a droga

Estatísticas	Droga	
	A	B
Número de pacientes	25	20
Média	5,5	5,0
Variância	2,25	1,69

11.4.8. 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 vinte caquis pelo novo método e vinte pelo método usual. Com base nas estatísticas apresentadas na Tabela 11.21, faça o teste *t*.

Tabela 11.21

Médias e variâncias do teor de ácido ascórbico em miligramas por 100 gramas da fruta, segundo o processo de armazenamento

Estatísticas	Armazenamento	
	Método usual	Novo método
Número de caquis	20	20
Média	33,4	41,0
Variância	4,0	6,0

11.4.9. Um nutricionista designa, ao acaso, 12 ciclistas para dois grupos: ambos os 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 leva para percorrer 10 km. Os dados estão apresentados na Tabela 11.22. Formule as hipóteses e faça o teste.

Tabela 11.22

Tempo, em minutos, para percorrer 10 km segundo o grupo

Grupo	
Suplemento de vitaminas	Placebo
15	16
18	12
20	15
14	15
16	14
19	18

11.4.10. Alguns estudos[13] indicam que o açúcar torna as crianças mais ativas, enquanto outros não encontram evidências 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 respectivas casas e retiraram todos os alimentos. Depois forneceram os alimentos por quatro 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 realizadas dentro de grupos. Esses dados constituem exemplo de dados pareados ou de grupos independentes? Quais são as hipóteses em teste?

[12]Ott, L e Mendenhall, W. *Understanding Statistics*. 6 ed. Belmont: Wadsworth, 1994, P. 305.

[13]Aliaga, M. e Gunderson, B. *Interactive Statistics*. 2 ed. New Jersey: Prentice Hall, 2003, p. 679.

---

[1]O *p*-valor pequeno indica que é muito improvável obter resultado igual ou menor do que o achado quando a hipótese da nulidade é verdadeira.

[2]Para 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.

[5]Para ver a metodologia desses ensaios: Vieira, S. e Hossne, W.S. *Metodologia científica para a área da saúde*. Rio de Janeiro: Elsevier, 2015.

[6]O programa Excel, muito conhecido pelos usuários de Estatística, pede que se indique o tipo de teste: *t* pareado; variâncias iguais das duas amostras (homocedástico); variâncias desiguais das duas amostras (heterocedástico).

[7]As duas populações das quais foram obtidas as amostras devem ter distribuição normal ou, pelo menos, simétrica.

[8]Aqui, a hipótese alternativa é necessariamente de um teste bilateral.

CAPÍTULO 12

Teste χ^2 para Variáveis Qualitativas

As pesquisas são feitas com o objetivo de responder a perguntas. E, para responder a perguntas, são necessárias informações obtidas por meio de amostras. Depois, com base nos dados da amostra e no resultado de um teste estatístico, os pesquisadores *generalizam seus achados para toda a população, aplicando testes estatísticos*. As tabelas 2 x 2 têm sido, possivelmente, a forma mais empregada para mostrar evidência estatística. O teste estatístico mais simples e mais conhecido é o teste de χ^2 (lê-se qui-quadrado). Neste capítulo, vamos ver como se faz esse teste.

12.1 Teste χ^2 para a associação de duas variáveis

Você aplica o teste de χ^2 (lê-se qui-quadrado) para verificar *se existe associação entre duas variáveis qualitativas*. Para isso, é preciso contar quantos participantes estão em cada uma das categorias de cada uma das variáveis. As contagens (*frequências*) são apresentadas em tabelas de contingência. Veja o Exemplo 12.1.

Exemplo 12.1 Uma tabela de contingência 2 x 2

Foram entrevistadas 1.091 pessoas residentes em uma área metropolitana da região Sul do Brasil. Cada pessoa foi classificada segundo duas variáveis: sexo (homem ou mulher) e tabagismo (tabagista ou não). Depois, foram feitas as contagens: havia seiscentos homens, dos quais 177 disseram ser tabagistas, e 491 mulheres, das quais 204 afirmaram ser tabagistas. Esses dados estão apresentados na Tabela 12.1.

Tabela 12.1

Tabagismo segundo sexo

Sexo	Tabagismo		Total
	Não	Sim	
Homens	423	177	600
Mulheres	287	204	491
Total	710	381	1.091

Fonte: Moreira, L. *et al*. Prevalência de tabagismo e fatores associados em área metropolitana da região Sul do Brasil. *Rev. Saúde Pública* 29 (1). São Paulo, 1995.

É importante apresentar as proporções observadas quando se faz um estudo transversal,[1] ou seja, quando se toma uma amostra da

população e se classifica cada pessoa segundo duas variáveis, *ao mesmo tempo*. Veja o Exemplo 12.2: para cada uma das 1.091 pessoas, foram registradas duas variáveis: 1. sexo (homem ou mulher) e 2. tabagismo (não ou sim).

Exemplo 12.2 Proporções obtidas por estudo transversal

Reveja o Exemplo 12.1. A Tabela 12.2 apresenta as proporções obtidas nesse estudo.

Tabela 12.2

Proporções obtidas por estudo transversal

Sexo	Tabagismo		Total
	Não	Sim	
Homens	0,39	0,16	0,55
Mulheres	0,26	0,19	0,45
Total	0,65	0,35	1,00

Vamos apresentar aqui o teste χ^2 (qui-quadrado), que se faz para estudar a associação de *duas variáveis* que se apresentam em apenas duas categorias. Para proceder a um *teste estatístico*, você já sabe: é preciso estabelecer as *hipóteses* em teste e o nível de significância. Em seguida, é preciso calcular a estatística de teste, que, no caso que estamos estudando, é o valor de χ^2.

As hipóteses em teste são:

H_0: as variáveis são independentes

H_1: as variáveis estão associadas

O nível de significância é α e a estatística de teste é:

$$\chi^2 = \frac{(ad - bc)^2 n}{(a+b)(c+d)(a+c)(b+d)}$$

Sob a hipótese da nulidade, a estatística calculada tem distribuição de χ^2. Mas o que significa isso tudo? Vamos devagar: veja a Tabela 12.3, que apresenta duas variáveis indicadas 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 12.3

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

De posse dos dados, você calcula o valor de χ^2. Se esse valor for maior do que o valor dado na tabela de χ^2 com 1 grau de liberdade e para o nível de significância estabelecido, você rejeita a hipótese de independência. Para calcular o valor de χ^2 na tabela, observe a Tabela 12.4, que reproduz parte da tabela de χ^2 do Apêndice. Foi sombreado o valor de χ^2 com três graus de liberdade, no nível de significância de 5%.

Tabela 12.4

Tabela (parcial) de χ^2 segundo os graus de liberdade e o valor do nível de significância

Graus de liberdade	Nível de significância		
	10%	5%	1%
1	2,71	3,84	6,64
2	4,6	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

Exemplo 12.3 Calculando o valor de χ^2

Reveja o Exemplo 12.1. A Tabela 12.1 está reproduzida aqui como Tabela 12.5, a fim de facilitar o acompanhamento dos cálculos.

Tabela 12.5

Tabagismo segundo sexo

Sexo	Tabagismo		
	Não	Sim	Total
Homens	423	177	600
Mulheres	287	204	491
Total	710	381	1.091

É preciso estabelecer as hipóteses e o nível de significância, bem como calcular o valor de χ^2. Então:

H_0: tabagismo independe do sexo.

H_1: tabagismo está associado ao sexo.

Nível de significância: 0,05.

$$\chi^2 = \frac{(ad - bc)^2 n}{(a+b)(c+d)(a+c)(b+d)}$$

$$\chi^2 = \frac{(423 \times 204 - 177 \times 287)^2 \times 1091}{600 \times 491 \times 710 \times 381} = 17,25$$

Como o valor calculado de χ^2 (17,25) é maior do que o valor dado na Tabela de χ^2 ao nível de 5% de significância (3,84), rejeita-se a hipótese de independência. A associação entre sexo e hábito de fumar é significante.

É mais correto calcular a estatística de teste com *correção de continuidade*. Fazendo essa correção,[2] que indicaremos por χ_c^2, a estatística de teste fica como segue:

$$\chi_c^2 = \frac{(|\, ad - bc \,| - \frac{1}{2}n)^2 n}{(a+b)(c+d)(a+c)(b+d)}$$

A correção de continuidade reduz o valor de χ^2 porque se reduz o numerador.[3] O efeito da correção de continuidade sobre o valor de χ^2 é maior quando a amostra é grande. Veja o cálculo para os dados apresentados na Tabela 12.2.

$$\chi_c^2 = \frac{(|\, 423 \times 204 - 287 \times 381 \,| - \frac{1}{2}1.091)^2 \times 1.091}{600 \times 491 \times 710 \times 381} = 16,72$$

Preste, portanto, muita atenção, *porque acontece* o seguinte: você aplica o teste χ^2 para testar a independência de duas variáveis a determinado conjunto de dados: sem a correção de continuidade, o resultado é significante; com a correção, é não significante. Fique, então, com a seguinte conclusão este último resultado: as variáveis são independentes.

12.1.1 Medidas de associação

É comum usar o valor de χ^2como medida de associação – o que está errado. O teste mede a *significância* da associação, mas não o *grau de associação*. O valor de χ^2 aumenta com o tamanho da amostra, desde que as proporções sejam mantidas. Então, se a amostra for grande, é mais certo encontrar significância mesmo que a associação seja apenas trivial.

12.1.1.1 Coeficiente fi

Uma medida do *grau de associação de duas variáveis* – no Exemplo 12.1, sexo e tabagismo – é o coeficiente φ (lê-se fi). Esse coeficiente não sofre influência do tamanho da amostra e é obtido facilmente a partir do valor não corrigido do χ^2. Veja a fórmula:

$$\varphi = \sqrt{\frac{\chi^2}{n}}$$

Você interpreta o resultado do coeficiente φ da seguinte forma:

1. o valor do coeficiente φ varia entre zero e um, ou seja, $0 \leq \varphi \leq 1$;
2. quanto mais próximo de 1 estiver o valor de φ, maior é o grau de associação entre as variáveis; quanto mais próximo de zero estiver o valor de φ, menor é a associação entre as variáveis;
3. $\varphi = 1$ significa associação perfeita;[4]
4. $\varphi = 0$ significa associação nula;

5. como regra prática, valores de φ menores do que 0,30 ou 0,35 podem ser tomados como indicadores de associação trivial[5] entre as duas variáveis.

Exemplo 12.4 Calculando o coeficiente fi

Para os dados do Exemplo 12.1, o tamanho da amostra é n =1.091. O valor de χ^2 sem correção de continuidade, apresentado no Exemplo 12.3, é 17,25. Então, o coeficiente de associação φ é

$$\varphi = \sqrt{\frac{\chi^2}{n}} = \sqrt{\frac{17,25}{1901}} = 0,126$$

A associação, embora significante como mostrada pelo teste χ^2, é apenas *trivial*. Não se pode considerar que a associação encontrada entre tabagismo e sexo feminino tenha maior importância.

12.1.1.2 Coeficiente gama

O coeficiente gama,[6] que se representa pela letra grega γ (lê-se gama), mede o grau de associação com que duas categorias ordenadas de variáveis tendem a crescer – e, portanto, decrescer – juntas. É definido por:

$$\gamma = \frac{ad - bc}{ad + bc}$$

em que a, b, c e d são os valores definidos na Tabela 12.3. O valor do coeficiente gama deve ser interpretado como segue:

- γ = 1: associação perfeita positiva

- $\gamma = -1$: associação perfeita negativa
- $\gamma = 0$: associação nula
- $0 < \gamma < 1$: associação positiva
- $-1 < \gamma < 0$: associação negativa

O coeficiente gama fica entre –1 e +1, inclusive, ou seja, $-1 \leq \gamma \leq +1$. Então, o coeficiente gama fornece, além do grau de associação entre duas variáveis qualitativas, o sentido da associação. Cuidado, portanto, ao desenhar a tabela para calcular o coeficiente γ, porque, ao inverter as linhas, muda o sinal do coeficiente (e, evidentemente, a interpretação).

Exemplo 12.5 Interpretando o valor do coeficiente gama

Para os dados do Exemplo 12.1, o coeficiente γ é:

$$\gamma = \frac{ad - bc}{ad + bc} = \frac{423 \times 204 - 177 \times 287}{423 \times 204 + 177 \times 287} = 0,259$$

Se a Tabela 12.1 estivesse na forma apresentada na Tabela 12.6, mostrada em seguida, o coeficiente γ seria:

Tabela 12.6

Tabagismo segundo o sexo

	Tabagismo		Total
Sexo	Sim	Não	
Homens	177	423	600
Mulheres	204	287	491
Total	381	710	1.091

$$\gamma = \frac{ad - bc}{ad + bc} = \frac{177 \times 287 - 423 \times 204}{177 \times 287 + 423 \times 204} = -0,259$$

Compare o coeficiente γ obtido para a Tabela 12.1 com o obtido para a Tabela 12.6: o primeiro mostra que, embora em pequeno grau, homens estão *positivamente associados ao hábito de não fumar*, enquanto o segundo mostra que, embora em pequeno grau, a associação entre *homens e hábito de fumar é negativa*.

12.1.2 Restrições ao uso do teste χ^2 para associação

É importante saber que o teste χ^2 apresenta muitas restrições de uso. Vejamos:

- os dados devem estar apresentados em tabelas de contingência;
- as variáveis em estudo são, obrigatoriamente, qualitativas;
- a amostra deve ter sido obtida por processo aleatório;
- a população deve ter, no mínimo, dez vezes o tamanho da amostra.

12.2 Teste χ^2 para comparar dois grupos em ensaios clínicos

Ensaio clínico[7] é um estudo no qual os pesquisadores avaliam, nos participantes da pesquisa, os efeitos de intervenções. Depois, comparam os resultados. Veja o Exemplo 12.6.

Exemplo 12.6 Comparando dois grupos nos ensaios clínicos

Para estudar a efetividade da betametasona no alívio da dor após a instrumentação endodôntica (tratamento de canal), um cirurgião-dentista fez um ensaio clínico. Antes do procedimento, administrou dois comprimidos de placebo para 17 pacientes (grupo placebo controlado) e dois comprimidos da droga para 21 pacientes (grupo tratado com betametasona). 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 são apresentados na Tabela 12.7.

Tabela 12.7

Distribuição dos pacientes segundo o grupo e o relato de alívio da dor

Grupo	Relato de alívio da dor		Total
	Sim	Não	
Placebo	2	15	17
Betametasona	12	9	21
Total	14	24	38

Fonte: Quintana-Gomes Jr. *et al*. Estudo clínico dos efeitos da betametasona sobre incidência da dor após a instrumentação endodôntica. *JBC – Jornal Brasileiro de Odontologia Clínica* (2):12, s. d.

12.2.1 Teste χ^2 nos ensaios clínicos

Para comparar as proporções de respostas positivas obtidas, por exemplo, por dois tratamentos concorrentes ou por um novo tratamento e um controle, é preciso fazer um *teste estatístico*. Neste caso, é possível aplicar o teste χ^2. Para proceder ao *teste estatístico*, estabelecem-se as hipóteses e o nível de significância. Depois, calcula-se a estatística de teste:

$$\chi^2 = \frac{(|\, ad - bc\,| - \frac{1}{2}n)^2\, n}{(a+b)(c+d)(a+c)(b+d)}$$

O teste consiste em rejeitar a hipótese de nulidade toda vez que o valor calculado de χ^2 for maior do que o valor dado na tabela de χ^2 com 1 grau de liberdade e para o nível estabelecido de significância.

Exemplo 12.7 O teste de χ^2 em ensaios clínicos

Reveja o Exemplo 12.6. Para aplicar o teste, é preciso estabelecer as hipóteses e o nível de significância. Então:

H_0: as probabilidades de relatos de dor são iguais em ambos os grupos, ou seja, $P_1 = P_2$.

H_1: a probabilidade de relatos de dor é diferente nos dois grupos, ou seja, $P_1 = P_2$.

Nível de significância: 0,05.

Depois, calcula-se

$$\chi^2 = \frac{(|ad - bc| - \frac{1}{2}n)^2 n}{(a+b)(c+d)(a+c)(b+d)}$$

$$\chi^2 = \frac{(|2 \times 9 - 15 \times 12| - \frac{1}{2}38)^2 38}{17 \times 21 \times 14 \times 24} = 6,48$$

Como o valor calculado de χ^2 (6,48) é maior do que o valor de χ^2 com um grau de liberdade e ao nível 5% de significância (3,84), rejeita-se H_0. Em termos do ensaio, o uso de betamesona após a instrumentação endodôntica diminui a probabilidade de dor.

12.2.2 Teste *z* nos ensaios clínicos

Embora seja comum apresentar dados de ensaios clínicos como na Tabela 12.7, há autores[8] que preferem fazê-lo na forma da Tabela 12.8, que exibe proporções. Assim, o tamanho da amostra (pequeno, no exemplo) e as proporções em comparação ficam mais visíveis.

Tabela 12.8

Proporção de pacientes com relato de dor após a instrumentação endodôntica, segundo o grupo

Grupo	Tamanho da amostra	Proporção de pacientes com relato de dor
Betametasona	17	0,118
Placebo	21	0,571
Total	38	0,368

Fonte: Quintana-Gomes Jr. *et al*. Estudo clínico dos efeitos da betametasona sobre incidência da dor após a instrumentação endodôntica. *JBC – Jornal Brasileiro de Odontologia Clínica* (2):12, s. d.

A significância estatística da diferença das proporções de respostas negativas (ou positivas) obtidas, por exemplo, por dois tratamentos concorrentes, ou por um novo tratamento e um controle, pode ser obtida por meio da estatística:

$$z = \frac{\left| p_2 - p_1 \right| - \frac{1}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}{\sqrt{\overline{p} \times \overline{q} \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Os valores n_1 e n_2 são os tamanhos das amostras de cada grupo; p_1 e p_2 são as proporções do evento em estudo nas respectivas amostras; $\overline{p}$ é a proporção média das duas amostras; e $\overline{q} = 1 - \overline{p}$. Para testar a hipótese H_0: $P_1 = P_2$ contra a hipótese H_1: $P_1 \neq P_2$, procura-se o valor de z numa tabela de distribuição normal. No entanto – aqui entra uma definição teórica importante –, como z^2 tem distribuição de χ^2 com 1 grau de liberdade, o valor obtido de z, elevado ao quadrado, pode ser comparado com o valor de χ^2 com 1 grau de liberdade – isso é mais fácil do que usar a tabela de distribuição normal padronizada. Rejeita-se a hipótese de nulidade se o valor calculado de z^2 for maior do que o valor dado na tabela de χ^2, com 1 grau de liberdade para o nível estabelecido de significância.

Exemplo 12.8 Outro teste para comparar duas proporções em ensaio clínico

Reveja o Exemplo 12.6. Para aplicar o teste:

H_0: as probabilidades de relatos de alívio de dor são iguais nos dois grupos, ou seja, $P_1 = P_2$.

H_1: a probabilidade de relatos de alívio de dor é menor no grupo que recebeu betametasona, ou seja, $P_2 > P_1$.

Nível de significância: 0,05.

Temos $\overline{q} = 1 - \overline{p} = 1 - 0,368 = 0,632$. Então:

$$z = \frac{\left| 0,571 - 0,118 \right| - \frac{1}{2}\left(\frac{1}{17} + \frac{1}{21}\right)}{\sqrt{0,368 \times 0,632 \times \left(\frac{1}{17} + \frac{1}{21}\right)}} = 2,54$$

O valor de z^2 é 6,48, maior que o valor de χ^2 com o nível 5% de significância. Rejeita-se H_0. Em termos da pesquisa, pode-se concluir que o uso de betamesona após a instrumentação endodôntica diminui a probabilidade de dor.

É importante lembrar que, em um trabalho de pesquisa, se deve fazer apenas um dos testes apresentados aqui. Aliás, ambos conduzem ao mesmo resultado. A questão é que os programas de computador oferecem várias opções – e alguém inexperiente pode achar que, ao colocar todas as opções, tornará seus resultados mais convincentes.

12.3 Teste χ^2 nos estudos prospectivos e retrospectivos

12.3.1 Teste χ^2 nos estudos prospectivos

A probabilidade de ocorrer determinado desfecho não é a mesma em todas as populações. Por exemplo, a probabilidade de morte violenta é maior entre jovens do sexo masculino do que entre jovens do sexo feminino. Para comparar probabilidades, pode-se fazer um *estudo prospectivo*.[9] No estudo prospectivo, uma das duas populações está exposta a um fator que se presume de risco (por exemplo, fumantes), enquanto a outra não está (não fumantes); o pesquisador, então, procura, nas amostras, determinado desfecho (câncer de pulmão). Veja a Figura 12.1.

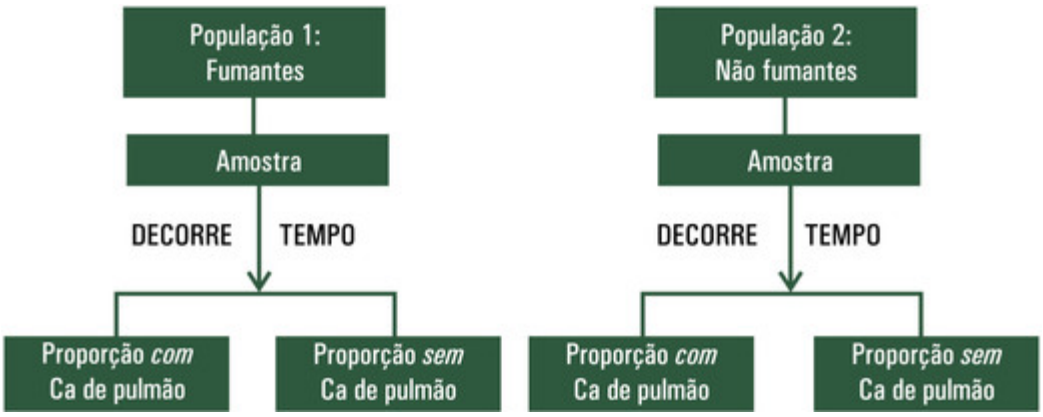


FIGURA 12.1 Estudo prospectivo

Exemplo 12.9 Um estudo prospectivo

Entre 2004 e 2006, foi feito um estudo prospectivo com 1.229 gestantes de Campinas, SP, para avaliar os fatores de risco comumente associados a desfechos desfavoráveis na saúde de recém-nascidos, como baixo peso ao nascer ou prematuridade.[10]

Veja, na Tabela 12.9, os dados para um desses fatores – consumo de cigarros durante a gestação – que permitem estimar riscos.

Tabela 12.9

Estimativas do risco de baixo peso ao nascer ou prematuridade segundo o consumo ou não de cigarros durante a gestação

Fumante na gestação	Baixo peso ou prematuro		Total
	Sim	Não	
Sim	44	121	165
Não	146	918	1.064
Total	190	1.039	1.229

Fonte: AUDI, C. A. F *et al*. Associação entre violência doméstica na gestação e peso ao nascer ou prematuridade. *J. Pediatr.*, v. 4, n. 1, Porto Alegre. Jan/fev. de 2008.

Para testar a hipótese de que a proporção de pessoas com uma característica específica é a mesma em duas amostras independentes, pode-se optar pelo teste de χ^2. Para proceder ao teste, estabelecem-se as hipóteses e o nível de significância. Em seguida, calcula-se a estatística de teste:

$$\chi^2 = \frac{(|ad - bc| - \frac{1}{2}n)^2 n}{(a+b)(c+d)(a+c)(b+d)}$$

[10]O teste tem mais poder quando os tamanhos de grupos são iguais ou, pelo menos, similares. Neste exemplo, há grande disparidade: os tamanhos de grupos para fumantes e não fumantes são, respectivamente, 165 e 1.065.

Exemplo 12.10 Teste χ^2 para um estudo prospectivo

Reveja o Exemplo 12.9. As hipóteses em teste são:

H_0: a proporção de nascituros com baixo peso ao nascer é a mesma entre gestantes fumantes e gestantes não fumantes, ou seja, $P_1 = P_2$.

H_1: a proporção de nascituros com baixo peso ao nascer é diferente entre gestantes fumantes e gestantes não fumantes, ou seja, $P_2 \neq P_1$..

Nível de significância: 0,05.

Agora, é preciso calcular:

$$\chi^2 = \frac{(|44 \times 918 - 121 \times 146| - \frac{1}{2}1229)^2 1229}{165 \times 1064 \times 146 \times 918} = 17,34$$

Como o valor calculado de χ^2 (17,34) é maior do que o valor de χ^2 com 1 grau de liberdade e ao nível de 5% de significância (3,84), rejeita-se H_0. Em termos do estudo, gestantes que fumam apresentam maior probabilidade de ter bebês de baixo peso ou prematuros.

12.3.1.1 Teste dos grupos com base na distribuição normal

Nos estudos prospectivos, deve ser apresentada a *proporção* dos que têm o desfecho buscado, tanto na amostra dos *expostos ao fator que se presume de risco* como na amostra dos *não expostos*. Veja o Exemplo 12.11, que exibe essas proporções. É mais comum apresentar dados de estudos prospectivos como na Tabela 12.9, mas há autores[11] que preferem fazê-lo na forma da Tabela 12.10, pois são essas proporções que estão em comparação.

Tabela 12.10

Proporção de nascituros com baixo peso ao nascer ou prematuros, segundo o fato de a mãe ter fumado ou não na gestação

Fumante na gestação	Amostra	Proporção de nascituros com baixo peso e prematuros
Sim	165	0,2677
Não	1.064	0,1372
Total	1.229	0,1546

Exemplo 12.11 Proporções obtidas por estudo retrospectivo

Reveja o Exemplo 12.9. As hipóteses em teste são:

H_0: a proporção de nascituros com baixo peso ao nascer é a mesma entre gestantes fumantes e gestantes não fumantes, ou seja, $P_1 = P_2$.

H_1: a proporção de nascituros com baixo peso ao nascer entre gestantes fumantes é diferente da proporção de nascituros com baixo peso ao nascer entre gestantes não fumantes, ou seja, $P_2 \neq P_1$.

Nível de significância: 0,05.

Para verificar a significância estatística da diferença de proporções em populações independentes, pode ser calculada a estatística:

$$z = \frac{\left| p_2 - p_1 \right| - \frac{1}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}{\sqrt{\overline{p} \times \overline{q} \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Os valores n_1 e n_2 são os tamanhos das amostras de cada grupo; p_1 e p_2 são as proporções de expostos ao fator que se presume de risco nas respectivas amostras (com e sem o problema); $\overline{p}$ é a proporção média das duas amostras; e $\overline{q} = 1 - \overline{p}$. Para testar a hipótese H_0: $P_1 = P_2$ contra a hipótese H_1: $P_1 \neq P_2$, procura-se o valor de z numa tabela de distribuição normal. Como z^2 tem distribuição de χ^2 com 1 grau de liberdade, o valor obtido de z, elevado ao quadrado, pode ser comparado com o valor de χ^2 com 1 grau de liberdade. Rejeita-se, portanto, a hipótese de nulidade, se o valor calculado de z^2 for maior do que o valor dado na tabela de χ^2 para o nível estabelecido de significância.

Exemplo 12.12 Teste para duas proporções em estudos prospectivos

Reveja o Exemplo 12.9. As hipóteses em teste são:

H_0: a proporção de nascituros com baixo peso ao nascer é a mesma entre gestantes fumantes e gestantes não fumantes, isto é, $P_1 = P_2$;

H_1: a proporção de nascituros com baixo peso ao nascer é diferente entre gestantes fumantes e gestantes não fumantes, ou seja, $P_2 \neq P_1$;

Nível de significância: 0,05.

Agora, é preciso calcular:

Temos $\overline{q} = 1 - \overline{p} = 1 - 0,1546 = 0,8454$. Então:

$$z = \frac{\left| 0,2667 - 0,1372 \right| - \frac{1}{2}\left(\frac{1}{165} + \frac{1}{1064}\right)}{\sqrt{0,1546 \times 0,8454 \times \left(\frac{1}{165} + \frac{1}{1064}\right)}} = 4,164$$

Como o valor calculado de z^2 é 17,34, maior do que o valor dado na tabela de χ^2 ao nível estabelecido de significância, rejeita-se H_0.

O hábito de fumar da gestante está relacionado com baixo peso ou prematuridade do nascituro.

12.3.2 Teste χ^2 nos estudos retrospectivos

No *estudo retrospectivo*, uma das populações é definida por ter (casos) de pulmão, enquanto e a outra por não ter (controles) determinado desfecho (por exemplo, câncer de pulmão); nas amostras, o pesquisador procura saber se houve exposição ao fator que se presume de risco (fumar). Então, o o estudo retrospectivo vai do efeito para a causa. Veja a Figura 12.2.

FIGURA 12.2 Estudo retrospectivo

Exemplo 12.13 Um estudo retrospectivo

Em uma pesquisa, perguntou-se a 142 jovens que apresentavam desordens mandibulares (o desfecho) se haviam ou não usado aparelho ortodôntico: 87 disseram que sim, ou seja, 87 foram expostos ao fator de risco. Também se perguntou a 228 jovens que *não* tinham desordens mandibulares se haviam ou não usado aparelho ortodôntico: 113 responderam que sim, ou seja, 113 foram expostos ao fator de risco. Esse é um estudo retrospectivo. Os dados estão apresentados na Tabela 12.11.

Tabela 12.11

Sintomas de desordens temporomandibulares (DTM) e uso de aparelho ortodôntico

DTM	Uso de aparelho		Total
	Sim	Não	
Sim	87	55	142
Não	113	115	228
Total	200	170	370

Fonte: Rizzati-Barbosa, C. M. *et al*. Correlação entre aparelho ortodôntico e desordens temporomandibulares. *J Bras. Ortodon. Ortop. Facial.* 7(39): 185-192, 2002.

Para testar a hipótese de que a proporção de pessoas com uma característica específica é a mesma em duas amostras independentes, pode-se optar pelo teste de χ^2. Para proceder ao teste no caso de estudos retrospectivos, estabelecem-se as hipóteses e o nível de significância. Depois se calcula a estatística de teste:

$$\chi^2 = \frac{(|ad - bc| - \frac{1}{2}n)^2 n}{(a+b)(c+d)(a+c)(b+d)}$$

Exemplo 12.14 Teste χ^2 para um estudo retrospectivo

Reveja o Exemplo 12.13. As hipóteses em teste são:

H_0: a proporção de jovens que usaram aparelho ortodôntico é a mesma entre os que apresentam e os que não apresentam DTM, isto é, $P_1 = P_2$;

H_1: a proporção de jovens que usaram aparelho ortodôntico é diferente para os que apresentam e os que não apresentam DTM, isto é, $P_2 \neq P_{1;}$.

Nível de significância: 0,05.

Agora, é preciso calcular:

$$\chi^2 = \frac{(|\,87 \times 115 - 55 \times 113\,| - \frac{1}{2}370)^2 \times 370}{142 \times 228 \times 200 \times 170} = 4,37$$

Como o valor calculado de χ^2 (4,37) é maior do que o valor de χ^2 com 1 grau de liberdade e com o nível de 5% de significância (3,84), rejeita-se H_0. Em termos do estudo, o uso de aparelho ortodôntico pode aumentar a probabilidade de DTM.

12.3.2.1 Teste dos grupos com base na distribuição normal

Nos estudos retrospectivos, deve ser apresentada a *proporção dos que foram expostos ao fator que se presume de risco,* tanto na amostra das pessoas que têm o problema em estudo como na amostra daquelas pessoas que não têm o problema. Veja o Exemplo 12.15, que exibe essas proporções. Embora seja mais comum apresentar dados de estudos retrospectivos como na Tabela 12.11, há quem[12] prefira fazê-lo na forma da Tabela 12.12, pois são essas proporções que estão em comparação.

Tabela 12.12

Proporção de jovens que usaram aparelho ortodôntico entre os que têm e os que não têm DTM

DTM	Amostra	Proporção de usuários
Sim	142	0,613
Não	228	0,496
Total	370	0,541

Exemplo 12.15 Proporções obtidas por estudo retrospectivo

Reveja a Tabela 12.12: p_1 = 0,613 dos 142 jovens com DTM foram expostos ao fator que se presume de risco, o uso de aparelho ortodôntico, e p_2 = 0,496 dos 228 jovens que não apresentavam DTM também foram expostos ao fator que se presume de risco, o uso de aparelho ortodôntico.

Para verificar a significância estatística da diferença de proporções em populações independentes, pode ser calculada a estatística:

$$z = \frac{\left| p_2 - p_1 - \frac{1}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right|}{\sqrt{\overline{p} \times \overline{q} \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Os valores n_1 e n_2 são os tamanhos das amostras de cada grupo; p_1 e p_2 são as proporções de expostos ao fator que se presume de risco nas respectivas amostras (com e sem o problema); $\overline{p}$ é a proporção média das duas amostras; e $\overline{q} = 1 - \overline{p}$. Para testar a hipótese H_0: $P_1 = P_2$ contra a hipótese H_1: $P_1 \neq P_2$, procura-se o valor de z numa tabela

de distribuição normal. Como z^2 tem distribuição de χ^2 com 1 grau de liberdade, o valor obtido de z, elevado ao quadrado, pode ser comparado com o valor de χ^2 com 1 grau de liberdade. Rejeita-se, portanto, a hipótese de nulidade, se o valor calculado de z^2 for maior do que o valor dado na tabela de χ^2 para o nível estabelecido de significância.

12.4 Risco relativo e razão de chances

Risco é a probabilidade da ocorrência de algum tipo de dano. Fator de risco é o fator que aumenta o risco, portanto, que afeta a probabilidade de ocorrer dano.

Por exemplo, sempre há risco de ocorrer um acidente de trânsito, mas o risco aumenta quando muita chuva ou quando o motorista ingere bebida alcoólica. Dizemos, então, que muita chuva ou motorista embriagado são *fatores de risco* para acidente de trânsito.

É possível estudar riscos por meio de ensaios clínicos e de estudos prospectivos. Reveja a Tabela 2.7, que apresenta pacientes com relato de dor após a instrumentação endodôntica. O objetivo do estudo foi o de comparar a proporção de pacientes com dor, em dois grupos: o tratado, que recebeu betametasona, e o controle, que não recebeu betametasona. Então, a estimativa de risco de dor para pacientes que receberam betametasona foi 11,8% e, para pacientes que não receberam betametasona, 5,71%.

Denomina-se *risco relativo*, que se indica por *RR*, a razão entre duas estimativas de risco. Veja o exemplo a seguir:

$$RR = \frac{57,1}{11,8} = 4,86 \cong 5$$

Neste exemplo, o risco relativo é de aproximadamente 5. Significa que é cinco vezes mais provável que pacientes que *não* receberam betametasona relatem dor após a instrumentação endodôntica.

Nos estudos prospectivos, o pesquisador acompanha um grupo de pessoas com uma característica específica (por exemplo, hipertensão arterial) e um grupo de pessoas sem essa característica (normotensos) por certo período, à espera da ocorrência de determinado desfecho (por exemplo, AVC). Depois, calcula a proporção de pessoas com o desfecho esperado, em ambos os grupos. Essas proporções são *estimativas de risco.*

Os estudos retrospectivos *não* permitem fazer estimativas de riscos. Nesses estudos, os pesquisadores procuram pessoas com uma doença – por exemplo, úlcera gástrica – e verificam quantas delas estiveram expostas a um fator que presumem de risco – por exemplo, comida apimentada por longo tempo. Depois, procuram pessoas sem a doença e verificam quantas estiveram expostas ao mesmo fator, para, depois, fazer comparações. Veja bem: são relatos históricos das pessoas, não são probabilidades. A situação já aconteceu. Então, não é possível calcular riscos, mas apenas as proporções de pessoas que foram expostas ao fator, entre casos e controles.

Os estudos transversais também *não* permitem fazer estimativas de riscos. Nesses estudos, os pesquisadores verificam, *ao mesmo tempo*, duas variáveis para a mesma pessoa. Por exemplo, o pesquisador verifica o tipo de infração de trânsito cometida e o sexo do motorista.

12.4.1 Razão de chances

12.4.1.1 O que é chance?

Os estudos retrospectivos não admitem estimar riscos, mas permitem estimar *chances*. Vamos entender isso por meio de dados fictícios. Imagine que exista um tratamento não muito eficiente para uma doença com alta taxa de mortalidade. Imagine que tenha sido proposto um novo tratamento. Em um ensaio em que se comparou o novo tratamento com o tratamento convencional, foram obtidos os dados mostrados na Tabela 12.13.

Tabela 12.13

Número de pacientes que morreram e dos que sobreviveram, segundo o tratamento

Tratamento	Morreram	Sobreviveram
Convencional	38	76
Novo	9	90

Com o tratamento convencional, 38 pacientes morreram para um número de 76 que sobreviveram. A chance de morrer é de 38 para 76 com o tratamento convencional. Então,

$$chance\ de\ morrer,\ dado\ convencional = \frac{38}{76} = \frac{1}{2}$$

Isso significa que, para cada paciente submetido ao tratamento convencional que morre, sobrevivem dois.

Com o novo tratamento, nove pacientes morreram para noventa que sobreviveram. Então,

$$chance\ de\ morrer,\ dado\ novo = \frac{9}{90} = \frac{1}{10}$$

Isso significa que, para cada paciente que morre submetido ao novo tratamento, sobrevivem dez.

Para obter a razão de chances, calcule:

$$raz\tilde{a}o\ de\ chance\ s = \frac{\frac{1}{2}}{\frac{1}{10}} = \frac{1 \times 10}{2 \times 1} = 5$$

Mas o que significa essa razão de chances? A chance de o paciente morrer é cinco vezes maior se receber o tratamento convencional em vez do novo. Para cada cinco pacientes que morrem recebendo tratamento convencional, apenas um morre recebendo o novo.

Considerando o evento morte mostrado no exemplo, se a razão de chances for igual a 1, significa que ambos os grupos têm a mesma chance de morrer. Se a razão de chances for maior que 1, significa que o *primeiro grupo tem maior* chance de morrer que o segundo. Se a razão de chances for menor que 1, significa que o *primeiro grupo tem menor* chance de morrer que o segundo, mas o número não é de fácil interpretação. Coloque o grupo que você espera ter maior chance em primeiro lugar.

O uso da razão de chances na área de saúde tem aumentado, mas, para muitos pesquisadores, a interpretação do resultado ainda é difícil. No Brasil, é comum o uso da expressão em inglês – *odds ratio* –, uma vez que os programas de Estatística para computador estão, em sua maioria, em inglês.

Exemplo 12.16 Cálculo da razão de chances

Em 1950, dois pesquisadores ingleses quiseram verificar se o hábito de fumar aumentava o risco de ter câncer do pulmão. Perguntaram, então, os hábitos de fumar dos 649 pacientes que tinham câncer do pulmão e os hábitos de fumar de outros 649 pacientes internados por outros motivos no mesmo hospital. Os dados estão apresentados na Tabela 12.14. Não era possível, para os pesquisadores, estimar riscos, porque os fatos já haviam acontecido (probabilidades referem-se a eventos futuros, nunca a eventos do passado).

Tabela 12.14

Distribuição dos participantes da pesquisa segundo ter ou não câncer de pulmão e ser ou não fumante

Motivo da internação	Hábito de fumar		Total
	Sim	Não	
Câncer no pulmão	27	622	649
Outra razão	2	647	649
Total	29	1.269	1.298

Fonte: Doll, R. e Hill, A.B. Smoking and carcinoma of the lung. *Br Med J 1950 (2)*: 739-48.

Dos pacientes que *tinham câncer de pulmão*, 27 eram fumantes e 622 eram não fumantes. Então, entre os pacientes que *tinham câncer de pulmão*, a chance era de encontrar 27 fumantes para cada 622 não fumantes.

$$chance\ de\ fumante,\ dado\ ter\ Ca = \frac{27}{622}$$

Dos pacientes que *não tinham câncer de pulmão*, havia dois fumantes e 647 pacientes que não fumavam. Logo, entre os pacientes que *não tinham câncer de pulmão*, a chance era de encontrar dois fumantes para cada 647 não fumantes.

$$chance\ de\ fumante,\ sem\ Ca = \frac{2}{647}$$

A razão de chance é:

$$razão\ de\ chances = \frac{\frac{27}{622}}{\frac{2}{247}} = \frac{27 \times 247}{622 \times 2} = 14,04 \cong 14$$

Mas o que significa essa razão de chances? A chance de ter câncer de pulmão é 14 vezes maior para fumantes do que para não fumantes. Para cada 14 fumantes com câncer de pulmão, há um não fumante na mesma condição.

A *razão de chances* também é conhecida como *razão dos produtos cruzados*. É fácil entender essa denominação. Usando os valores literais definidos na Tabela 3.10 (Cap. 3), a razão de chances é dada por

$$razão\ de\ chances = \frac{ad}{bc} = \frac{27 \times 247}{622 \times 2} = 14,04 \cong 14$$

12.5 Teste de uma proporção

As taxas e os coeficientes de prevalência são, basicamente, proporções. Vamos mostrar aqui como se faz um teste estatístico para estabelecer se uma proporção tem um valor especificado. Portanto, o teste também se aplica às taxas e aos coeficientes de prevalência, desde que expressos em proporções (e não por mil ou cem mil indivíduos). Considere, então, que um pesquisador tenha contado o número X de portadores de determinada característica em uma amostra de tamanho n. Pode, então, calcular a *proporção* de portadores dessa característica na amostra, como segue:

$$p = \frac{X}{n}$$

Exemplo 12.17 Obtendo prevalência

Em Campinas, um médico[13] examinou 2.964 recém-nascidos e verificou que 73 apresentavam anomalias, no ano de 1977. Para obter a prevalência de anomalia nessa amostra, divide o número de recém-nascidos que apresentavam anomalia pelo tamanho da amostra. Multiplicando o resultado por 100, obtém a prevalência em porcentagem:

$$p - \frac{73}{2964} \times 100 - 2,46\,\%$$

[13] Arena, JFP. Incidência de malformações em uma população brasileira. *Rev. Paul. Med.* 89 (3/4): 42-9. 1977.

Imagine agora que o pesquisador pretenda testar a hipótese de que a proporção P de portadores com essa característica, na população da qual a amostra proveio, tem o valor θ especificado na literatura. É preciso, então, fazer um *teste estatístico*.

Para proceder a um *teste estatístico*, estabelecem-se as hipóteses e o nível de significância. Depois se calcula a estatística de teste:

$$z = \frac{p - \theta}{\sqrt{\frac{\theta \times (1-\theta)}{n}}}$$

Sob a hipótese da nulidade, a variável z tem, aproximadamente, distribuição normal padronizada, desde que $np > 5$ e $n\ (1 - p) > 5$. Se o valor calculado de z for maior do que o valor dado na tabela de distribuição normal padronizada para o nível estabelecido de significância, deve-se rejeitar a hipótese de que a proporção de portadores da característica em estudo, na população da qual a amostra proveio, tem o valor que foi especificado.

Exemplo 12.18 Comparando a prevalência com o valor especificado

Reveja o Exemplo 12.17: o médico quis testar a hipótese de que a prevalência de recém-nascidos com anomalia em Campinas no ano de 1977 era o valor especificado na literatura internacional, ou seja, 3%. Então, foi preciso estabelecer as hipóteses e o nível de significância:

H_0: a prevalência de recém-nascidos com anomalia em Campinas no ano de 1977 era o valor especificado de 3%;

H_1: a prevalência de recém-nascidos com anomalia em Campinas no ano de 1977 era diferente do valor especificado de 3%;

Nível de significância: 0,05.

A prevalência observada na amostra é:

$$p = \frac{73}{2964} = 0,02464$$

A estatística de teste é:

$$z = \frac{0,02464 - 0,03}{\sqrt{\frac{0,03 \times (1 - 0,03)}{2964}}} = -1,714$$

Como o valor calculado de *z* (–1,714) é, em valor absoluto, menor do que o valor de z com o nível de 5% de significância (1,96, para teste bilateral), não há evidência para rejeitar a hipótese de que a prevalência de recém-nascidos com anomalia na região de Campinas em 1977 era de 3%, compatível com a prevalência citada na literatura internacional.

É recomendável calcular a estatística de teste com *correção de continuidade*, principalmente quando a amostra é pequena. Ao fazer essa correção, a estatística de teste fica como segue:

$$z = \frac{\left(|p - \theta| - \frac{1}{2n}\right)}{\sigma(p)}$$

A correção de continuidade reduz o valor de *z* porque, subtraindo 1/2*n* da diferença entre a proporção observada e a proporção esperada, reduz o numerador.[14] Além disso, o efeito da correção de continuidade sobre a estatística de teste é maior quando a amostra é grande: o valor da estatística diminui com o aumento da amostra.

Exemplo 12.19 Correção de continuidade

Reveja o Exemplo 12.18. O valor da estatística de teste com a correção de continuidade é:

$$z = \frac{\left(|\,0,02463 - 0,03\,| - \frac{1}{2\times 2964}\right)}{0,0031333} = 0,1660,$$

menor do que o valor calculado anteriormente sem a correção de continuidade, uma vez que o tamanho da amostra é bem grande.

12.6 Exercícios resolvidos

12.6.1. O Estudo do Coração de Helsinque (Helsinki Heart Study)[15] mostrou redução na incidência de eventos cardíacos em homens de meia-idade com nível alto de colesterol, mas sem diagnóstico de doença coronariana. Dos 2.051 participantes que, durante cinco anos, receberam uma 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.

a. Qual é a proporção de participantes que registraram evento cardíaco no grupo tratado?

b. Qual é a proporção de participantes que registraram evento cardíaco no grupo placebo?

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,b. Veja a Tabela 12.15.

c. É preciso fazer um teste estatístico. Então:

Tabela 12.15

Participantes da pesquisa segundo o tratamento e o registro ou não de evento cardíaco

Tratamento	Evento cardíaco		Total	Proporção com registro de evento
	Sim	Não		
Droga	56	1.995	2.051	0,0273
Placebo	84	1.946	2.030	0,0414
Total	140	3.941	4.081	

H_0: $P_1 = P_2$

H_1: $P_1 \neq P_2$

Nível de significância: 5%

Calcule a estatística de teste:

$$\chi^2 = \frac{(ad - bc)^2 n}{(a+b)(c+d)(a+c)(b+d)}$$

$$\chi^2 = \frac{(56 \times 1946 - 1995 \times 84)^2 \times 4081}{(56+1995)(84+1946)(56+84)(1995+1946)}$$

$$= \frac{(-58604)^2 \times 4081}{(2051)\,(2030)\,(140)\,(3941)} = 6,10$$

H_0 deve ser rejeitada com o nível de 5% de significância; temos, portanto, a evidência de que a droga surtiu efeito.

d. Faça a diferença entre as duas proporções e divida pela proporção do grupo que recebeu placebo. Multiplique por 100, para obter a diferença em relação ao placebo expressa em porcentagem.

$$\frac{0,0414 - 0,0273}{0,0414} \times 100 = 34\,\%$$

O uso da droga reduziu a incidência de eventos cardíacos em 34%.

12.6.2. Foi elaborado um questionário para comparar a sexualidade de pacientes jovens com doença de Parkinson com a sexualidade de controles sadios.[16] As respostas para

uma das questões, que avaliou o sentimento de solidão, são apresentadas na Tabela 12.16. Construa uma tabela para apresentar a proporção de pessoas que relatam sentir solidão, em ambos os grupos. Compare com o teste estatístico.

Tabela 12.16

Pacientes que relatam sentir solidão segundo o grupo

Grupo	Sentem solidão		
	Sim	Não	Total
Parkinsoniano	56	65	121
Controle sadio	23	103	126
Total	79	168	247

Tabela 12.17

Proporções obtidas por estudo transversal

Grupo	Amostra	Sentem solidão
Parkinsoniano	121	0,463
Controle sadio	126	0,183
Total	247	0,320

H_0: a probabilidade de sentir solidão é a mesma para um jovem parkinsoniano e um jovem sadio, ou seja, $P_1 = P_2$;

H_1: a probabilidade de sentir solidão é maior para um jovem parkinsoniano do que para um jovem sadio, ou seja, $P_2 > P_1$;

Nível de significância: 0,05.

$$z = \frac{\left|0,463 - 0,183\right| - \frac{1}{2}\left(\frac{1}{121} + \frac{1}{126}\right)}{\sqrt{0,320 \times 0,680 \times \left(\frac{1}{121} + \frac{1}{126}\right)}} = 4,58$$

Como o valor calculado de *z* (4,58) é maior do que o valor de z com o nível de 5% de significância (2,54, para teste unilateral), rejeita-se H_0. Logo, a conclusão da pesquisa é a de que parkinsonianos jovens sentem mais solidão do que jovens sadios.

12.6.3. Realizou-se um estudo[17] com 263 adolescentes que aparentavam comportamento suicida. Eles fizeram avaliação psiquiátrica e foram acompanhados durante seis meses. Desse grupo, 86 adolescentes foram avaliados como apresentando comportamento suicida, embora sem depressão no início do estudo. Dos 77 jovens com comportamento suicida persistente no *follow-up*, 45 foram avaliados como apresentando depressão no início do estudo. Cem jovens não apresentavam nem depressão nem comportamento suicida. A) Construa uma tabela para apresentar os dados; B) calcule a razão de chances; C) interprete.

Em primeiro lugar, é preciso obter os valores de *b* e *d*. Veja em seguida.

Comportamento suicida	Depressão		Total
	Sim	Não	
Sim	$a = 45$	$b = 86$	$a + b = ?$
Não	$c = ?$	$d = ?$	$c + d = ?$
Total	$a + c = 77$	$b + d = ?$	263

$a + b = 131$
$c = 77 - 45 = 32$
$c + d = 263 - 131 = 132$
$d = 132 - 32 = 100$

Agora, é preciso construir a Tabela 12.18.

A razão de chances é:

Tabela 12.18

Depressão como fator de risco para comportamento suicida

Comportamento suicida	Depressão		Total
	Sim	Não	
Sim	45	86	131
Não	32	100	132
Total	77	186	263

$$OR = \frac{45 \times 100}{32 \times 86} = \frac{4500}{2752} = 1,63$$

Usando a razão de chances como estimativa de risco, podemos dizer que é 1,63 vez mais provável que um adolescente com depressão apresente comportamento suicida do que o adolescente que não tem depressão.

---

[15]Marshall, 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.

[16]Jacobs, H.; Vieregge, A.; Vieregge, P. Sexuality in young patients with Parkinson's disease: a population based comparison with healthy controls. *Neurol Neurosurg Psychiatry* 2000; 550-552 doi:10.1136/jnnp.69.4.550.

[17]Greenfield B., Henry M., Weiss M., Tse S. M., Guile J. M., Dougherty G., Zhang X., Fombonne E., Lis E., Lapalme-Remis, Harnden, B. Previously suicidal adolescents: Predictors of six-month outcome. *Journal of the Canadian Association of Child and Adolescent Psychiatry*. 2008;17(4):197-201. [PMC free article] [PubMed].

12.7 Exercícios propostos

12.7.1. A proporção de recém-nascidos com defeito ou doença séria é de 3%. Imagine que um médico suspeite que essa proporção tenha aumentado. Então, examinou 1.000 recém-nascidos e encontrou 34 com defeito ou doença séria. Você acha que a suspeita do médico é procedente?

12.7.2. Com base nos dados apresentados na Tabela 12.20, com o nível de significância de 5%, teste a hipótese de que a proporção de recém-nascidos vivos portadores de anomalia é a mesma em ambos os sexos.

Tabela 12.20

Recém-nascidos vivos segundo o sexo e a presença ou não de anomalia

Sexo	Anomalia	
	Sim	Não
Masculino	28	1.485
Feminino	45	1.406

Fonte: Arena, J. F. P. Incidência de malformações em uma população brasileira. *Rev. Paul. Med*. 89 (3/4):42-9, 1977.

12.7.3. Com base nos dados apresentados na Tabela 12.21, teste, com o nível de significância de 1%, a hipótese de que a ausência congênita de dentes independe do sexo.

Tabela 12.21

Escolares segundo o sexo e a ausência congênita de dentes

Sexo	Ausência congênita de dentes	
	Sim	Não
Masculino	23	1.078
Feminino	40	859

Fonte: Vedovelo Filho, M. Prevalência de agenesias dentárias em escolares de Piracicaba, 1972. [[Tese (mestrado)] FOP-Unicamp].

12.7.4. 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 um coeficiente de associação para os dados do Exercício 12.7.3. Você considera grande a associação?

12.7.5. Com base nos dados apresentados na Tabela 12.22, calcule o coeficiente de associação. Faça o teste de qui-quadrado.

Tabela 12.22

Resultados de casos de diagnóstico pré-natal segundo a idade da gestante e a presença ou a ausência de aberração cromossômica

Sexo	Aberração cromossômica	
	Sim	Não
De 35 até 40 anos	10	447
40 anos ou mais	18	510

12.7.6. Para determinar se existe associação entre implantes mamários e doenças do tecido conjuntivo e outras doenças,[18] foram observadas, durante vários anos, 749 mulheres que haviam recebido implante e exatamente o dobro de mulheres que não haviam recebido implante. Os pesquisadores, então, verificaram que cinco mulheres que receberam implantes e dez das que não receberam tiveram doenças do tecido

conjuntivo. Quais são as hipóteses em teste? Quais são as proporções de mulheres doentes, em ambos os grupos?

12.7.7. Com base nos dados apresentados na Tabela 12.23, você rejeita a hipótese de que a probabilidade de natimorto é a mesma para ambos os sexos?

Tabela 12.23

Recém-nascidos segundo o sexo e a condição de vivo ou natimorto

Sexo	Condição	
	Vivo	Natimorto
Masculino	1513	37
Feminino	1451	27

Fonte: Arena, J. F. P. Incidência de malformações em uma população brasileira. *Rev. Paul. Med*. 89 (3/4):42-9, 1977.

12.7.8. Com base nos dados apresentados na Tabela 12.24, obtenha o coeficiente de associação. O que significa?

Tabela 12.24

Recém-nascidos segundo a idade materna e o tempo de gestação

Idade materna	Tempo de gestação		Total
	Até 36 semanas	De 37 a 41 semanas	
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, G. D. *et al*. Efeito da idade materna sobre os resultados perinatais. *RBGO 24* (3): 2002.

12.7.9. Com base nos dados apresentados na Tabela 12.25, você rejeita a hipótese de que a probabilidade de dormir mais de oito horas é a mesma para as duas faixas etárias?

Tabela 12.25

Participantes da pesquisa segundo o tempo de sono, em horas, e a faixa etária

Faixa etária	Tempo de sono	
	Menos de 8 horas	8 horas ou mais
De 30 a 40 anos	172	78
De 60 a 70 anos	120	130

12.7.10. Com base nos dados apresentados na Tabela 12.26, você rejeita a hipótese de que a probabilidade de ter gripe é a mesma para pessoas vacinadas e não vacinadas?

Tabela 12.26

Participantes da pesquisa segundo o fato de ter sido vacinada contra gripe e ter tido gripe

Vacina	Gripe	
	Sim	Não
Sim	11	538
Não	70	464

[18]Gabriel, S. E. *et al*. Risk of connective tissues diseases and other disorders after breast implantation. *New Engl J Med 330*:1.697-1.702, 1994. Apud: Motulsky, H. *Intuitive Biostatistics*. Nova York: Oxford University Press, 1995, p. 318.

[1]Veja Vieira, S. e Hossne, WS. *Metodologia científica para a área da saúde*. 2 ed. Rio de Janeiro: Elsevier, 2015.

[2]Alguns programas de computador dão o valor de χ^2 com e sem correção de continuidade. É preciso optar por um deles.

[3]Nem sempre se faz a correção de continuidade, embora seja teoricamente recomendada. De qualquer forma, o uso da correção diminui a probabilidade de

encontrar valor significante.

[4]Esse valor, porém, só ocorre quando as amostras são de mesmo tamanho.

[5]Veja Fleiss, JL. *Statistical methods for rates and proportions*. Nova York: Wiley, 1981, p. 60.

[6]O coeficiente γ também é conhecido como coeficiente de Yule.

[7]Veja a metodologia em: Vieira, S. e Hossne, WS. *Metodologia científica para a área da saúde*. Rio de Janeiro: Elsevier, 2015.

[8]Fleiss, J L. Statistical methods for rates and proportions. Nova York: Wiley, 1981.

[9]Veja mais sobre esses estudos em Vieira, S. e Hossne, WS. *Metodologia científica para a área da saúde*. 2 ed. Rio de Janeiro: Elsevier, 2015.

[11]Fleiss, J L. Statistical methods for rates and proportions. Nova York: Wiley, 1981.

[12]Fleiss, J. L. Statistical methods for rates and proportions. Nova York: Wiley, 1981.

[14]A correção de continuidade, embora teoricamente recomendada, nem sempre é feita. De qualquer forma, o uso da correção diminui a probabilidade de encontrar valor significante.

Apêndices

ESBOÇO

Apêndice Capítulo 13: Probabilidades

Apêndice Capítulo 14: Distribuição Binomial

APÊNDICE

CAPÍTULO 13

Probabilidades

Lidamos com ideias sobre probabilidade em nosso dia a dia. Dizemos, por exemplo: “É provável que chova amanhã” ou “Carlos provavelmente chega hoje”. Mas também calculamos probabilidades. Quando alguém pergunta qual é a probabilidade de sair cara no jogo de moeda, a resposta é fácil: ½ ou 50%. Como encontramos essa probabilidade? Pensamos assim: quando uma moeda é lançada, pode sair tanto cara quanto coroa; as duas faces não podem ocorrer ao mesmo tempo, mas têm a mesma chance. Portanto, cara ocorre na metade vezes. Mas será que, se você jogar uma moeda duas vezes, é *certo* que sairá cara uma das vezes? Claro que não. Quando dizemos que a probabilidade de sair cara num jogo de moeda é ½, estamos apenas afirmando que, se uma moeda for lançada *um grande número de vezes*, *espera-se* que ocorra cara na metade delas.

13.1 A linguagem para o estudo de probabilidades

O estudo de probabilidades tem muita aplicação em todas as ciências, mas começou com os jogos de azar. As pessoas queriam entender a "lei" que rege esses jogos para ganharem dinheiro nos cassinos.[1] E os matemáticos acabaram estabelecendo a teoria das probabilidades, que trata dos fenômenos aleatórios. Muitos fenômenos têm padrão de comportamento *previsível no longo prazo*, mas comportamento imprevisível quando observados por pouco tempo. Lembre-se de que você *não sabe*, quando joga uma moeda, se sairá cara ou coroa. No entanto, *pode prever* que, em mil lançamentos, ocorrerá cara em cerca de metade das vezes. As ocorrências possíveis em dado fenômeno aleatório são até bem conhecidas.

Espaço amostral é o conjunto dos resultados possíveis de um fenômeno aleatório. Para um lançamento de moeda, o espaço amostral, que indicaremos por E, é cara e coroa. Escrevemos:

$$E = \{\text{cara e coroa}\}$$

Evento é qualquer subconjunto do espaço amostral. Diversos resultados podem constituir o evento de interesse. Por exemplo, imagine um jogo em que se lançam duas moedas e o jogador ganha se a mesma face ocorrer em ambas. O espaço amostral é

$$E = \{\text{cara e cara, cara e coroa, coroa e cara, coroa e coroa}\}$$

O jogador ganha se ocorrer qualquer um dos dois resultados do evento A:

$$A = \{\text{cara e cara, coroa e coroa}\}$$

Dado o evento *A*, denomina-se o *complemento de A*, que se indica por A^c, o conjunto de eventos que *não são A*.

No exemplo que acabamos de ver:

$$A^c = \{\text{cara e coroa, coroa e cara}\}$$

Dois eventos são chamados de *mutuamente exclusivos* quando não têm elementos em comum. É o caso das pesquisas de opinião em que o entrevistador deve buscar grupos que são definidos por características excludentes – quem está em determinado grupo não pode estar em outro. Por exemplo:

1. *A* = adultos, com idade de 18 a 60 anos
2. *B* = idosos, com mais de 60 anos

Às vezes, estamos interessados em eventos que não são simples. Quando interessam tanto o evento *A* como o evento *B*, ou seja, *A* ou *B*, dizemos estar interessados na *união de A* e *B*, matematicamente indicada por $A \cup B$. Por exemplo, quando você diz que aceita sorvete de creme *ou* de chocolate, significa que aceita qualquer um deles: um ou outro.

Quando interessam os resultados que sejam simultaneamente evento *A* e evento *B*, dizemos estar interessados na *intersecção A* e *B*, matematicamente indicada por $A \cap B$. A ideia de *d*ois *eventos que ocorrem juntos* é expressa pela conjunção "e". Por exemplo, quando o entrevistador pergunta a um morador da cidade de São Paulo se tem moto e é favorável à implantação de ciclovias, pode estar interessado na interseção dos eventos:

$$A = \{\text{ter moto}\} \text{ e } B = \{\text{ser favorável ás ciclovias}\}$$
$$A \cap B = \{\text{ter moto } e \text{ ser favorável ás ciclovias}\}$$

Dois eventos são *independentes* se a ocorrência de um deles não tem influência na ocorrência do outro. Lembre-se do exemplo dado anteriormente, do jogo em que se lançam duas moedas e o jogador ganha se a mesma face ocorrer em ambas: a ocorrência de determinada face em uma das moedas não tem qualquer efeito sobre o que ocorre na outra moeda.

É importante considerar aqui o risco 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 ½, mas, dada a condição de que ocorreu coroa, é zero. Então, a probabilidade de sair cara muda se sair coroa.

Eventos são indicados pelas primeiras letras do alfabeto, escritas em itálico: *A*, *B*, *C* etc. Muitas vezes, o espaço amostral e os eventos são apresentados em diagrama de Venn. Para desenhar esse diagrama, você traça um retângulo que representará o espaço amostral e, dentro do retângulo, círculos que representarão os eventos. Veja a Figura 13.1.

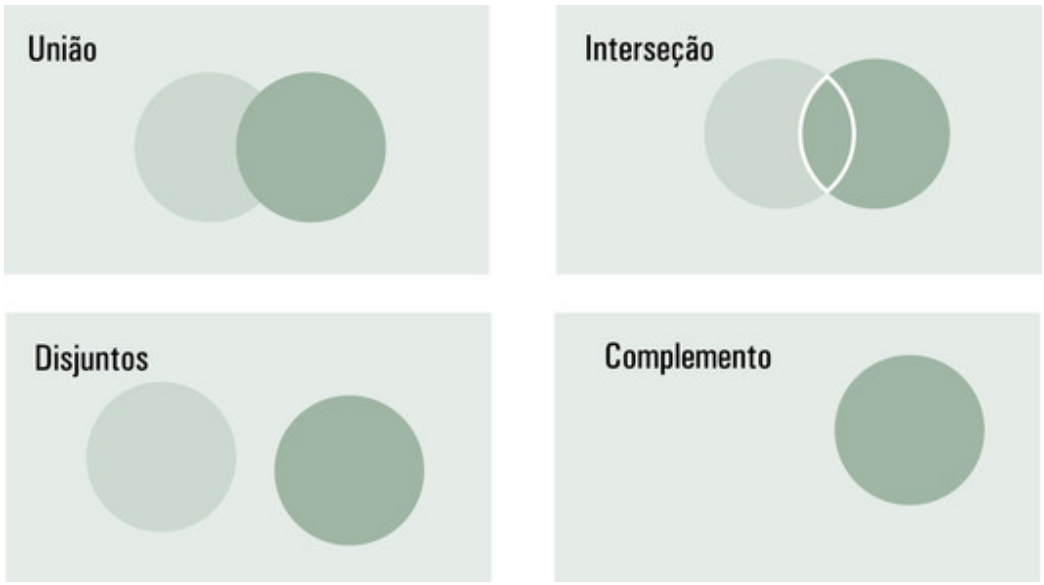


FIGURA 13.1 Diagramas de Venn

13.2 Definições de probabilidade

13.2.1 Definição frequentista de probabilidade

PROBABILIDADE de ocorrer um evento com a característica A, indicada por P(A), é dada pela frequência relativa desse evento em uma série de n observações feitas sob as mesmas condições.

$$P(A) = \frac{m}{n}$$

Exemplo 13.1

Um médico[2] verificou que, de 2.964 nascidos vivos, 73 tinham algum defeito ou uma doença séria. Com base nessa amostra, a estimativa da probabilidade de um recém-nascido ter defeito ou doença séria é

$$\frac{73}{2964} = 0,0246$$

[2]Arena, J. F. P. Estudo clínico-epidemiológico prospectivo das anomalias congênitas na população de Campinas, 1977. Tese (Doutorado) – FCM, Unicamp, Campinas.

A palavra *probabilidade* é entendida neste texto como uma proporção, ou seja, o número de vezes em que um evento ocorre dividido pelo número de vezes em que o processo é repetido nas mesmas condições – muitas e muitas vezes.

13.2.1.1 Regras a que as definições de probabilidade devem obedecer

1. Probabilidade é um valor numérico que varia entre zero e 1, inclusive.[3] *Eventos impossíveis* têm probabilidade zero, enquanto eventos *certos* têm probabilidade 1.
2. A soma das probabilidades de todos os eventos possíveis é igual a 1.
3. A probabilidade de um evento é igual a 1 menos a probabilidade de esse evento não ocorrer.

Exemplo 13.2

Evento *certo*: a probabilidade de que qualquer um de nós venha a morrer um dia é 1. Evento *impossí*vel: a probabilidade de que qualquer um de nós seja imortal é zero.

A definição de probabilidade que acabamos de ver, chamada por muitos de *definição frequentista*, é aplicada às situações que podem ser pensadas como repetíveis sob condições específicas, no mundo das ciências. Tiramos amostras da população para *ter dados que permitam estimar probabilidades*.

Na área de saúde, as probabilidades de danos e eventos adversos são referidas como *riscos*. Muitos estudos já foram feitos para estimar o risco de um fumante ter câncer do pulmão, de sobreviver a um acidente de carro ou de um nascituro ser menino. O Exemplo 13.3 estima o risco de ocorrer erro médico em um hospital, em determinado período limitado, em condições específicas (por exemplo, mantidos o mesmo equipamento e a mesma equipe).

Exemplo 13.3

Numa amostra de 30.195 registros hospitalares selecionados ao acaso, foram identificados 1.133 pacientes com lesões sérias causadas por imprudência, negligência ou imperícia do médico.[4] **O risco estimado** de lesão séria por erro médico nesse hospital é

$$\frac{1.133}{30.195} = 0,0375$$

[4]Leape, L. *et al*. The nature of adverse events in hospitalized patients: Results of the Harvard Medical Practice Study II. *The New England Journal of Medicine*, v. 324, n. 6, Feb. 7, 1991.

É comum que as pessoas pensem em probabilidades como porcentagens. Os estatísticos preferem sempre expressar valores de probabilidade por números entre zero e 1 porque, em cálculos mais avançados, isso se faz necessário. Mas, se você quiser expressar probabilidade em porcentagem, basta multiplicar o valor dado pela definição por 100 e acrescentar o símbolo de porcentagem (%) ao resultado. Aliás, na prática, as probabilidades são mais bem-compreendidas quando expressas em porcentagem.

Exemplo 13.4

No Exemplo 13.3, foi estimada a probabilidade de lesão séria por erro médico em determinado hospital:

$$\frac{1.133}{30.195} = 0,0375$$

Para ser dada em porcentagem, essa estimativa é multiplicada por 100. Em porcentagem, a estimativa do risco de lesão séria por erro médico nesse hospital é de 3,75%, expressão mais facilmente entendida.

13.2.2 Definição clássica de probabilidade

A definição frequentista de probabilidade atende bem ao conhecimento da área de saúde quando o pesquisador quer estimar riscos. Por meio de observações de muitos casos, é possível estimar o risco de efeitos adversos. Mas é preciso que o número de eventos observados possa crescer indefinidamente. Quando o espaço amostral contém um número finito de eventos contáveis – desde que igualmente prováveis –, é fácil usar a definição clássica.

DEFINIÇÃO CLÁSSICA: Se forem possíveis n resultados mutuamente exclusivos e igualmente prováveis, se m desses resultados forem favoráveis, a probabilidade de resultado favorável é

$$P(A) = \frac{n^{o}\ de\ resultados\ favor\acute{a}veis}{n^{o}\ de\ resultados\ poss\acute{i}veis} = \frac{m}{n}$$

Exemplo 13.5 Cálculo de probabilidade

Qual é a probabilidade de ocorrer face 6 quando se joga um dado? Os $n = 6$ resultados possíveis compõem o espaço amostral:

$S = \{1, 2, 3, 4, 5, 6\}$.

Só um resultado ($m = 1$) atende à característica pedida: face 6. Então, a probabilidade de ocorrer 6 é:

$$P(6) = \frac{1}{6} = 0,1667$$

13.2.3 Definição de probabilidade subjetiva

É impossível encaixar, dentro da ideia de probabilidade, afirmativas como "a probabilidade de o Brasil ganhar a próxima Copa Mundial de Futebol é 0,80". Nesses casos, é preciso usar a definição subjetiva de probabilidade.

PROBABILIDADE SUBJETIVA é um valor entre zero e 1 que representa um ponto de vista pessoal sobre a possibilidade de ocorrer determinado evento.

É importante entender que probabilidade subjetiva não é apenas uma forma de pensar logicamente sobre fenômenos aleatórios. É a maneira como uma pessoa descreve seu grau de crença em determinado desfecho. É, portanto, racional, embora não se baseie em técnicas computacionais. E tem sentido quando fornecida por alguém que conhece o assunto. Logo, probabilidade subjetiva é de enorme importância quando as informações são apenas parciais e é preciso ter intuição.

A grande *desvantagem* da definição subjetiva de probabilidade é o fato de ser pessoal. Em função disso, nos casos em que a frequência relativa pode ser calculada, a probabilidade subjetiva pode não ter relação alguma com os resultados realmente obtidos. Mas a probabilidade subjetiva predomina nas decisões administrativas, nas aplicações financeiras e nos jogos de azar.

13.3 Teorema da soma ou a regra do ou

13.3.1 Regra 1 da soma: eventos mutuamente exclusivos

Se A e B são *eventos mutuamente exclusivos*, a probabilidade de ocorrer A ou B é igual à soma das probabilidades de ocorrer cada um deles. Escreve-se:

$$P(A \cup B) = P(A) + P(B)$$

Exemplo 13.6 Soma de eventos mutuamente exclusivos

Foi feito um estudo de caso-controle com pacientes hospitalizados (7.804 casos e 15.207 controles) para determinar os fatores de risco de câncer do pulmão.[5] Os dados apresentados na Tabela 13.1 foram obtidos para saber se o risco de câncer do pulmão aumenta com o número de cigarros fumados por dia. Qual é a probabilidade de uma pessoa, tomada ao acaso dessa amostra, fumar um maço de cigarros (20) ou mais por dia?

Tabela 13.1

Distribuição de casos e controles segundo o número de cigarros fumados por dia

Nº de cigarros/dia	Casos	Controles	Total	Risco
Nenhum	164	2.616	2.780	0,059
De 1 a 9	664	2.194	2.858	0,232
De 10 a 19	1.704	3.385	5.089	0,335
De 20 a 29	2.127	3.108	5.235	0,406
30 ou mais	1.369	1.746	3.115	0,439
	6.028	13.049	19.077	

A probabilidade de uma pessoa, tomada ao acaso, fumar um maço de cigarros (20) ou mais por dia é dada, usando os dados da Tabela 13.1, pela probabilidade de fumar de 20 a 29 cigarros por dia, somada a probabilidade de fumar 30 cigarros ou mais por dia.

$$P(de\ 20\ a\ 29) = \frac{5.235}{19.077} = 0,274$$

$$P(30\ ou\ mais) = \frac{3.115}{19.077} = 0,163$$

A probabilidade de a pessoa fumar um maço ou mais de cigarros por dia, nessa amostra, é:

$$P = 0,274 + 0,163 = 0,437$$

[5]Assessment of Lung Cancer Risk Factors by Histologic Category1 *JNCI*, v. 73, n. 2, agosto de 1984.

13.3.2 Regra 2 da soma: eventos não mutuamente exclusivos

Se os eventos não são mutuamente exclusivos, ou seja, se A e B podem ocorrer ao mesmo tempo, a probabilidade de ocorrer A ou B é dada pela probabilidade de A, mais a probabilidade de B, menos a probabilidade de A e B. Escreve-se:

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

É preciso subtrair o conjunto interseção porque, quando somamos $P(A) + P(B)$, a probabilidade do conjunto interseção $P(A \cap B)$ é somada duas vezes. No caso de eventos *mutuamente exclusivos*, não se faz a subtração, porque a probabilidade de os eventos ocorrerem ao mesmo tempo é zero. Veja o diagrama da Figura 13.1: eventos mutuamente exclusivos não têm interseção.

Exemplo 13.7 Soma de eventos não mutuamente exclusivos

Foi feito um estudo de caso-controle (299 casos e 292 controles) para determinar os fatores de risco para infarto do miocárdio. Os dados da Tabela 13.2 foram obtidos para saber se pacientes diabéticos apresentam maior risco de infarto do miocárdio. Qual é a probabilidade de uma pessoa, tomada ao acaso dessa amostra, ser ou diabética ou infartada?

Tabela 13.2

Distribuição dos casos de infarto e controles, segundo a presença ou não de diabetes

Diabetes	Infartados		Total
	Casos	Controles	
Sim	59	29	88
Não	240	263	503
Total	299	292	591

Fonte: Silva, MAD; Sousa, AGMR; Schargodsky, H. Fatores de Risco para Infarto do Miocárdio no Brasil. *Arq Bras Cardiol*, v. 71 (n° 5), 667-675, 1998.

Probabilidade de ter tido infarto:

$$P(infartado) = \frac{299}{591} = 0,506$$

Probabilidade de ser diabético:

$$P(diab\acute{e}tico) = \frac{88}{591} = 0,149$$

Veja que as pessoas que tiveram infarto e são diabéticas estão no conjunto interseção e, portanto, foram consideradas nos dois cálculos. Então

$$P(infartado \cap diab\acute{e}tico) = \frac{59}{591} = 0,0998$$

Probabilidade de ter tido infarto e ser diabético:

$$P(infartado \cup diab\acute{e}tico) = \frac{299}{591} + \frac{88}{591} - \frac{59}{591} = \frac{328}{591} = 0,555$$

13.4 Teorema da multiplicação ou a regra do e

Antes de estudar o teorema da multiplicação, é importante entender bem a questão da independência de eventos. Já vimos que dois eventos, *A* e *B*, são *independentes* se a ocorrência de um deles (*A* ou *B*) não tem efeito sobre a ocorrência do outro (*B* ou *A*). Por exemplo, quando se joga uma moeda duas vezes, o resultado da primeira jogada não tem qualquer efeito sobre o resultado da segunda. São eventos independentes.

13.4.1 Regra 1 da multiplicação: eventos independentes

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:

$$P(A \cap B) = P(A) \times P(B)$$

Exemplo 13.8 Ocorrência conjunta de eventos independentes

Qual é a probabilidade de ocorrerem duas caras quando se joga uma moeda duas vezes? Veja a Tabela 13.3.

Tabela 13.3

Resultados de dois lançamentos de uma moeda

2ª moeda	1ª moeda	
	Cara	Coroa
Cara	Cara; cara	Coroa; cara
Coroa	Cara; coroa	Coroa; coroa

A probabilidade de ocorrer cara na primeira jogada é:

$$P(cara\ 1^{a}\ moeda) = \frac{1}{2} = 0,5$$

O fato de ter ocorrido cara na primeira jogada *não modifica* a probabilidade de ocorrer cara na segunda jogada (os eventos são independentes). Então, a probabilidade de ocorrer cara na segunda jogada é:

$$P(cara\ 2^{a}\ moeda) = \frac{1}{2} = 0,5$$

Para obter a probabilidade de ocorrer cara nas duas jogadas (primeira e segunda), faz-se o produto:

$$P(cara \cap cara) = \frac{1}{2} \times \frac{1}{2} = 0,25$$

Na vida real, encontramos muitos exemplos de eventos independentes como o que vimos, ou seja, "sair cara no primeiro lançamento de uma moeda" e "sair cara no segundo lançamento da mesma moeda". Por exemplo, "chover hoje" e "ser feriado amanhã" são eventos independentes, porque o fato de "chover hoje" não muda a possibilidade de "ser feriado amanhã", nem o fato de "ser feriado amanhã" altera a possibilidade de "chover hoje". No entanto, a ocorrência de certos eventos tem efeito sobre a ocorrência de outros. Por exemplo, "estar alcoolizado" aumenta a probabilidade de "provocar acidente de trânsito". "Vida sedentária" aumenta a probabilidade de "sobrepeso". Dizemos que esses eventos são *dependentes*. Portanto, dois eventos A e B são *dependentes* quando a ocorrência de um deles (por exemplo, a ocorrência de A) *modifica* a probabilidade de o outro ocorrer (no caso, de B).

13.4.2 Regra 2 da multiplicação: eventos dependentes

Antes de estudar a regra 2 da multiplicação, vamos entender por que alguns eventos estão condicionados a outros. Denomina-se *probabilidade condicional* de B dado A a probabilidade de ocorrer o evento B sob a condição de A ter ocorrido. Escreve-se $P(B|A)$, que se lê "probabilidade de B dado A". Pense: você só entra no cinema se comprar a entrada – então, comprar entrada é *condição* para entrar no cinema.

Exemplo 13.9 Probabilidade condicional

Um casal tem dois filhos. a) Qual é a probabilidade de os dois serem meninos? b) Qual é a probabilidade de os dois serem meninos, dado que o primeiro é menino?

Para obter a probabilidade de os dois serem meninos, você calcula:

$$P(menino \cap menino) = \frac{1}{2} \times \frac{1}{2} = 0,25$$

No entanto, quando se pergunta a probabilidade de os dois serem meninos dado que o primeiro é menino, você calcula:

$$P(menino) = \frac{1}{2} = 0,5$$

De acordo com a regra 2 da multiplicação, se A e B são *eventos dependentes*, a probabilidade de ocorrer A e B é dada pela probabilidade de ocorrer A multiplicada pela probabilidade de ocorrer B dado que A ocorreu (essa probabilidade é condicional). Escreve-se:

$$P(A \cap B) = P(A) \times P(B \mid A)$$

Exemplo 13.10 Ocorrência conjunta de eventos dependentes

Uma caixa contém duas bolas brancas e três bolas azuis. Duas bolas são retiradas ao acaso, uma em seguida da outra e sem que a primeira tenha sido recolocada. Qual é a probabilidade de que as duas sejam brancas?

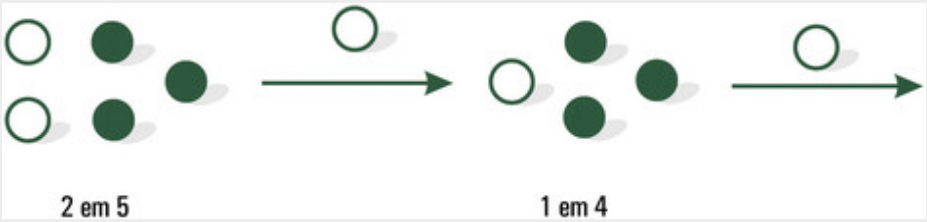


FIGURA 13.2 Retirada de duas bolas brancas, sem reposição

A caixa contém cinco bolas: duas são brancas. Então, a probabilidade de a primeira bola retirada ser branca é

$$P(branca) = \frac{2}{5}$$

Como a bola retirada não foi recolocada, restam quatro bolas na caixa. Para que as duas bolas retiradas da urna sejam brancas, é preciso que a primeira bola retirada seja branca. *Dado que primeira bola retirada era branca*, das quatro bolas que estão na caixa, uma é branca. A probabilidade (condicional) de a segunda bola retirada ser branca é:

$$P(branca \mid branca) = \frac{1}{4}$$

A probabilidade de as duas bolas retiradas serem brancas é dada pelo produto:

$$P(branca\ e\ branca) = \frac{2}{5} \times \frac{1}{4} = \frac{2}{20} = \frac{1}{10}$$

13.4.3 Condição de independência

No dia a dia, muitas vezes dizemos "uma coisa não tem nada a ver com a outra". Em linguagem técnica, queremos dizer que os eventos são *independentes*. O Exemplo 13.8 ilustra a *condição de independência*: quando se jogam duas moedas, o resultado da primeira não influencia o resultado da segunda. Então, dois eventos são independentes se a probabilidade de ocorrerem juntos for igual ao produto das probabilidades de que ocorram em separado, uma vez que a ocorrência de um deles em nada ajuda a ocorrência do outro. Essa é a *condição de independência* de dois eventos. Escreve-se:

$$P(A \cap B) = P(A) \times P(B)$$

Aprendemos que a probabilidade de ocorrer determinado evento *depende*, muitas vezes, das condições em que ocorre esse evento. Isso é conhecido na área de saúde e é importante para a prevenção: a probabilidade de câncer do pulmão depende de ter ou não o hábito de fumar; a probabilidade de ter algumas doenças depende de ter ou não sido imunizado; a probabilidade de ocorrer um acidente automobilístico depende das condições dos pneus. Outras vezes, a probabilidade de ocorrer determinado evento *não depende* da ocorrência de outro. Por exemplo, a probabilidade de ter cárie dentária não depende de a pessoa ser ou não míope; a probabilidade de ter cálculos renais não depende da profissão; a probabilidade de ser calvo não depende do estado civil.

Muitas pesquisas são realizadas para estudar se há ou não dependência entre determinados eventos, o que significa buscar os *fatores que modificam as probabilidades*. Veja um exemplo em que o valor de probabilidade não se modifica em dada condição.

Exemplo 13.11 Condição de independência

Para determinar se existe associação entre implantes mamários e doenças do tecido conjuntivo e outras doenças,[6] durante vários anos foram observadas 749 mulheres que haviam recebido implante e 1.498 que não haviam recebido implante. Verificou-se que cinco das mulheres que haviam recebido implantes e dez das que não haviam recebido implante tiveram doenças do tecido conjuntivo. Você acha que ter doenças do tecido conjuntivo depende ou não de a mulher ter implantes mamários?

A Tabela 13.3 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:

$$\frac{749}{2247}$$

A Tabela 13.4 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:

Tabela 13.4

Distribuição de mulheres com implante mamário e o fato de terem ou não doenças do tecido conjuntivo e outras

Implante mamário	Doenças do tecido conjuntivo e outras		Total	Proporção daquelas que receberam implante mamário
	Sim	Não		
Sim	5	744	749	
Não	10	1.488	1.498	
Total	15	2.232	2.247	
Proporção de mulheres que tiveram doença				

$$\frac{15}{2247}$$

Como 5 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:

$$\frac{5}{2247}$$

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:

$$P(implante \cap doen\,ç\,a) = P(implante) \times P(doen\,ç\,a)$$

[6]Gabriel SE *et al*. Risk of connective tissues diseases and other disorders after breast implantation. *New Engl J Med* 330:1.697-1.702, 1994. Apud: Motulsky, H. *Intuitive Biostatistics*. Nova York: Oxford University Press, 1995, p. 318.

13.5 Exercícios resolvidos

13.5.1. De uma classe com trinta alunos, dos quais 14 são meninos, um aluno é escolhido ao acaso para apresentar um trabalho. Qual é a probabilidade de: a) o aluno escolhido ser um menino? b) o aluno escolhido ser uma menina?

A classe tem trinta alunos ($n = 30$) e todos têm a mesma probabilidade de ser 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.

13.5.2. Uma pessoa comprou um número de rifa que tem cem números e irá sortear cinco prêmios. Qual é a probabilidade de essa pessoa: a) ganhar um prêmio? b) de não ganhar?

Todos os cem números ($n = 100$) da rifa têm igual probabilidade de serem sorteados. Serão sorteados cinco 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.

13.5.3. Uma urna tem dez 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 dez 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.

13.5.4. Joga-se um dado. Qual é a probabilidade de sair: a) o número 3? b) um número maior do que 3? c) um número menor do que 3? d) um número par?

Quando se joga um dado, pode ocorrer um dos seguintes 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 um número maior do que 3 é ½;

c) dos seis eventos, dois ($m = 2$) são menores do que 3 (1; 2). Então, a probabilidade de ocorrer um 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 um número par é ½.

13.5.5. 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 13.5.

Tabela 13.5

Resultados possíveis no jogo de duas moedas

Evento	1ª moeda	2ª moeda
1	Cara	Coroa
2	Coroa	Cara
3	Cara	Cara
4	Coroa	Coroa

A Tabela 13.5 mostra $n = 4$ eventos mutuamente exclusivos e igualmente prováveis. A probabilidade de saírem:

a) duas caras (evento 3 na tabela) é ¼;

b) duas coroas (evento 4 na tabela) é ¼;

c) uma cara e uma coroa (eventos 1 e 2 na tabela) é 2/4.

13.5.6. Em uma família com três filhos, qual é a probabilidade de os três serem homens? Suponha que meninos e meninas tenham a mesma probabilidade de nascer.

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:

$$\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}$$

13.5.7. 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 13.6.

Tabela 13.6

Resultados possíveis no jogo de duas moedas

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.

13.5.8. 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 ½ e que o sexo do segundo filho não depende do sexo do primeiro. Então:

a) a probabilidade de o primogênito ser homem é ½;

b) a probabilidade de os dois filhos serem homens pode ser obtida pelo teorema do produto (de 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):

13.5.9. No cruzamento de ervilhas amarelas homozigotas (AA) com ervilhas verdes homozigotas (aa), ocorrem ervilhas amarelas heterozigotas (Aa). Se essas ervilhas forem cruzadas entre si, ocorrem três ervilhas amarelas para cada ervilha verde (a proporção é de três para um). Suponha que tenham sido pegas, ao acaso, três ervilhas resultantes do cruzamento de ervilhas amarelas heterozigotas. Qual é 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:

$$\frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} = \frac{1}{64}$$

13.5.10. 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 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, ou seja, ½.

13.5.11. 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:

Cara-cara; Cara-coroa; coroa-cara; 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

$$P(A \cap B) = \frac{1}{4}$$

Como

$$P(A) = \frac{2}{4} = \frac{1}{2}$$

$$P(B) = \frac{2}{4} = \frac{1}{2}$$

$$P(A \cap B) = P(A) \times P(B)$$

A condição de independência foi, portanto, satisfeita. Os eventos "cara na primeira moeda" e "faces iguais nas duas moedas" são independentes.

13.6 Exercícios propostos

13.6.1. 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?

13.6.2. Uma urna tem dez bolas numeradas de 1 a 10. Retira-se uma bola ao acaso. Qual é a probabilidade de essa bola: a) ter um número maior do que 7? b) ter um número menor do que 7? c) ter número 1 ou 10?

13.6.3. 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 um número maior do que 15?

13.6.4. 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?

13.6.5. Dois dados, um vermelho, outro azul, são lançados ao mesmo tempo e se pergunta: a) qual é a probabilidade de ocorrer a face 6 no dado vermelho? b) qual é a probabilidade de ocorrer a face 6 no dado vermelho, sabendo que saiu a face 6 no dado azul?

13.6.6. Um exame realizado em jovens que concluíram 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?

13.6.7. 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?

13.6.8. 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 de forma independente, 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?

13.6.9. 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)?

13.6.10. Suponha que a probabilidade de uma pessoa ser do tipo sanguíneo O é de 40%, ser A é de 30% e ser B é de 20%. Suponha ainda que o fator Rh não dependa do tipo sanguíneo e que a probabilidade de Rh^+ é de 90%. Nessas condições, calcule a probabilidade de uma pessoa tomada ao acaso da população ser:

a) O, Rh^+

b) AB, Rh^-

---

[1]Os jogos de azar são antiquíssimos e foram praticados não só como apostas, mas também como um modo de prever o futuro, decidir conflitos ou dividir heranças.

[3]*Não* existe, por exemplo, 200% de probabilidade. Expressões desse tipo aparecem na linguagem coloquial, na intenção de enfatizar uma certeza.

APÊNDICE

CAPÍTULO 14

Distribuição Binomial

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*. E também 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*. Além disso, todos nós sabemos que ganhar na loteria não é fácil. Temos, portanto, percepção sobre *probabilidade*. A seguir, definiremos o que é variável aleatória – que, intuitivamente, você talvez já conheça.

14.1 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, de copas. O acaso determina o resultado.

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 responder que tem 42 anos, a variável aleatória que representa idade de pessoas assumiu, nesse caso, o valor 42.

As variáveis aleatórias são, portanto, *numéricas*. Portanto, podem ser *discretas* e *contínuas*. Neste capítulo, vamos estudar as variáveis aleatórias discretas.

14.1.1 Variável aleatória binária

Alguns experimentos só podem resultar em uma de duas possibilidades: o evento no qual estamos interessados, o "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 o valor zero ao “fracasso”.

14.1.2 Variável aleatória binomial

Muitas vezes, contamos o número de vezes em 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 dez moedas;
- um pesquisador conta *quantos*, dos quinhentos chefes de família que entrevistou, eram mulheres;
- um médico conta *quantos*, dos cem pacientes que tratou com uma nova droga, ficaram curados;
- um biomédico conta *quantos*, dos 32 hemogramas feitos no dia, indicaram doença contagiosa;
- uma enfermeira conta *quantos*, dos 3.052 nascidos vivos em determinado ano em uma maternidade, tinham doença ou defeito grave.

A variável que resulta da soma dos resultados de uma variável aleatória binária em n tentativas é uma *variável aleatória binomial*.

Exemplo 14.1 Variável aleatória binomial

Uma moeda é lançada duas vezes. O número X de caras que podem ocorrer estão apresentados na Tabela 14.1.

Tabela 14.1

Eventos possíveis e número de caras quando uma moeda é lançada duas vezes

Eventos possíveis	Valor de X
Coroa e coroa	0
Coroa e cara	1
Cara e coroa	1
Cara e cara	2

14.2 Distribuição de probabilidades

Os valores observados da variável aleatória X são indicados por x_1, x_2,..., x_k e as respectivas probabilidades por $p(x_1)$, $p(x_2)$,..., Px_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.

Distribuição de probabilidades de uma variável aleatória discreta X é a lista dos valores que X pode assumir e suas respectivas probabilidades.

Exemplo 14.2 Distribuição de probabilidades

Seja X a variável aleatória que representa o número de caras obtidas quando se lança uma moeda duas vezes, vamos calcular a distribuição de probabilidades de X.

Se saírem duas coroas, $X = 0$. A probabilidade de $X = 0$ é:

$$P\,(coroa) \times P\,(coroa) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} = 0,25$$

Se saírem uma coroa e uma cara, a variável X assume valor um. A probabilidade $X = 1$ é:

$$P\,(coroa) \times P\,(cara) + P\,(cara) \times P\,(coroa) = \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} + \frac{1}{4} = 0,50$$

Se saírem duas caras, a variável X assume valor dois. A probabilidade de $X = 2$ é:

$$P\,(cara) \times P\,(cara) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} = 0,25$$

A Tabela 14.2 e a Figura 14.1 apresentam um resumo desses cálculos, ou seja, apresentam a distribuição de probabilidades de X. A soma das probabilidades é 1.

Tabela 14.2

Distribuição de probabilidades do número de caras em dois lançamentos de uma moeda

Evento	Valor de X	$P(X)$
Coroa e coroa	0	
Coroa e cara ou cara e coroa	1	
Cara e cara	2	
Total		1

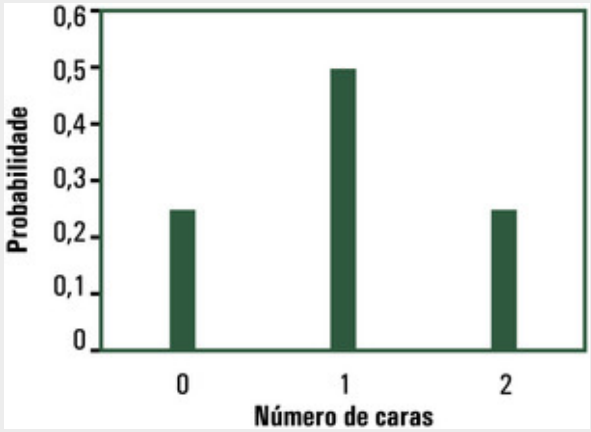


FIGURA 14.1 Distribuição de probabilidades do número de caras em dois lançamentos de uma moeda

Neste ponto, é importante deixar claro que existe diferença entre *distribuição de probabilidades* e *distribuição de frequências*. As distribuições de frequências, tratadas no Capítulo 1, 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 nos dados de toda a população. A distribuição de probabilidades é estável.

14.3 Distribuição binomial

Uma distribuição de probabilidades bem conhecida é a *distribuição binomial*, que estuda o número X de sucessos em n tentativas e suas respectivas probabilidades.

Para aprender a trabalhar com a distribuição binomial, imagine que, em determinada maternidade, tenham nascido três bebês em um dia. Vamos estudar a distribuição de meninos em três nascimentos. Fazendo A indicar menina e O indicar menino, os eventos possíveis são os seguintes:

AAA AAO AOO OOO
AOA OAO
OAA OOA

O número de meninos que podem ocorrer em três nascimentos é uma *variável aleatória binomial*, que indicaremos por X. A Tabela 14.3 apresenta os valores possíveis de X e o número de vezes que cada um deles ocorre.

Tabela 14.3

Números possíveis de meninos em três nascimentos

Valor de X	Frequência
0	1
1	3
2	3
3	1

Seja p a probabilidade de nascer menino e q a probabilidade de nascer menina. Então, $p + q = 1$.

Se nascerem três meninas, ou seja, se ocorrer o evento AAA, a variável aleatória X assume valor zero, com probabilidade:

$$\mathrm{P}\,[X=0] = \mathrm{P}\,[\mathrm{A}] \times \mathrm{P}[\mathrm{A}] \times \mathrm{P}[\mathrm{A}] = q \times q \times q = q^3$$

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:

$$\mathrm{P}[\mathrm{A}] \times \mathrm{P}[\mathrm{A}] \times \mathrm{P}[\mathrm{O}] = q \times q \times p = pq^2$$
$$\mathrm{P}[\mathrm{A}] \times \mathrm{P}[\mathrm{O}] \times \mathrm{P}[\mathrm{A}] = q \times p \times q = pq^2$$
$$\mathrm{P}[\mathrm{O}] \times \mathrm{P}[\mathrm{A}] \times \mathrm{P}[\mathrm{A}] = p \times q \times q = pq^2$$

Então,

$$\mathrm{P}\ [X=1] = 3pq^2$$

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:

$$\mathrm{P}[\mathrm{A}] \times \mathrm{P}[\mathrm{O}] \times \mathrm{P}[\mathrm{O}] = q \times p \times q = p^2 q$$
$$\mathrm{P}[\mathrm{O}] \times \mathrm{P}[\mathrm{A}] \times \mathrm{P}[\mathrm{O}] = q \times q \times p = p^2 q$$
$$\mathrm{P}[\mathrm{O}] \times \mathrm{P}[\mathrm{O}] \times \mathrm{P}[\mathrm{A}] = p \times p \times q = p^2 q$$

Então,

$$P[X = 2] = 3p^2q$$

Se nascerem três meninos, isto é, se ocorrer o evento OOO, a variável aleatória X assume valor 3, com probabilidade:

$$P[X = 3] = P[O] \times P[O] \times P[O] = p \times p \times p = p^3$$

A distribuição binomial do número X de meninos em n = 3 nascimentos está na Tabela 14.4. São dados os resultados possíveis de X e suas respectivas probabilidades.

Tabela 14.4

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 seja p = 0,5 e que a probabilidade de nascer menina seja 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

$$p = 0{,}5$$
$$q = 0,5,$$

obtemos a distribuição de probabilidades do número de meninos em três nascimentos, apresentada na Tabela 14.5 e na Figura 14.2.

Tabela 14.5

Distribuição de probabilidades do número de meninos em três nascimentos $(p = q = 0,5)$

Valor de X	$P(X)$
0	
1	
2	
3	
Total	1

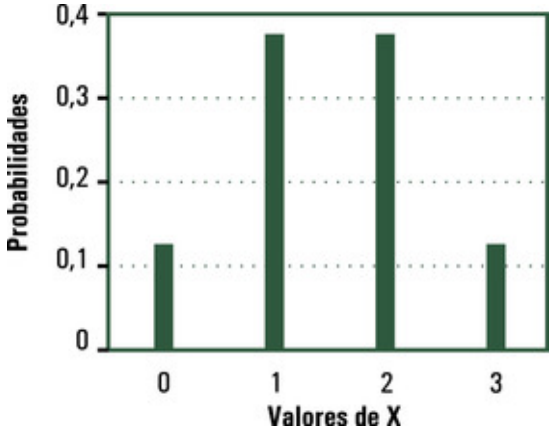


FIGURA 14.2 Distribuição de probabilidades do número de meninos em três nascimentos

14.3.1 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*:

1. n, ou seja, o *número de ensaios* (por exemplo, se uma moeda for lançada dez vezes)
2. p, ou seja, a probabilidade de sucesso em uma tentativa (por exemplo, sair cara quando se joga uma moeda).

*14.3.2 Função de distribuição na distribuição binomial

Um parâmetro de interesse é a probabilidade de sucesso numa distribuição binomial. Lembre-se de que a distribuição binomial surge quando se conta o número X de sucessos em n ensaios.

Considere um experimento em que fazemos n observações independentes da variável aleatória X que segue uma distribuição f (x | $\boldsymbol{p}$), onde $\boldsymbol{p}$ é o vetor de parâmetros (ou seja, $\{p_1, p_2, ..., p_k\}$) para o de distribuição. A probabilidade de obter os resultados específicos para essa experiência é dada pela

Distribuição de probabilidades de uma variável aleatória discreta X, que é a lista dos valores que X pode assumir e suas respectivas probabilidades.

Vamos aceitar, sem demonstração, que, dada uma distribuição binomial de parâmetros n e p, a probabilidade de ocorrerem x eventos favoráveis é dada pela seguinte fórmula:

$$\binom{n}{x} p^x q^{(n-x)}$$

em que $\binom{n}{x}$ é a combinação[1] de n, x a x. Portanto, a probabilidade de ocorrerem x eventos favoráveis em n tentativas é dada pela seguinte fórmula:

$$P(X = x) = \frac{n!}{x!(n-x)!} p^x q^{(n-x)}$$

Veja, agora, um exemplo que ajuda a entender como trabalhamos com a distribuição binomial.

Exemplo 14.3 Eventos em uma distribuição binomial

Um dentista vai examinar uma amostra de quatro crianças de 6 anos para saber se elas têm (Sim, indicado por S) ou não (Não, indicado por N) cárie. Quais são os eventos possíveis?

Os eventos possíveis são os que seguem:

NNNN	NNNS	NNSS	NSSS	SSSS
	NNSN	NSNS	SNSS	
	NSNS	NSSN	SSNS	
	SNNN	SNNS	SSSN	
		SNSN		
		SSNN		

Exemplo 14.4 Distribuição binomial

Reveja o Exemplo 14.3. 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.

Tabela 14.6

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 14.5 Distribuição binomial (n = 4; p = 0,4)

Reveja o Exemplo 14.4. 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 cáries?

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 14.6 Cálculo de probabilidades na distribuição binomial

Reveja o Exemplo 14.4. 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 cáries 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 cáries nessa população é de 0,3456.

*14.3.3 Média e variância na distribuição binomial

A média μ (lê-se mi) de uma distribuição binomial é dada pela seguinte fórmula:

$$\mu = np$$

e a variância σ^2 (lê-se sigma ao quadrado) é dada pela fórmula a seguir:

$$\sigma^2 = npq$$

Exemplo 14.7 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.

A média é

$$\mu = np = 1.000 \times 0,5 = 500 \text{ meninos},$$

e a variância é

$$\sigma^2 = npq = 1.000 \times 0,5 \times 0,5 = 250.$$

14.4 Revisão sobre análise combinatória

Se n é um número inteiro positivo maior do que zero, por definição, o *fatorial de n*, que se indica por $n!$, é dado por:

$$n! = n(n-1)(n-2)\ldots 1.$$

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 $x \leq n$, a *combinação de n*, x a x, é indicada por:

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

Essa 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:

$$\binom{n}{n} = \frac{n!}{n!(n-n)!} = \frac{n!}{n!0!} = 1$$

14.5 Exercícios resolvidos

14.5.1. Encontre 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%.

14.5.2. Numa prova,[2] 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 alternativas 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 respostas que ele consegue acertar?

A resposta *e* seria correta: Descoberta do Brasil (1500), Descoberta da América (1492), Independência do Brasil (1822). Outras respostas têm as datas de um ou dois acontecimentos na ordem correta. Veja a distribuição de probabilidades na Tabela 14.7.

Tabela 14.7

Distribuição de probabilidades do número de respostas que o aluno acerta

Resposta	Probabilidade	N° de respostas corretas
a	1/6	0
b	1/6	1
c	1/6	1
d	1/6	0
e	1/6	3
f	1/6	1

14.5.3. 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− : $p = 0,15$.

Tabela 14.8

Cálculos intermediários para se obter a distribuição binomial

Eventos	Valores possíveis de X	Cálculos	Probabilidade
Rh+, Rh+, Rh+	3	0,85 × 0,85 × 0,85	0,614125
Rh+, Rh+, Rh-	2	0,85 × 0,85 × 0,15	0,108375
Rh+, Rh-, Rh+	2	0,85 × 0,15 × 0,85	0,108375
Rh-, Rh+, Rh+	2	0,15 × 0,85 × 0,85	0,108375
Rh+, Rh-, Rh-	1	0,85 × 0,15 × 0,15	0,019125
Rh-, Rh+, Rh-	1	0,15 × 0,85 × 0,15	0,019125
Rh-, Rh-, Rh+	1	0,15 × 0,15 × 0,85	0,019125
Rh-, Rh-, Rh-	0	0,15 × 0,15 × 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 apresentada na Tabela 9.9.

Tabela 14.9

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 14.3 Distribuição de probabilidades do número de pessoas com Rh+, em três pessoas.

14.5.4. 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:

$$P(X=0)=\binom{6}{0}\times\left(\frac{1}{2}\right)^{0}\times\left(\frac{1}{2}\right)^{6}=$$

$$=\frac{6!}{1!(6-1)!}\times\frac{1}{2^{0}}\times\frac{1}{2^{6}}=\frac{1}{64}$$

Para obter a probabilidade de X assumir o valor 1, ou seja, 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, ou seja, 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 14.10, que apresenta uma distribuição binomial para $n = 6$ e $p = 0{,}5$. O gráfico de barras é apresentado na Figura 14.4.

Tabela 14.10

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

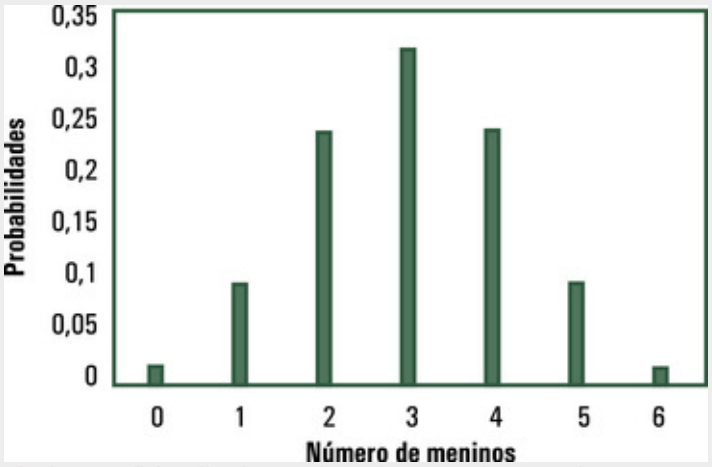


FIGURA 14.4 Distribuição do número de meninos em uma família com seis crianças

14.5.5. 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 seguinte fórmula:

$$P(X = x) = \binom{n}{x} p^x q^{(n-x)}$$

Então:

$$P(X = 4) = \binom{4}{4} \times 0,8^4 \times 0,92^0 =$$
$$= 0,00004096 \text{ ou } 0,004096\,\%$$

14.5.6. O resultado do cruzamento de ervilhas amarelas homozigotas (AA) com ervilhas verdes homozigotas (aa) são ervilhas amarelas heterozigotas (Aa). Se essas 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:

$$P(X=2) = \binom{4}{2} \times \left(\frac{3}{4}\right)^2 \left(\frac{1}{4}\right)^2 = \\ = 0,2109 \text{ ou } 21,09\,\%$$

14.5.7. Considere novamente o cruzamento de ervilhas amarelas e verdes, descrito no Exercício 14.5.6. 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:

$$\sigma^2 = 100 \times \frac{3}{4} \times \frac{1}{4} = 18,75$$

14.5.8. Um exame é constituído de cem testes com cinco alternativas, em que 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 ao acaso é p = 1/5. Existem n = 100 testes. Então, aplicando a fórmula, tem-se:

$$\mu = 100 \times \frac{1}{5} = 20$$

ou seja, um aluno que nada sabe sobre a matéria acerta, em média, vinte testes. A variância da distribuição é:

$$\sigma^2 = 100 \times \frac{1}{5} \times \frac{4}{5} = 16$$

14.5.9. Um pesquisador de mercado quer saber a proporção de consumidores que preferem café sem cafeína. Se ele pergunta a quinhentas pessoas que tipo de café adquiriram 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 a questionários”. Se as respostas do tipo “Sim” e “Não” chegarem a 70%, ou seja, se a taxa de resposta for de 70% (quando a quantidade de não respondentes é grande, a pesquisa não responde à pergunta feita, ou seja, não tem validade), terá uma distribuição binomial. A média será obtida pela seguinte fórmula:

$$u = np$$

e a variância σ^2 pela fórmula a seguir:

$$\sigma^2 = npq$$

O valor de *p* é obtido dividindo o número de consumidores que preferem café sem cafeína pelo número *n* de respondentes.

14.5.10. Numa cirurgia experimental, uma cobaia pode sobreviver (S) ou morrer (M). O pesquisador não sabe (é exatamente isso que ele está pesquisando), mas considere que a probabilidade de uma cobaia sobreviver na cirurgia é de 0,25. A cirurgia será feita em duas cobaias. Se ambas sobreviverem, operam-se mais duas. Se apenas uma sobreviver, outra será operada. Se as duas morrerem, o pesquisador interrompe o experimento. Qual é a probabilidade de não se fazer uma segunda sequência de cirurgias (de as duas primeiras cobaias operadas morrerem)? Qual é a probabilidade de quatro cobaias serem operadas e as quatro sobreviverem?

As respostas são dadas na Tabela 14.11. Se as duas cobaias morrerem (sobrevivência zero), o pesquisador interrompe o experimento. A probabilidade de isso ocorrer é de 0,5625. Se as duas cobaias sobreviverem (sobrevivência 2), o pesquisador opera mais duas. A probabilidade de isso ocorrer é:

Tabela 14.11

Probabilidade de sobrevivência de cobaias submetidas a uma cirurgia experimental

Nº de cobaias operadas (n)	1ª sequência		2ª sequência	
	Sobrevivência	Probabilidade	Sobrevivência	Probabilidade
2	0	0,5625	Interrompe o experimento	
2	1	0,375	0	0,75
			1	0,25
2	2	0,0625	0	0,5625
			1	0,375
			2	0,0625

$$0{,}0625 \times 0{,}0625 = 0{,}0039$$

[2]Adaptado de Mosteller, F. Rourke, R. E. K., Thomas JR, G. B. *Probability and Statistics.* Reading, Addison- Wesley, 1961, p. 160.

14.6 Exercícios propostos

14.6.1. Há três bolas numeradas em uma caixa, cada qual com um número diferente. Os números são 1, 2 e 3. Tira-se uma bola da caixa; em seguida, outra. Forma-se, então, um número de dois dígitos com os números das bolas retiradas. Por exemplo, se saiu o número 3 e, em seguida, o 2, foi formado o número 32. Um jogador ganha se sair número par. Nesse jogo, ganha-se mais do que se perde ou é justamente o contrário?

14.6.2. 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.

14.6.3. Um exame é constituído de dez testes tipo certo-errado. Um aluno que nada sabe sobre a matéria do exame, quantos testes, em média, acerta? Qual é a variância dessa distribuição?

14.6.4. Um exame é constituído de dez testes com cinco alternativas, em que 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?

14.6.5. Suponha que determinado medicamento usado no diagnóstico precoce da gravidez é capaz de confirmar casos positivos em apenas 90% das gestantes muito jovens. Isso porque, em 10% das gestantes muito jovens, ocorre descamação do epitélio do útero, que é confundida com menstruação. Nessas condições, qual é a probabilidade de duas, de três gestantes muito jovens que fizeram uso desse medicamento, não terem confirmado precocemente a gravidez?

14.6.6. A probabilidade de um casal heterozigoto para o gene da fenilcetonúria ($Aa \times Aa$) ter um filho afetado (aa) é de 1/4. Se o casal tiver três filhos, qual é a probabilidade de ter um filho com essa doença?

14.6.7. A probabilidade de um indivíduo ter sangue Rh^- é 10%, na população brasileira toda. Qual é a possibilidade de se terem apresentado, em determinado dia em um banco de sangue, cinco doadores de sangue, todos Rh^-?

14.6.8. Foi feito um levantamento acerca da opinião de 1.000 enfermeiras que trabalhavam em determinado hospital sobre dada 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?

14.6.9. 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 sejam submetidos ao detector de mentiras. Todos os suspeitos se afirmam inocentes e estão dizendo a verdade. Qual é a probabilidade de ocorrer uma resposta positiva?

14.6.10. 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?

---

[1]Uma rápida revisão sobre análise combinatória é dada ao final deste Apêndice.

Anexos

ESBOÇO

Anexos Capítulo 15: Tabelas

ANEXOS

CAPÍTULO 15

Tabelas

Tabela 1

Distribuição normal reduzida P(0 < Z < z)

	Último dígito									
	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 c^2, segundo os graus de liberdade e o valor de α

Graus de liberdade	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 α = 2,5%, segundo o número de graus de liberdade do numerador e do denominador

Nº de g. l do denominador	Número de graus de liberdade do numerador																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	648	800	864	900	922	937	948	957	963	969	977	985	993	997	1.000	1.010	1.010	1.010	1.020
2	38,5	39,0	39,2	39,2	39,3	39,3	39,4	39,4	39,4	39,4	39,4	39,4	39,4	39,5	39,5	39,5	39,5	39,5	39,5
3	17,4	16,0	15,4	15,1	14,9	14,7	14,6	14,5	14,5	14,4	14,3	14,3	14,2	14,1	14,1	14,0	14,0	13,9	13,9
4	12,2	10,6	9,98	9,60	9,36	9,20	9,07	8,98	8,90	8,84	8,75	8,66	8,56	8,51	8,46	8,41	8,36	8,31	8,26
5	10,0	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68	6,62	6,52	6,43	6,33	6,28	6,23	6,18	6,12	6,07	6,02
6	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52	5,46	5,37	5,27	5,17	5,12	5,07	5,01	4,96	4,90	4,85
7	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82	4,76	4,67	4,57	4,47	4,42	4,36	4,31	4,25	4,20	4,14
8	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36	4,30	4,20	4,10	4,00	3,95	3,89	3,84	3,78	3,73	3,67
9	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03	3,96	3,87	3,77	3,67	3,61	3,56	3,51	3,45	3,39	3,33
10	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78	3,72	3,62	3,52	3,42	3,37	3,31	3,26	3,20	3,14	3,08
11	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,59	3,53	3,43	3,33	3,23	3,17	3,12	3,06	3,00	2,94	2,88
12	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44	3,37	3,28	3,18	3,07	3,02	2,96	2,91	2,85	2,79	2,72
13	6,41	4,97	4,35	4,00	3,77	3,60	3,48	3,39	3,31	3,25	3,15	3,05	2,95	2,89	2,84	2,78	2,72	2,66	2,60
14	6,30	4,86	4,24	3,89	3,66	3,50	3,38	3,29	3,21	3,15	3,05	2,95	2,84	2,79	2,73	2,67	2,61	2,55	2,49
15	6,20	4,77	4,15	3,80	3,58	3,41	3,29	3,20	3,12	3,06	2,96	2,86	2,76	2,70	2,64	2,59	2,52	2,46	2,40
16	6,12	4,69	4,08	3,73	3,50	3,34	3,22	3,12	3,05	2,99	2,89	2,79	2,68	2,63	2,57	2,51	2,45	2,38	2,32
17	6,04	4,62	4,01	3,66	3,44	3,28	3,16	3,06	2,98	2,92	2,82	2,72	2,62	2,56	2,50	2,44	2,38	2,32	2,25
18	5,98	4,56	3,95	3,61	3,38	3,22	3,10	3,01	2,93	2,87	2,77	2,67	2,56	2,50	2,44	2,38	2,32	2,26	2,19
19	5,92	4,51	3,90	3,56	3,33	3,17	3,05	2,96	2,88	2,82	2,72	2,62	2,51	2,45	2,39	2,33	2,27	2,20	2,13
20	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84	2,77	2,68	2,57	2,46	2,41	2,35	2,29	2,22	2,16	2,09
21	5,83	4,42	3,82	3,48	3,25	3,09	2,97	2,87	2,80	2,73	2,64	2,53	2,42	2,37	2,31	2,25	2,18	2,11	2,04
22	5,79	4,38	3,78	3,44	3,22	3,05	2,93	2,84	2,76	2,70	2,60	2,50	2,39	2,33	2,27	2,21	2,14	2,08	2,00
23	5,75	4,35	3,75	3,41	3,18	3,02	2,90	2,81	2,73	2,67	2,57	2,47	2,36	2,30	2,24	2,18	2,11	2,04	1,97
24	5,72	4,32	3,72	3,38	3,15	2,99	2,87	2,78	2,70	2,64	2,54	2,44	2,33	2,27	2,21	2,15	2,08	2,01	1,94
25	5,69	4,29	3,69	3,35	3,13	2,97	2,85	2,75	2,68	2,61	2,51	2,41	2,30	2,24	2,18	2,12	2,05	1,98	1,91
26	5,66	4,27	3,67	3,33	3,10	2,94	2,82	2,73	2,65	2,59	2,49	2,39	2,28	2,22	2,16	2,09	2,03	1,95	1,88
27	5,63	4,24	3,65	3,31	3,08	2,92	2,80	2,71	2,63	2,57	2,47	2,36	2,25	2,19	2,13	2,07	2,00	1,93	1,85
28	5,61	4,22	3,63	3,29	3,06	2,90	2,78	2,69	2,61	2,55	2,45	2,34	2,23	2,17	2,11	2,05	1,98	1,91	1,83
29	5,59	4,20	3,61	3,27	3,04	2,88	2,76	2,67	2,59	2,53	2,43	2,32	2,21	2,15	2,09	2,03	1,96	1,89	1,81
30	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57	2,51	2,41	2,31	2,20	2,14	2,07	2,01	1,94	1,87	1,79
40	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,45	2,39	2,29	2,18	2,07	2,01	1,94	1,88	1,80	1,72	1,64
60	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33	2,27	2,17	2,06	1,94	1,88	1,82	1,74	1,67	1,58	1,48
120	5,15	3,80	3,23	2,89	2,67	2,52	2,39	2,30	2,22	2,16	2,05	1,94	1,82	1,76	1,69	1,61	1,53	1,43	1,31
∞	5,02	3,69	3,12	2,79	2,57	2,41	2,29	2,19	2,11	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 α = 5%, segundo o número de graus de liberdade do numerador e do denominador

Nº de g. l do denominador	Número de graus de liberdade do numerador																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4	19,4	19,4	19,4	19,4	19,4	19,5	19,5	19,5	19,5	19,5	19,5
3	10,1	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	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 α = 10%, segundo o número de graus de liberdade do numerador e do denominador

Nº de g. l do denominador	Número de graus de liberdade do numerador																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	39,9	49,5	53,6	55,8	57,2	58,2	58,9	59,4	59,9	60,2	60,7	61,2	61,7	62,0	62,3	62,5	62,8	63,1	63,3
2	8,53	9,00	9,16	9,24	9,29	9,33	9,35	9,37	9,38	9,39	9,41	9,42	9,44	9,45	9,46	9,47	9,47	9,48	9,49
3	5,54	5,46	5,39	5,34	5,31	5,28	5,27	5,25	5,24	5,23	5,22	5,20	5,18	5,18	5,17	5,16	5,15	5,14	5,13
4	4,54	4,32	4,19	4,11	4,05	4,01	3,98	3,95	3,94	3,92	3,90	3,87	3,84	3,83	3,82	3,80	3,79	3,78	3,76
5	4,06	3,78	3,62	3,52	3,45	3,40	3,37	3,34	3,32	3,30	3,27	3,24	3,21	3,19	3,17	3,16	3,14	3,12	3,10
6	3,78	3,46	3,29	3,18	3,11	3,05	3,01	2,98	2,96	2,94	2,90	2,87	2,84	2,82	2,80	2,78	2,76	2,74	2,72
7	3,59	3,26	3,07	2,96	2,88	2,83	2,78	2,75	2,72	2,70	2,67	2,63	2,59	2,58	2,56	2,54	2,51	2,49	2,47
8	3,46	3,11	2,92	2,81	2,73	2,67	2,62	2,59	2,56	2,54	2,50	2,46	2,42	2,40	2,38	2,36	2,34	2,32	2,29
9	3,36	3,01	2,81	2,69	2,61	2,55	2,51	2,47	2,44	2,42	2,38	2,34	2,30	2,28	2,25	2,23	2,21	2,18	2,16
10	3,29	2,92	2,73	2,61	2,52	2,46	2,41	2,38	2,35	2,32	2,28	2,24	2,20	2,18	2,16	2,13	2,11	2,08	2,06
11	3,23	2,86	2,66	2,54	2,45	2,39	2,34	2,30	2,27	2,25	2,21	2,17	2,12	2,10	2,08	2,05	2,03	2,00	1,97
12	3,18	2,81	2,61	2,48	2,39	2,33	2,28	2,24	2,21	2,19	2,15	2,10	2,06	2,04	2,01	1,99	1,96	1,93	1,90
13	3,14	2,76	2,56	2,43	2,35	2,28	2,23	2,20	2,16	2,14	2,10	2,05	2,01	1,98	1,96	1,93	1,90	1,88	1,85
14	3,10	2,73	2,52	2,39	2,31	2,24	2,19	2,15	2,12	2,10	2,05	2,01	1,96	1,94	1,91	1,89	1,86	1,83	1,80
15	3,07	2,70	2,49	2,36	2,27	2,21	2,16	2,12	2,09	2,06	2,02	1,97	1,92	1,90	1,87	1,85	1,82	1,79	1,76
16	3,05	2,67	2,46	2,33	2,24	2,18	2,13	2,09	2,06	2,03	1,99	1,94	1,89	1,87	1,84	1,81	1,78	1,75	1,72
17	3,03	2,64	2,44	2,31	2,22	2,15	2,10	2,06	2,03	2,00	1,96	1,91	1,86	1,84	1,81	1,78	1,75	1,72	1,69
18	3,01	2,62	2,42	2,29	2,20	2,13	2,08	2,04	2,00	1,98	1,93	1,89	1,84	1,81	1,78	1,75	1,72	1,69	1,66
19	2,99	2,61	2,40	2,27	2,18	2,11	2,06	2,02	1,98	1,96	1,91	1,86	1,81	1,79	1,76	1,73	1,70	1,67	1,63
20	2,97	2,59	2,38	2,25	2,16	2,09	2,04	2,00	1,96	1,94	1,89	1,84	1,79	1,77	1,74	1,71	1,68	1,64	1,61
21	2,96	2,57	2,36	2,23	2,14	2,08	2,02	1,98	1,95	1,92	1,88	1,83	1,78	1,75	1,72	1,69	1,66	1,62	1,59
22	2,95	2,56	2,35	2,22	2,13	2,06	2,01	1,97	1,93	1,90	1,86	1,81	1,76	1,73	1,70	1,67	1,64	1,60	1,57
23	2,94	2,55	2,34	2,21	2,11	2,05	1,99	1,95	1,92	1,89	1,84	1,80	1,74	1,72	1,69	1,66	1,62	1,59	1,55
24	2,93	2,54	2,33	2,19	2,10	2,04	1,98	1,94	1,91	1,88	1,83	1,78	1,73	1,70	1,67	1,64	1,61	1,57	1,53
25	2,92	2,53	2,32	2,18	2,09	2,02	1,97	1,93	1,89	1,87	1,82	1,77	1,72	1,69	1,66	1,63	1,59	1,56	1,52
26	2,91	2,52	2,31	2,17	2,08	2,01	1,96	1,92	1,88	1,86	1,81	1,76	1,71	1,68	1,65	1,61	1,58	1,54	1,50
27	2,90	2,51	2,30	2,17	2,07	2,00	1,95	1,91	1,87	1,85	1,80	1,75	1,70	1,67	1,64	1,60	1,57	1,53	1,49
28	2,89	2,50	2,29	2,16	2,06	2,00	1,94	1,90	1,87	1,84	1,79	1,74	1,69	1,66	1,63	1,59	1,56	1,52	1,48
29	2,89	2,50	2,28	2,15	2,06	1,99	1,93	1,89	1,86	1,83	1,78	1,73	1,68	1,65	1,62	1,58	1,55	1,51	1,47
30	2,88	2,49	2,28	2,14	2,05	1,98	1,93	1,88	1,85	1,82	1,77	1,72	1,67	1,64	1,61	1,57	1,54	1,50	1,46
40	2,84	2,44	2,23	2,09	2,00	1,93	1,87	1,83	1,79	1,76	1,71	1,66	1,61	1,57	1,54	1,51	1,47	1,42	1,38
60	2,79	2,39	2,18	2,04	1,95	1,87	1,82	1,77	1,74	1,71	1,66	1,60	1,54	1,51	1,48	1,44	1,40	1,35	1,29
120	2,75	2,35	2,13	1,99	1,90	1,82	1,77	1,72	1,68	1,65	1,60	1,55	1,48	1,45	1,41	1,37	1,32	1,26	1,19
∞	2,71	2,30	2,08	1,94	1,85	1,77	1,72	1,67	1,63	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

Graus de liberdade	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

Respostas aos Exercícios Propostos

Capítulo 1: Apresentação de Dados em Tabelas

1.5.1. 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.

1.5.2. Distribuição das pessoas segundo a opinião

Opinião	Frequência	Percentual
Favorável	425	49,9%
Contrária	368	43,2%
Não tem/não sabe	59	6,9%
Total	852	100,0%

1.5.3. Distribuição das notas de duzentos alunos

Nota do aluno	Frequência	Frequência relativa
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

1.5.4. Distribuição dos pacientes segundo o estágio da doença

Estágio da doença	Frequência	Frequência relativa
Leve	8	0,40
Moderado	9	0,45
Severo	3	0,15
Total	20	1,00

1.5.5. 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.

1.5.6. Doadores de sangue segundo o tipo de sangue

Tipo de sangue	Frequência	Frequência relativa
O	15	0,375
A	16	0,4
B	6	0,15
AB	3	0,075
Total	40	1

1.5.7. Vinte alunos.

1.5.8. Distribuição das crianças segundo o hábito de sucção

Hábito de sucção	Frequência	Percentual
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%

1.5.9.

Classe

$70 \vdash 75$
$75 \vdash 80$
$80 \vdash 85$
$85 \vdash 90$

90 ⊢ 95
95 ⊢ 100
100 ⊢ 105
105 ⊢ 110
110 ⊢ 115
115 ⊢ 120

1.5.10. O intervalo de classes é 5 (enfermeiros em serviço). O intervalo de toda a distribuição é 30.

1.5.11. Distribuição de pacientes acidentados no trabalho segundo o tempo de internação, em dias.

Classe	Frequência
1 ⊢ 3	5
3 ⊢ 6	8
6 ⊢ 9	11
9 ⊢ 12	4
12 ⊢ 15	6
15 ⊢ 18	2
Total	36

Distribuição de pacientes acidentados no trabalho segundo o tempo de internação, em dias.

Classe	Frequência
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

1.5.12. 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 calcular o número de classes: $\sqrt{100} \approx 10$; amplitude dos dados: 821-187 = 634. Dividindo a amplitude total pelo número de classes, encontra-se o intervalo de classe: $634 \div 10 = 63{,}4 \approx 65$.

185 ⊢ 250
250 ⊢ 315
315 ⊢ 380
380 ⊢ 445
445 ⊢ 510
510 ⊢ 575
575 ⊢ 640
640 ⊢ 705
705 ⊢ 770
770 ⊢ 835

1.5.13. 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%

1.5.14. Distribuição dos dentistas segundo a adoção de métodos de prevenção de cáries e doenças gengivais no consultório

Prevenção	Frequência	Porcentual
Sim	78	78,0%
Não	22	22,0%
Total	100	100,0%

A prática da prevenção deveria ser adotada por 100% dos dentistas.

1.5.15. Número e proporção de óbitos por grupos de causas. Brasil, 2004.

Grupos de causas	Número		Porcentagem	
	Masculino	Feminino	Masculino	Feminino
Doenças infecciosas e parasitárias	27.437	18.615	5,2%	5,0%
Neoplasias	76.065	64.724	14,5%	17,3%
Doenças do aparelho circulatório	150.383	135.119	28,8%	36,2%
Doenças do aparelho respiratório	55.785	46.369	10,7%	12,4%
Afecções originadas no período perinatal	17.530	13.165	3,4%	3,5%
Causas externas	107.032	20.368	20,5%	5,4%
Demais causas definidas	88.563	75.399	16,9%	20,2%
Total	522.795	373.759	100,0%	100,0%

Houve 896.554 óbitos com causa definida, 58,3% homens e 41,7% mulheres. Doenças do aparelho circulatório respondem pela maior proporção de mortes. Chama a atenção a grande proporção de óbitos de homens por causas externas (acidentes e homicídios).

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

Faixa etária	Número	Frequência relativa
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.

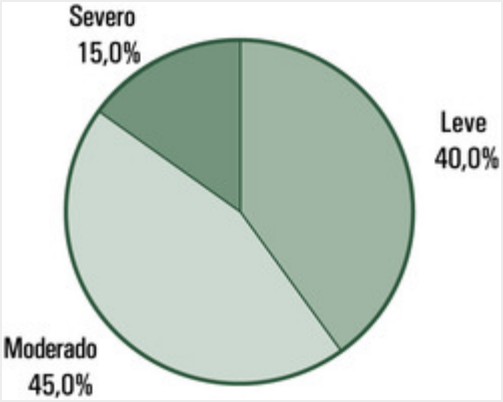
1.5.17. Número de órgãos obtidos de doadores cadáveres.

Órgão	Número de doadores	Número de órgãos aproveitados	Taxa de aproveitamento
Rim	105	210	100,0%
Coração	105	45	42,9%
Fígado	105	20	19,0%
Pulmões	105	17	8,1%

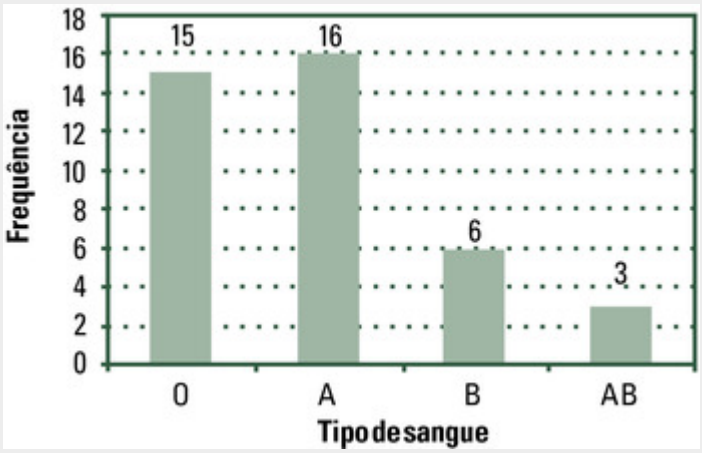
Nota: Cada cadáver é potencialmente doador de dois rins, um coração, um fígado e dois pulmões. A taxa de aproveitamento é sobre número de órgãos – não de cadáveres.

Capítulo 2: Apresentação de Dados em Gráficos

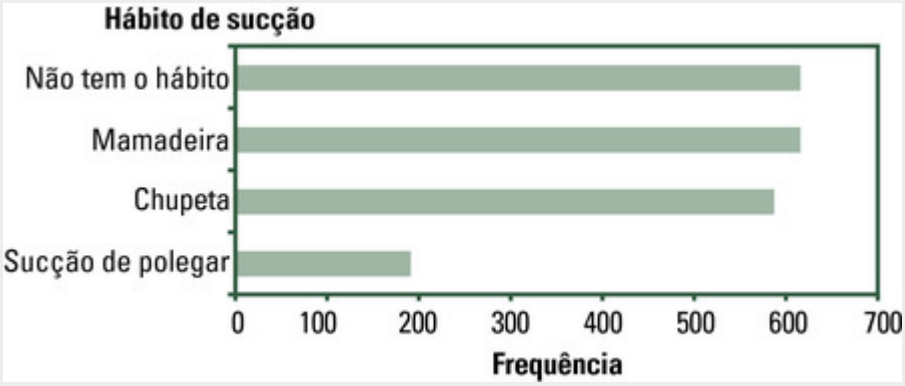
2.4.1. Distribuição dos pacientes segundo o estágio da doença



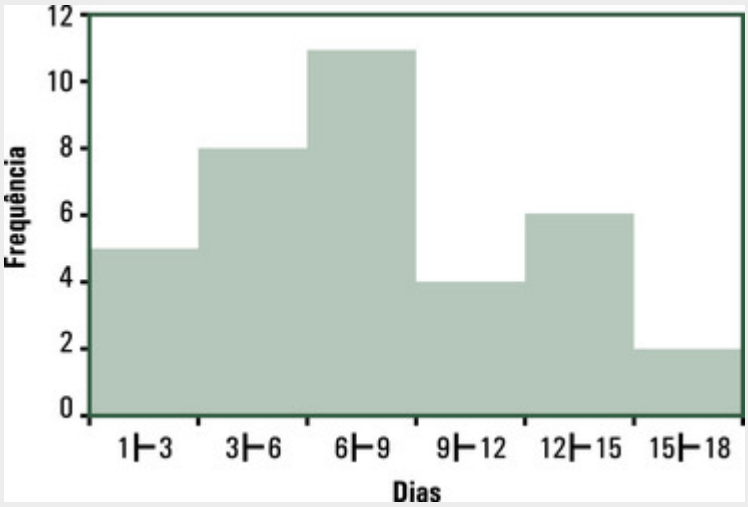
2.4.2. Distribuição dos doadores de sangue segundo o tipo de sangue



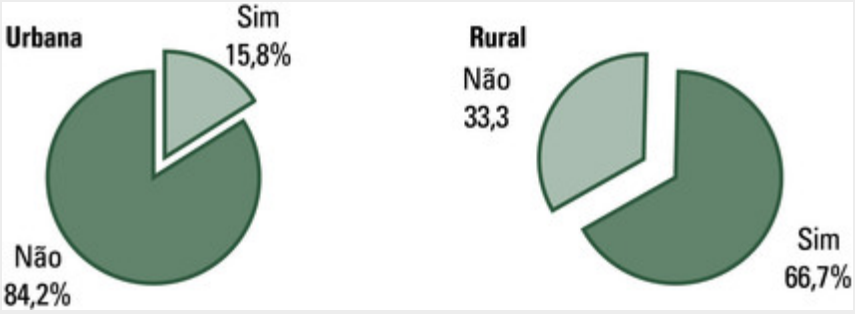
2.4.3. Distribuição das crianças segundo o hábito de sucção



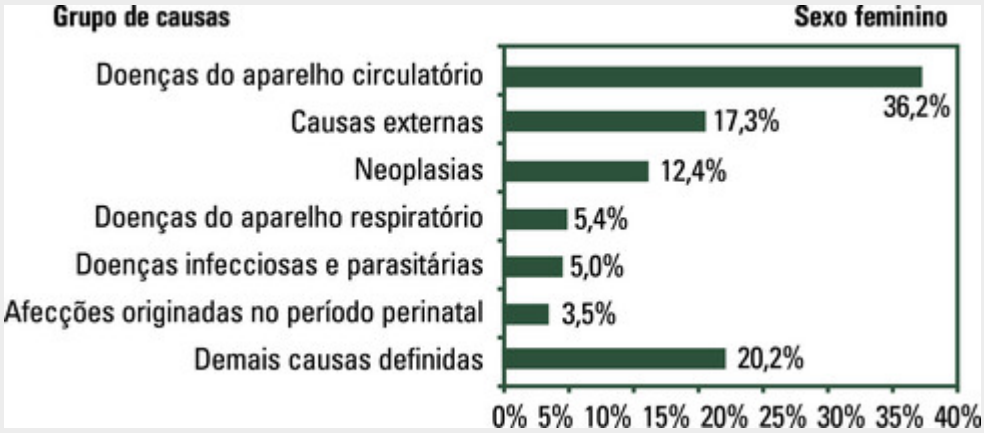
2.4.4. Distribuição de pacientes acidentados no trabalho segundo o tempo de internação, em dias

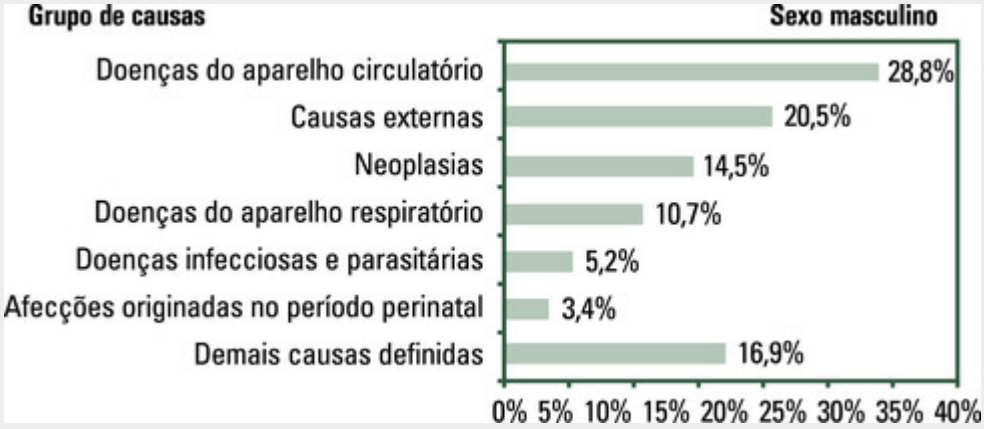


2.4.5. Taxa de abandono do tratamento contra tuberculose pulmonar segundo a zona de moradia

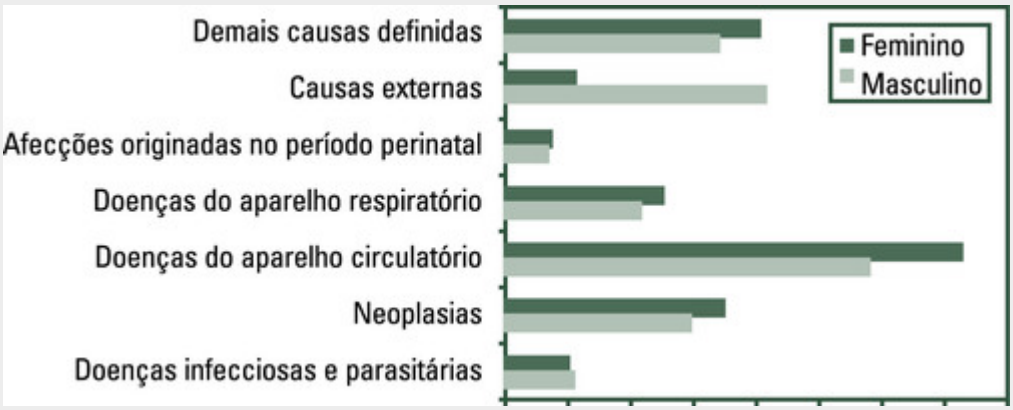


2.4.6. Proporção de óbitos por grupos de causas. Brasil, 2004

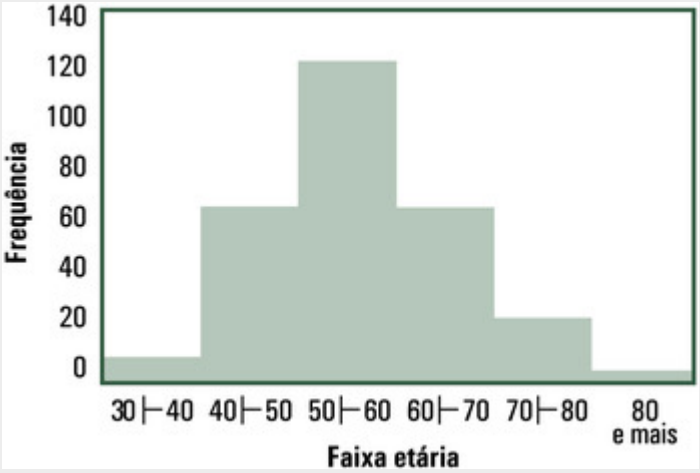




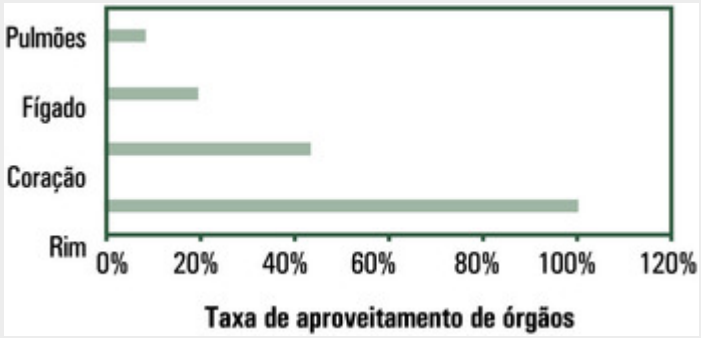
Nesses gráficos, as grandes causas foram colocadas em ordem decrescente, considerando as porcentagens. Mas os dois gráficos podem ser reunidos em um só, como na figura que se segue.



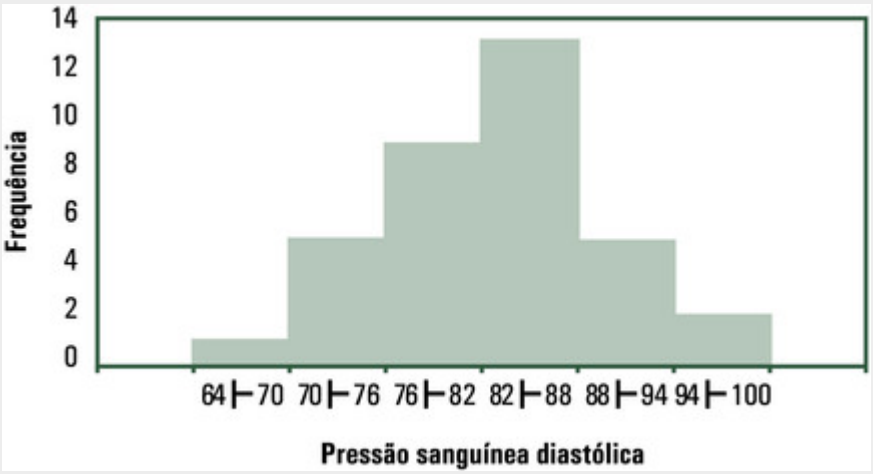
2.4.7. Pacientes portadores de carcinoma epidermoide de base de língua, segundo a faixa etária, em anos



2.4.8. Taxa de aproveitamento de órgãos obtidos de doadores cadáveres

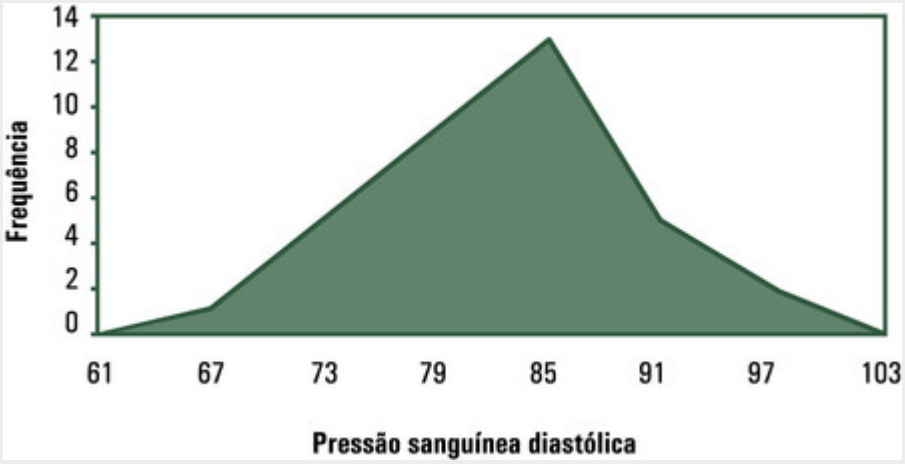


2.4.9. Pressão sanguínea diastólica de 35 enfermeiros que trabalham em um hospital



Classe	Frequência	Frequência relativa
64 ⊢ 70	1	2,9%
70 ⊢ 76	5	14,3%
76 ⊢ 82	9	25,7%
82 ⊢ 88	13	37,1%
88 ⊢ 94	5	14,3%
94 ⊢ 100	2	5,7%
Total	35	100,0%

2.4.10. Pressão sanguínea diastólica de 35 enfermeiros que trabalham em um hospital



Capítulo 3: Medidas de Tendência Central

3.6.1.

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.

3.6.2. Mediana.

3.6.3. Moda.

3.6.4. 24 anos.

3.6.5. A média é 100 mg por 100 mL de sangue e a mediana é 99,5 mg por 100 mL de sangue.

3.6.6. Estatura: Média = 1,70 m; mediana = 1,68 m.
Peso: Média = 72,5 kg; mediana = 70 kg.
Pressão arterial: Média = 165,5 mL de mercúrio; mediana = 160 mL de mercúrio.

3.6.7. Menino: média = 0,88 dentes cariados; meninas: média = 1 dente cariado.

3.6.8. 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.

3.6.9. 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.

3.6.10. Masculino: Média = 0,90 L por dia; mediana = 0,85 L por dia.
Feminino: Média = 0,80 L por dia; mediana = 0,75 L por dia.

3.6.11. Metade das pacientes retornou às atividades menos de 27,5 dias depois de submetidas a histerectomias; o conjunto de dados não tem moda, ou seja, nenhum número de dias foi mais frequente.

3.6.12. 3,62 mg de ácido ascórbico em 100 mL

3.6.13. Sim, exemplo: 1; 2; 3; 3; 3; 4; 5; para esse conjunto de dados, a média, a mediana e a moda são iguais a 3.

3.6.14. A média, porque a última classe não tem o extremo superior definido.

Capítulo 4: Medidas de Dispersão

4.6.1. a) 1; b) 5; c) 4.

4.6.2. a) $\Sigma x = 35$; b)

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

4.6.3. A média é 4 e o desvio padrão é 3.

4.6.4. O tamanho da amostra é 6.

4.6.5. A média é 24, e a variância, 80.

4.6.6. 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.

4.6.7. a) O desvio padrão pode ser maior do que o valor da média; exemplo: a)-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.

4.6.8. A variância é 16, o desvio padrão é 4 e o coeficiente de variação é 4,00%.

4.6.9. A média é 5 e a variância é 0,8.

4.6.10. 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.

4.6.11. 1° ano: média = 74,6; desvio padrão = 7,4.
2° ano: média = 95,6; desvio padrão = 7,9.
As variabilidades são praticamente iguais, mas a diferença é que a média do 2º ano é aproximadamente 28% maior do que a média do 1° ano, o que justifica a ideia de que alunos que começam a atender pacientes em disciplinas clínicas têm aumento na frequência do batimento cardíaco.

4.6.12. A diferença de médias não é muito grande (6 e 7, respectivamente), mas a diferença de variabilidades é tão grande (2 e 11,2, respectivamente) que justifica preferir a primeira dieta para perda de peso. Como na primeira dieta as respostas são mais homogêneas, a expectativa do resultado é mais previsível.

Capítulo 5: Noções sobre Correlação

5.6.1.

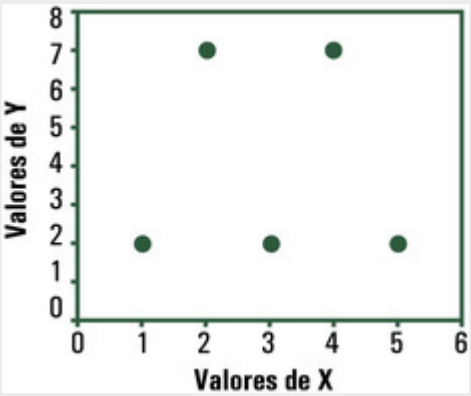
a) $r = 1$: correlação perfeita positiva;
b) $r = -1$: correlação perfeita negativa;
c) $r = 0$: correlação nula;
d) $r = 0{,}90$: correlação positiva alta;
e) $r = -0{,}90$: correlação negativa alta.

5.6.2.

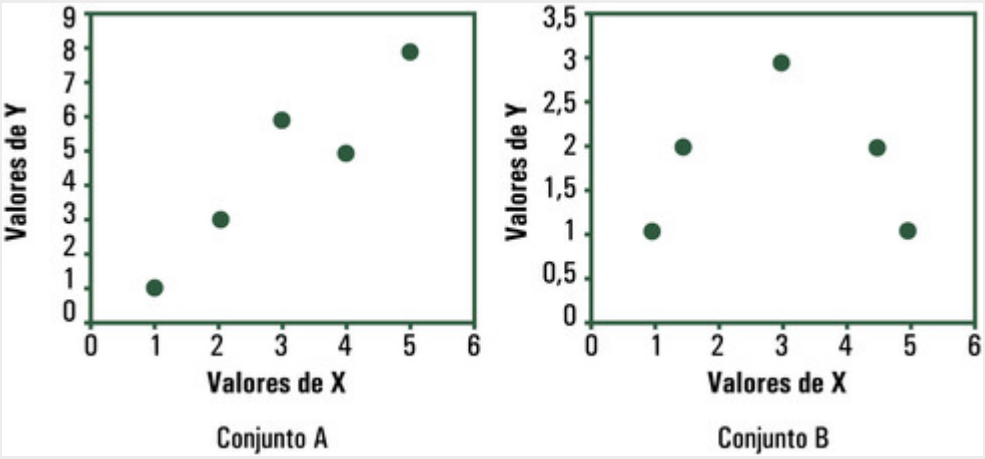
a) correlação negativa;
b) correlação positiva;
c) correlação nula.

5.6.3. O sobrepeso pode ser um fator de risco para morte por doenças do coração.

5.6.4. Não.
5.6.5. Correlação perfeita negativa (7)
Forte correlação positiva (1)
Correlação nula ou próxima de nula (3)
5.6.6. 1; 1 ou -1; positiva ou negativa; zero; maior.
5.6.7. Negativa.
5.6.8. Se as variáveis estão ou não correlacionadas.
5.6.9. Não existe correlação entre as variáveis: r = 0. O diagrama de dispersão mostra isso.



Dados relativos a duas variáveis, *X* e *Y*
5.6.10. 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.

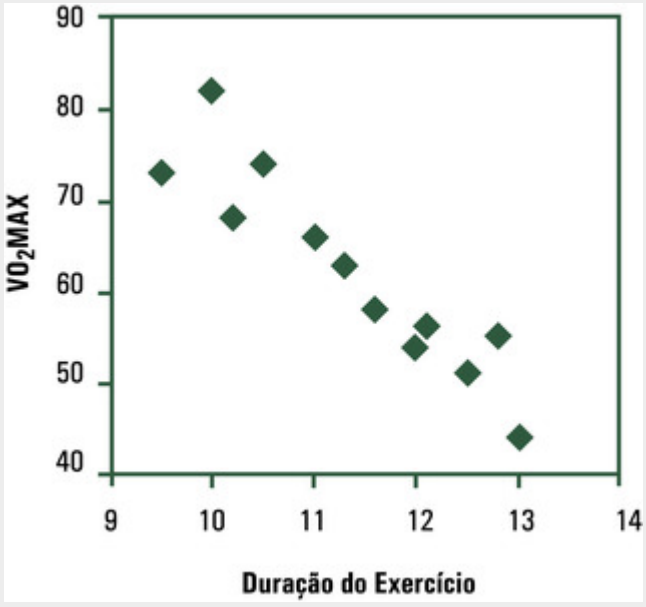
5.6.11. Não é possível[1] calcular o valor de r, mas, obviamente, não existe correlação entre as variáveis: X cresce e Y permanece constante.

5.6.12. $\Sigma x = 255$, $\Sigma x^2 = 9443$, $\Sigma y = 17{,}25$, $\Sigma y^2 = 50{,}4375$, $\Sigma xy = 660{,}25$. Logo, $r = 0{,}913$.

5.6.13. 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.

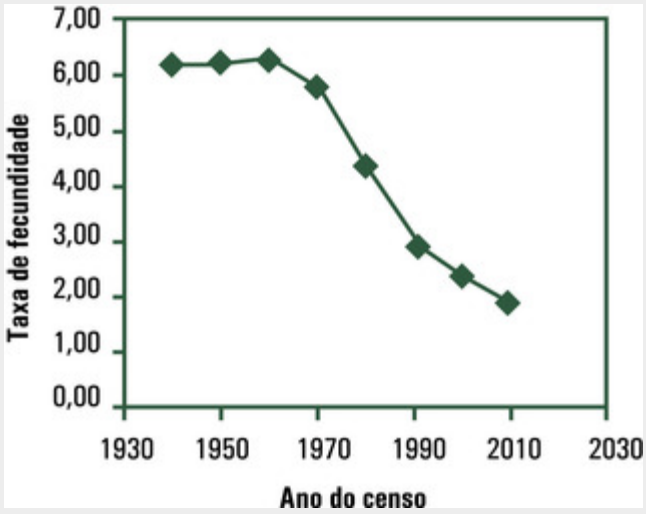
5.6.14. O valor de r é 0,774 (correlação positiva alta).

5.6.15. Duração do exercício, em minutos, e VO_2 MAX em mililitros por quilograma por minuto para 12 homens saudáveis.



Olhando o diagrama, é razoável afirmar que VO_2MAX diminui quando aumenta o tempo da atividade.

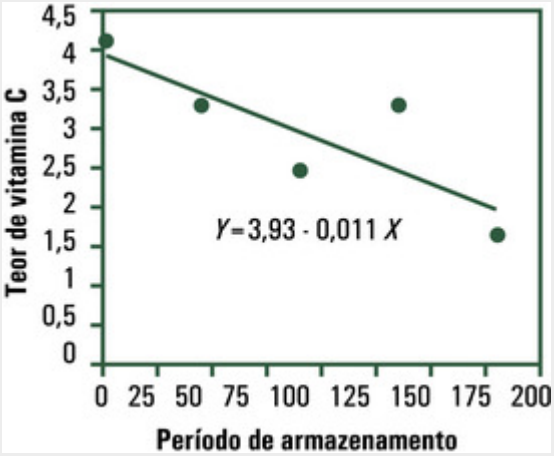
5.6.16. Taxas de fecundidade total no Brasil, segundo o ano do censo



[1]Divisão por zero, uma vez que a variância de Y, que aparece no denominador, é zero.

Capítulo 6: Noções sobre Regressão

6.7.1. 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.

O coeficiente de correlação.

6.7.2. 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.

6.7.3. Não.

6.7.4. $\hat{Y} = 5 + X$

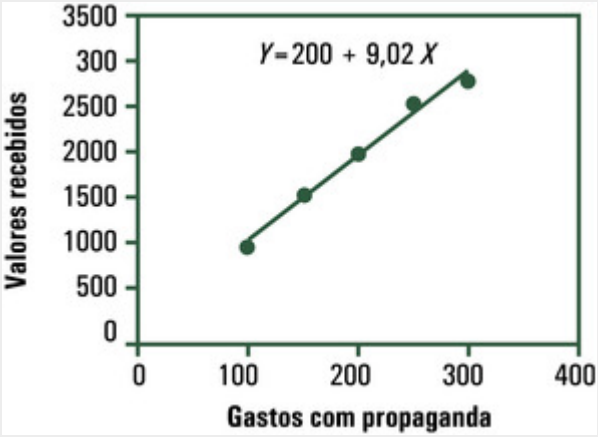
6.7.5. Não seria possível achar o valor de b pela fórmula, uma vez que o denominador seria zero. Mas a ideia é de uma reta paralela ao eixo das ordenadas.

6.7.6. Os dados são poucos para discutir um assunto tão complexo, mas, em geral, é possível afirmar que escolaridade está associada a nível de renda, que significa maiores gastos com produtos de higiene e maior busca por profissionais de saúde, além da facilidade de ter e buscar novos conhecimentos. De qualquer forma, ensinar métodos preventivos produz bons resultados. O que não se pode é usar estatísticas de má qualidade, ainda que se tenha por

objetivo "provar" assuntos já comprovados ou demonstrar boas intenções.

6.7.7. 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.



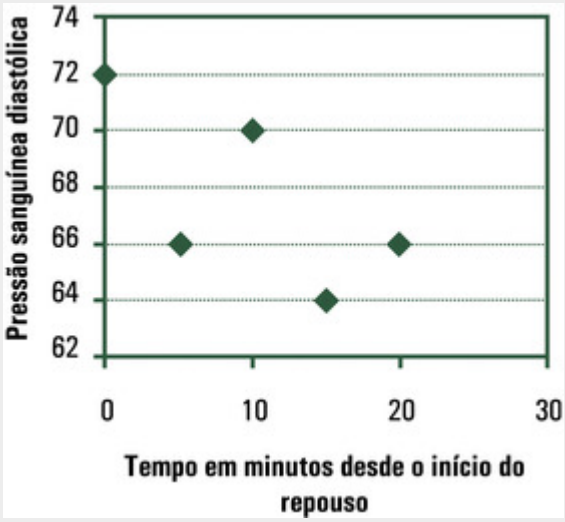
Gastos com propaganda, em reais, na semana, e valores recebidos, em reais, nas vendas.

6.7.8. $\hat{Y} = 11{,}23 + 1{,}309X$

6.7.9. $\hat{Y} = 162{,}5 - 8{,}841X$

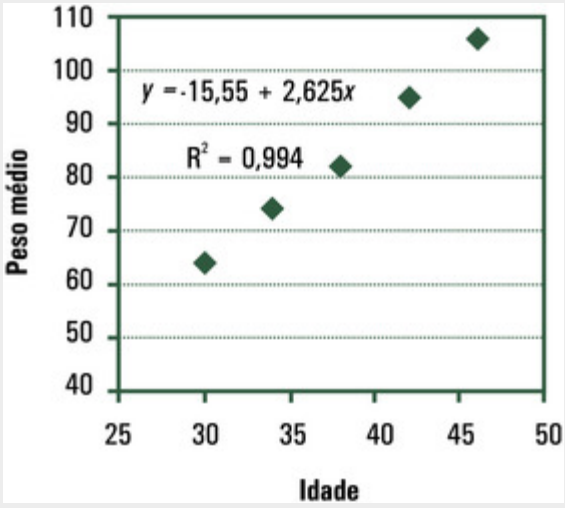
Sim, existe tendência de queda. O coeficiente de determinação é $R^2 = 0{,}859$. Então, o VO_2MAX inalado diminui linearmente quando aumenta a atividade, no intervalo estudado.

6.7.10. Tempo em minutos desde o início do repouso e pressão sanguínea diastólica, em milímetros de mercúrio.



A simples inspeção do gráfico mostra que a pressão sanguínea diastólica diminui com o tempo de repouso, mas há outros fatores que explicam a variação. A maior crítica, aqui, é pelo fato de as observações feitas ao longo do tempo não serem independentes (foram tomadas na mesma pessoa, ao longo do tempo). Para se ajustar uma reta de regressão aos dados, é preciso que as observações sejam *independentes*.

6.7.11. Para 32 dias, a estimativa é 68,85 g.



6.7.12. A regressão *exponencial* traz a variável explanatória no expoente. Escreve-se:

$$Y = a\, e^{bX}$$

Para ajustá-la, é preciso calcular o logaritmo neperiano de Y. Ajusta-se:

$$\ln Y = \ln a + bX$$

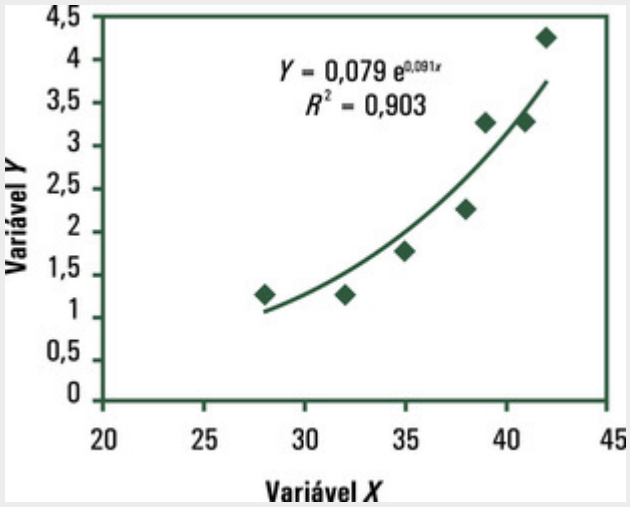
Cálculos auxiliares

X	Y	lnY	X ln Y	X^2
28	1,25	0,22314	6,24802	784
32	1,25	0,22314	7,14059	1024
35	1,75	0,55962	19,58655	1225
38	2,25	0,81093	30,81535	1444
39	3,25	1,17865	45,96754	1521
41	3,25	1,17865	48,32485	1681
42	4,25	1,44692	60,77060	1764
255		5,62106	218,85351	9443

Aplicando as fórmulas, obtém-se:

$$\ln \hat{Y} = -2,535 + 0,09164 \ln X$$

$$\hat{Y} = 0,0792 e^{0,0916x}$$



Equação exponencial ajustada aos dados das variáveis *X* e *Y*

Capítulo 7: Noções sobre Amostragem

7.8.1. 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.

7.8.2. Podem ser selecionados: a) os elementos de ordem par; b) os elementos de ordem ímpar; c) os quatro primeiros elementos.

7.8.3. Numeram-se os alunos e sorteiam-se seis.

7.8.4. Divida dez por cinco e obterá dois. Sorteie um dos dois primeiros números, ou seja, 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.

7.8.5. 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.

7.8.6. Questão fechada: Você costuma escovar os dentes todos os dias?

Sim; Não.

Questão aberta: Como você limpa seus dentes?

7.8.7. 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.

7.8.8. Leitores de livros técnicos.

7.8.9. O costume é escolher uma cidade "representativa" de todo o estado.

7.8.10. a) qualquer conjunto de dez 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).

Capítulo 8: Distribuição Normal

8.9.1. De acordo com a regra empírica, 95% dos dados estarão no entorno da média, a menos de dois desvios padrões de distância da média μ. No caso, dois desvios padrões valem 2 x 15 = 30. A proporção de pessoas com quociente de inteligência acima da média, que é 100, é 95/2 = 47,5. Então, 2,5% de pessoas têm quociente de inteligência acima de 130.

8.9.2. 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, 139,5 + 3 =142,5); média mais dois desvios padrões (no caso, 139,5 + 2 x 3 =145,5) seria o ponto de corte para dizer que a concentração de sódio no plasma de uma pessoa está além do limite de normalidade.

8.9.3. a) ±0,67; b) ±1,64; c) ±1,96.

8.9.4. a) 78,88%; b) 10,56%.

8.9.5. a) 4,75%; b) 45,25%.

8.9.6. a) 97,72%; b) 2,28%.

8.9.7. a) 21,19%; b) 21,19%.

8.9.8. a) 0,1587 ou 15,87%; *b*) 0,0228 ou 2,28%; *c*) 0,5 ou 50%; *d*) 0,1003 ou, aproximadamente 10%.

8.9.9. Sim, metade dos escores é positiva e metade é negativa, porque a distribuição normal reduzida é simétrica em torno da média zero.

8.9.10. 0,0475 ou 4,75%

Capítulo 9: Intervalo de Confiança

9.7.1. a) Se forem tomadas repetidamente muitas amostras e calculados seus intervalos de confiança, 95% deles devem conter a média.

9.7.2. Resposta: falso, pois podem ser obtidos para qualquer parâmetro, usando os dados de uma amostra.

9.7.3. O intervalo de 90% de confiança obtido para a média da pressão sanguínea sistólica (em mm Hg) de uma amostra de cem indivíduos sadios com idade entre 20 e 25 anos é

$$121,7 < \mu < 124,3$$

9.7.4. O intervalo de 95% de confiança calculado para a média de Hb (em g/dL) medida em uma amostra de duzentas mulheres adultas sadias é

$$11,84 < \mu < 16,16$$

9.7.5. O intervalo de 90% de confiança calculado para a média de comprimento (em cm) ao nascer para o sexo masculino, dos filhos de mães sadias com período completo de gestação, foi

$$49,20 < \mu < 50,80$$

9.7.6. O intervalo de 95% de confiança calculado para a média de glicose por 100 mL de sangue em uma amostra de 25

normoglicêmicos é

$$85,32 < \mu < 104,68$$

9.7.7. A amostra de trinta homens sadios com idade entre 30 e 48 anos, não fumantes e que tinham atividade física regular forneceu, em repouso, o intervalo de 95% de confiança para a média de frequência cardíaca

$$61,2 < \mu < 66,6$$

9.7.8. A estimativa por intervalo da média da quantidade de gordura em cem hambúrgueres de determinada cadeia de restaurantes, com 95% de confiança, é

$$31,0 < \mu < 29,4$$

9.7.9. A estimativa por intervalo da média da quantidade de gordura em cem hambúrgueres de determinada cadeia de restaurantes, com 95% de confiança, é

$$645,7 < \mu < 670,3$$

9.7.10.

a. não necessariamente;
b. sim;
c. não necessariamente;
d. não.

Capítulo 10: Teste *t* para uma Amostra

10.4.1. Hipóteses

a. chove;
b. não chove.

Decisões possíveis

a. levar o guarda-chuva;
b. não levar o guarda-chuva.

Erros possíveis

a. chover e não ter guarda-chuva;
b. não chover e carregar o guarda-chuva.

10.4.2. Hipótese da nulidade: o peso médio ao nascer de filhos de gestantes que vivem em extrema pobreza e participaram do programa é *igual* ao peso médio ao nascer histórico (μ = 2.800 g) de filhos de gestantes que vivem em extrema pobreza e não participaram do programa.

Hipótese alternativa: o peso médio ao nascer de filhos de gestantes que vivem em extrema pobreza e participaram do programa é *diferente* do peso médio ao nascer histórico (μ = 2.800 g) de filhos de gestantes que vivem em extrema pobreza e não participaram do programa.

Nível de significância de 5%.

Considerando-se peso médio ao nascer de 3.075 g e desvio padrão 500 g na amostra de 25 mulheres, calcula-se o valor de *t*:

$$t = \frac{\mu - 2800}{\frac{s}{\sqrt{n}}} = \frac{3075 - 2800}{\frac{500}{\sqrt{25}}} == 2,75$$

Com n -1 = 25-1 = 24 graus de liberdade, o valor crítico na tabela de t para um teste bilateral é 2,064. Como o valor absoluto de t calculado é maior que o da tabela, rejeita-se a hipótese da nulidade, ou seja, o peso médio ao nascer de filhos de gestantes que vivem em extrema pobreza e participaram do programa é *diferente* do peso médio ao nascer histórico (μ = 2.800 g) de filhos de gestantes que vivem em extrema pobreza e não participaram do programa.

10.4.3. Estabeleça as hipóteses:

$$H_0 : \mu = 7.0$$
$$H_1 : \mu \geq 7.0$$

Calcule a média e o desvio padrão:

$$\overline{x} = \frac{\Sigma x}{n} = \frac{47,5}{6} = 7,917$$

$$s^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1} = \frac{384,71 - 376,0417}{6-1} = 1,7337$$

$$s = 1,317$$

Calcule o valor de *t*:

$$t = \frac{\mu - 7,00}{\frac{s}{\sqrt{n}}} = \frac{7,92 - 7,00}{\frac{1,32}{\sqrt{6}}} = 1,71$$

Como a hipótese da nulidade será rejeitada apenas em uma direção, o teste é unilateral. Com $n - 1 = 6 - 1 = 5$ graus de liberdade, o valor crítico na tabela de *t* no nível de 10% é 1,476. Como o valor absoluto de *t* calculado é maior que o da tabela, rejeita-se a hipótese da nulidade no nível de 5%, ou seja, em média, as notas dos alunos são significantemente maiores do que o valor especificado.

10.4.4. Estabeleça as seguintes hipóteses:

$$\mathrm{H}_0 : \mu = 4,7$$
$$\mathrm{H}_1 : \mu > 4,7$$

Calcule o valor de *t*:

$$t = \frac{\mu - 4,7}{\frac{s}{\sqrt{n}}} = \frac{5,0 - 4,7}{\frac{0,452}{\sqrt{22}}} = 3,11$$

A hipótese da nulidade será rejeitada apenas em uma direção; o teste é unilateral. Com $n - 1 = 22 - 1 = 21$ graus de liberdade, o valor crítico na tabela de t no nível de 5% é 1,721. Como o valor absoluto de t calculado é maior que o da tabela, rejeita-se a hipótese da nulidade, ou seja, as notas dos alunos são, em média, significantemente maiores do que o valor especificado.

10.4.5. Estabeleça as seguintes hipóteses:

$$H_0 : \mu = 90$$
$$H_1 : \mu \neq 90$$

Estabeleça o nível de significância: 5%

Calcule o valor de t:

$$t = \frac{\mu - 90,00}{\frac{s}{\sqrt{n}}} = \frac{92 - 90}{\frac{14}{\sqrt{100}}} == 1,43$$

Com $n - 1 = 22 - 1 = 21$ graus de liberdade, o valor crítico na tabela de t no nível de 5% é 1,721. Como o valor absoluto de t calculado é menor que o da tabela, não se rejeita a hipótese da nulidade, ou seja, não se pode concluir que, em média, o escore para depressão seja menor em crianças com baixa estima do que nas crianças em geral.

10.4.6. Estabeleça as hipóteses:

$$H_0 : \mu = 20$$
$$H_1 : \mu \neq 20$$

Estabeleça o nível de significância: 5%
Calcule o valor de *t*:

$$t = \frac{\mu - 20}{\frac{s}{\sqrt{n}}} = \frac{18 - 20}{\frac{9}{\sqrt{81}}} == -2,0a$$

Com *n* -1 = 81-1 = 80 graus de liberdade, o valor crítico na tabela de *t* no nível de 5% é 1,960. Como o valor absoluto de *t* calculado é maior que o da tabela, rejeita-se a hipótese da nulidade, ou seja, em média, a terapia proposta reduz a ansiedade em alunos do curso fundamental.

10.4.7. Usando o Minitab, *p* = 0,074 < 0,10. Rejeita-se a hipótese da nulidade.

One-Sample T: Notas dos alunos

Test of μ = 7 vs > 7							
Variable	N	Mean	StDev	SE Mean	90% Lower Bound	T	P
Notas dos alunos	6	7,917	1,317	0,538	7,123	1,71	0,074

10.4.8. Errado. Um teste estatístico não faz hipóteses sobre médias de amostras. O teste *t* para uma amostra é usado para verificar se a média da população de onde a amostra proveio é significativamente diferente de um valor especificado.

10.4.9.

Test of μ = 7 vs ≠ 7							
Variable	N	Mean	StDev	SE Mean	95% CI	T	P
Notas dos alunos	6	7,917	1,317	0,538	(6,535; 9,298)	1,71	0,149

10.4.10. O *p*-valor calculado usando o programa Minitab é 1,00. Não se rejeita a hipótese de que a média dos escores seja 5,0.

One-Sample T: Escore

Test of μ = 5 vs ≠ 5							
Variable	N	Mean	StDev	SE Mean	95% CI	T	P
Notas dos alunos	22	5,0000	0,4629	0,0987	(4,7948; 5,2052)	0,00	1,000

10.4.11. A hipótese da nulidade é a de que, em média, o tempo de alívio de dor é 100 minutos, como acontece com as outras formulações. A hipótese alternativa é a de que o tempo médio para alívio de dor é diferente de 100 minutos.

$$H_0 : \mu = 100 \text{ minutos}$$
$$H_1 : \mu \neq 100 \text{ minutos}$$

Para um teste bilateral no nível de 5% de significância, temos que a média é 98,1; a variância, 21,87778; o desvio padrão, 4,67737; a variância da média, 2,18778; o erro padrão da média, 1,47911; o valor de *t*, -1,28455; e o *p*-valor é 0,231026. O tempo médio de alívio da dor com a nova formulação não difere estatisticamente do tempo médio de outras formulações ($p > 0,05$)

Capítulo 11: Teste t para Comparação de Médias

11.4.1. Médias e desvios padrões de pesos de ratos

Estatísticas	Ração	
	Padrão	Experimental
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 dieta de ração experimental ganharam mais peso.

11.4.2. Observações pareadas; *t* = 4,226, significante no nível de 5%. O teste B dá, em média, resultados significativamente maiores de QI do que o teste A.

11.4.3. *t* = 1,642, não significante a 5%. Os dados não mostram que o uso de anticoncepcionais orais aumente a pressão sanguínea sistólica.

11.4.4. *t* = 0,623, não significante a 5%. Os dados não mostram diferença de peso ao nascer entre sexos.

11.4.5. Médias, variâncias e desvios padrões da pressão sanguínea dos ratos

Valores de *F* e *t*

Estatística	Temperatura	
	26°C	5°C
Média	133,17	165,83
Variância	136,97	218,17
Desvio padrão	11,70	14,77
Valor de F	1,24 n.s.	
Valor de t (unilateral)	4,25*	

Nota: n.s pt indica não significância e o asterisco indica significância no nível de 5%.

Não se rejeita a hipótese de variâncias iguais ($p > 0,05$). Rejeita-se a hipótese de médias iguais ($p < 0,05$). A pressão sanguínea dos ratos ficou mais baixa em baixa temperatura.

11.4.6. Estatísticas para comparar o tempo despendido pelas drogas

Estatística	Resultado
Valor de *F*	1,16
p-valor	0,4097
Variância ponderada	17,457
Valor de *t*	2,99
p-valor (bilateral)	0,0097

Não se rejeita a hipótese de variâncias iguais ($p > 0,05$). Rejeita-se a hipótese de médias iguais ($p = 0,00974 < 0,05$).

11.4.7. Estatísticas para comparar o tempo de alívio da dor obtido com a droga A (nova) em relação à droga B (mais usada)

Estatística	**Resultado**
Valor de *F*	1,33
p-valor	0,2644
Variância ponderada	2,003
Valor de *t*	-1,16
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 ($p > 0,05$).

11.4.8. Estatísticas para comparar os dois métodos de processamento

Estatística	**Resultado**
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 ($p = 0,0000 < 0,05$).

11.4.9. Estatísticas para comparar as duas dietas

Estatística	**Resultado**
Valor de *F*	1,18
p-valor	0,4290
Variância ponderada	2,183
Valor de *t*	-2,34
p-valor (unilateral)	0,0205

Não se rejeita a hipótese de variâncias iguais ($p > 0,05$). Rejeita-se a hipótese de médias iguais ($p = 0,0205 < 0,05$).

11.4.10. Teste *t* pareado, porque a mesma criança foi observada em duas ocasiões: 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.

Capítulo 12: Teste χ^2

12.7.1. Um teste de qui-quadrado no nível de 5% de significância não rejeita a hipótese de que a proporção de recém-nascidos com defeito ou doença séria seja de 3%.

12.7.2. χ^2= 4,82. A proporção de recém-nascidos portadores de anomalia congênita é maior no sexo feminino.

12.7.3. χ^2= 9,04. A ausência congênita de dentes ocorre com mais frequência em meninas.

12.7.4. O coeficiente gama é -0,372. A associação positiva entre anodontia e sexo feminino, na ordem de 37%, é pequena.

12.7.5. χ^2= 1,32. A associação é -0,22, pequena. O teste não rejeita a hipótese de que a presença de aberração cromossômica no feto não depende de a faixa de idade da gestante ser de 35 a 40 anos, ou de 40 anos ou mais.

12.7.6. 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.

12.7.7. Hipótese da nulidade: a probabilidade de natimorto é idêntica para ambos os sexos. Hipótese alternativa: a probabilidade de natimorto é maior para um dos sexos. α = 5%; χ^2 =1,15, portanto não se rejeita H_0.

12.7.8. O coeficiente gama é 0,0816. Associação praticamente inexistente.

12.7.9. Hipótese da nulidade: a probabilidade de dormir mais de oito horas é idêntica para as duas faixas etárias; hipótese alternativa: a probabilidade de dormir mais de oito horas é diferente para as duas faixas etárias no nível de 1% de significância; χ^2 =22,26, portanto se rejeita H_0.

12.7.10. $\chi^2 = 48{,}24$; rejeita-se H_0 no nível de 1%.

Apêndices

Capítulo 13: Probabilidades

13.8.1.

a) $\frac{4}{52} = \frac{1}{13}$

b) $\frac{13}{52} = \frac{1}{4}$

c) $\frac{1}{52}$

13.8.2.

a) $\frac{2}{9}$

b) $\frac{6}{9} = \frac{1}{3}$

c) $\frac{2}{9}$

13.8.3.

a) $\frac{7}{15}$

b) $\frac{8}{15}$

c) zero

13.8.4. 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) $\frac{3}{10}$

13.8.5.

a) $\frac{1}{6}$

b) $\frac{1}{6}$

13.8.6. 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(Reprovado Português) = 0,10
P(Reprovado Matemática) = 0,20
P(Reprovado Português ∩ Reprovado Matemática) = 0,05

$$0{,}05 \neq 0{,}10 \times 0{,}20$$

13.8.7. a) 50% b) 50%
13.8.8. 0,1%
13.8.9. 50%
13.8.10. a) 36% b) 1%

Capítulo 14: Distribuição Binomial

14.6.1. Eventos e respectivos resultados no jogo

Eventos	Resultados possíveis
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.

14.6.2. 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

14.6.3. $\mu = 5$, $\sigma^2 = 2,5$
14.6.4. $\mu = 2$, $\sigma^2 = 1,6$
14.6.5. 2,7%
14.6.6. 27/64 ou 42,2%
14.6.7. 0,001%
14.6.8. a) as respostas têm distribuição binomial; b) depende da taxa de respostas, que deve ser igual ou superior a 70%, ou seja, 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 ofender seus colegas). Nesse caso, as respostas poderiam, eventualmente, ser tendenciosas, e a taxa de respostas, pequena.
14.6.9. 0,59049.
14.6.10. Se considerarmos cada dia um ensaio, em cada dia podem ocorrer mais de dois eventos (ocorreu acidente ou não). Interessa saber o número de acidentes por dia e, em seguida, também o estudo da distribuição de frequências: em quantos dias houve um acidente, dois, três etc. e o estudo das respectivas causas.

Eventos e respectivos resultados no jogo

Eventos	Resultados possíveis
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.

14.6.2. 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

14.6.3. $\mu = 5$, $\sigma^2 = 2{,}5$

14.6.4. $\mu = 2$, $\sigma^2 = 1{,}6$

14.6.5. 2,7%

14.6.6. 27/64 ou 42,2%

14.6.7. 0,001%

14.6.8. 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 ofender seus colegas). Nesse caso, as respostas poderiam, eventualmente, ser tendenciosas, e a taxa de respostas, pequena.

14.6.9. 0,59049.

14.6.10. Se considerarmos cada dia um ensaio, em cada dia podem ocorrer mais de dois eventos (ocorreu acidente ou não).

Interessa saber o número de acidentes por dia e, em seguida, também o estudo da distribuição de frequências: em quantos dias houve um acidente, dois, três etc. e o estudo das respectivas causas.

Sugestões para leitura

Aliaga, M., Gunderson, B. *Interactive Statistics*, 2 ed. New Jersey: Prentice Hall; 2003.

Armitage, P. *Statistical methods in medical research*, 4 ed. Oxford: Blackwel Scientific Publications; 2002.

Bland, M. *An introduction to medical statistics*, 3 ed. Oxford: Oxford Medical Publications; 2000.

Brown, B. W., Hollander, M. *Statistics: a biomedical introduction*. New York: Wiley; 1977.

Bishop, V. M.M., et al. *Discrete multivariate analysis, theory and practice*. Cambridge: MIT Press; 1977.

Bussab, W., Morettin, P. A. *Estatística Básica*. São Paulo: Saraiva; 2002.

Cochran, W. *Sampling techniques*. New York: Wiley; 1977.

Chow, S. C., 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*, 10 ed. New York: Wiley; 2013.

Dawson, B., Trapp, R. G. *Bioestatística básica e clínica*, 3 ed. Rio de Janeiro: McGraw; 1994.

Dean, A., Voss, D. *Design and analysis of experiments*. New York: Springer; 1999.

Elston, R. C., Johnson, W. D. *Essentials of biostatistics*. Philadelphia: F.A. Davis Company; 1994.

Freund, J. E., E Smith, R. M. *Statistics: a first course*, 4 ed. Englewood Cliffs: Prentice Hall; 1986.

Glantz, S. A. *Primer of biostatistics*, 7 ed. New York: McGraw; 2011.

Johnson, R., E Tsui, K. W. *Statistical reasoning and methods*. New York: Wiley; 1998.

Lohr, S. L. *Sampling: Design and analysis*, 2 ed. Pacific Grove: Brooks; 2010.

Matthews, D. E., Farewell, V. *Using and understanding medical statistics*, 4 ed. New York: Karger; 2007.

Minium, E. W., Clarke, R. C., Coladarci, T. *Elements of Statistical Reasoning*, 2 ed. New York: Wiley; 1999.

Motulsky, H. *Intuitive Biostatistics*. New York: Oxford Press; 1995.

Ott, L., Mendenhall, W. *Understanding Statistics*, 6 ed. Belmont: Wadsworth; 1994.

Schork, M. A., Remington, R. D. *Statistics with applications to the biological and health sciences*, 3 ed. New Jersey: Prentice Hall; 2000.

Vieira, S. *Elementos de Estatística*, 5 ed. São Paulo: Atlas; 2012.

Vieira, S. *Bioestatística: Tópicos Avançados*, 2 ed. Rio de Janeiro: Campus-Elsevier; 2008. [5ª tiragem].

Vieira, S., E Hossne, W. S. *Metodologia científica para a área de saúde*, 2 ed. São Paulo, Rio de Janeiro: Elsevier; 2015.

Vieira, S. *Análise de variância*. São Paulo: Atlas; 2006.

Vieira, S., Hossne, W. S. *Experimentação com seres humanos*, 3 ed. São Paulo: Moderna; 1988.

Zar, J. H. *Biostatistical analysis*, 5 ed. New Jersey: Prentice Hall; 2010.

Índice remissivo

A

Ajuste de regressão não linear, 85

Amostra, 91

- casual simples, 93
- estratificada, 94
- não probabilística ou de conveniência, 97
- por conglomerados, 95
- por quotas, 96
- probabilística, 93
- semiprobabilística, 95
- sistemática, 95
- tendenciosa, 99

Amplitude, 43

Análise combinatória, 199

Apuração de dados, 2

Áreas sob a curva normal, 108

Avaliação das técnicas de amostragem, 97

C

Cabeçalho, tabela, 4

Cálculo

- da razão de chances, 170
- da variância, 47
- das probabilidades sob a distribuição normal, 111
- de probabilidade, 182

do coeficiente de correlação, 63
do intervalo de confiança para uma média, 121
do número de classes, 11
dos coeficientes de regressão, 78
Caracterização da distribuição binomial, 197
Caudas da curva, 106
Censo, 92
Chances, 169
Classe modal, 35
Coeficiente(s)
angular da reta, 77
de correlação, 63
de correlação de Pearson, 63
de determinação, 81, 82
de regressão, cálculo dos, 78
de variação, 52
de Yule, 160
fi, 160
gama, 160
linear da reta, 76
Colunas, tabela, 4
Comparação de duas médias, 139
Condição de independência, 187
Confiança, 122
Conglomerados, 95
Construção de tabelas, 3
Correção de continuidade, 172
Correlação
de Pearson, coeficiente de, 83
forte, 60
fraca, 60

negativa, 61
nula, 60
positiva, 61

D

Dado(s), 23
apuração de, 2
contínuos, 9
discrepantes, 34
discretos, 8, 9
estatístico, 1
numéricos, apresentação de, 4, 87
pareados, 140
qualitativos, 19
quantitativos, 8, 24
Desfecho, 66
Desvio médio, 48
Desvio padrão, 47, 51
Diagrama
de caixa (Box plot), 47
de dispersão, 59
de linhas, 24
Dispersão
dos dados em relação à média, 53
relativa, 53
Distância interquartílica, 46
Distribuição
binomial, 195, 197, 198
das médias das amostras, 120
de frequências, 5, 8, 9, 31
de Gauss, 104
de probabilidades, 194, 198

teórica, 103
Distribuição normal, 103
cálculo das probabilidades, 111
características, 104
probabilidades associadas à, 106
reduzida ou padronizada, 107
usos da, 112

E

Ensaio
clínico, 161
com dados pareados, 140
Equação da reta, 76
Erro(s), 130
definindo os, 130
padrão da média, 117, 119
tipo I, 130
Escolha da variável explanatória, 80
Espaço amostral, 179
Estatística, 1, 91
Estimativa(s)
da média por intervalo, 123
da média por ponto, 123
da variável resposta, 79
de risco, 169
por ponto, 117
Estudo
prospectivo, 164
retrospectivo, 166
Evento(s), 179
dependentes, 186
impossíveis, 181

independentes, 185
não mutuamente exclusivos, 184
Extração de raiz quadrada, 86
Extrapolação, 79
Extremos de classe, 10

F

Falácia, 82
Fator, 66
de risco, 168
Frequência relativa, 6, 183

G

Gerador de números aleatórios, 93
Gráfico
de linhas, 66
de série temporal, 66
de barras, 19
de pontos, 25
de setores, 22
Grau
de associação, 160
de correlação linear, 63
de dispersão das médias das amostras, 118
de liberdade, 49, 122

H

Hipótese(s), 128
alternativa, 129
da nulidade, 129
Histograma, 25

I

Inferência, 117, 127
  estatística, 130
Intervalo(s)
  de classe, 10
  de confiança, 117
    interpretação dos, 124
Inversão, 86

L

Levantamento de dados, 1
Limites dos intervalos de classe, 10
Logaritmo neperiano da velocidade, 86

M

Margens de erro, 91, 121
Máximo, 43
Média
  aritmética, 30
  da amostra, 117
  da população, 117, 118
  dos desvios, 48
  na distribuição binomial, 199
Mediana, 33
Medida(s)
  de associação, 160
  de dispersão, 43
  de tendência central, 29
  de variabilidade, 43
Métodos de amostragem, 93
Mínimo, 43
Moda, 35

N

Nível
  de confiança, 122
  de significância, 122, 131
Notação de somatório, 30
Número de classes, 11

P

p–valor, 133
Parâmetro, 91
Polígonos de frequências, 26
População(ões), 91
  -alvo, 91
  configurada, 91
  independentes, 143
Prevalência, 171
Probabilidade, 179
  cálculo de, 111
  condicional, 186
  definições de, 181, 182
  distribuição de, 194
  na distribuição normal reduzida, 107
  na distribuição normal, 106
  subjetiva, 183
Proporção, 171

Q

Qualidade de uma estimativa, 98
Quartil, 44

R

Razão(ões)

de chances, 168, 169
para o uso de amostras, 92
Regra
do “e”, 185
do “ou”, 183
empírica, 107
Regressão, 75
linear simples, 76, 87
não linear, 83
Relação(ões)
determinísticas, 81
linear, 75
probabilísticas, 81
Representatividade, 99
Reta de regressão, 76
Risco relativo, 168

S

Símbolos matemáticos, 29
Soma
de eventos mutuamente exclusivos, 183
de eventos não mutuamente exclusivos, 184
de quadrados dos desvios, 48
de variáveis aleatórias independentes, 105
Somatório, notação de, 30

T

Tabela(s)
de distribuição de frequências, 5, 31
dados quantitativos, 8
variância de dados agrupados, 50
de contingência, 7, 157

Tamanho da amostra, 63, 98
Tendência, 99
  central, medidas de, 29
Teorema
  da multiplicação, 185
  da soma, 183
  do limite central, 105, 106
Teoria das probabilidades, 179
Teste
  bilateral, 129
  de hipóteses, 134
  de uma proporção, 171
  dos grupos com base na distribuição normal, 165, 167
  estatístico, 63, 127, 128, 158, 171
  F, 144
  t, 132
    na comparação de grupos independentes, 143
    nos estudos com dados pareados, 139
    para comparar médias, 139, 145
    para uma amostra, 127
  unilateral, 129
  Z nos ensaios clínicos, 163
  χ^2
    nos ensaios clínicos, 162
    nos estudos prospectivos, 164
    nos estudos retrospectivos, 166
    para a associação de duas variáveis, 157
    para comparar dois grupos em ensaios clínicos, 161
Tomada de decisão em condições de incerteza, 127
Transformação
  dos dados, 84

logarítmica, 86

V

Valor
científico, 93
discrepante, 44
máximo, 9
mínimo, 9
Variabilidade, 43
das médias das amostras, 117
Variação conjunta das variáveis, 60
Variância(s), 48
da média, 118
de dados agrupados, 50
na distribuição binomial, 199
desiguais, 147
dos grupos, 143
iguais, 145
Variável, 1
aleatória, 103, 193
aleatória binária, 193
aleatória binomial, 194
explanatória, 66, 80
resposta, 66
Z, 108


BIOESTATÍSTICA
SONIA VIEIRA
TÓPICOS AVANÇADOS
Testes não paramétricos,
Testes diagnósticos,
Medidas de Associação e
Concordância.
ELSEVIER
4ª EDIÇÃO

Bioestatística

Vieira, Sonia

9788535289824

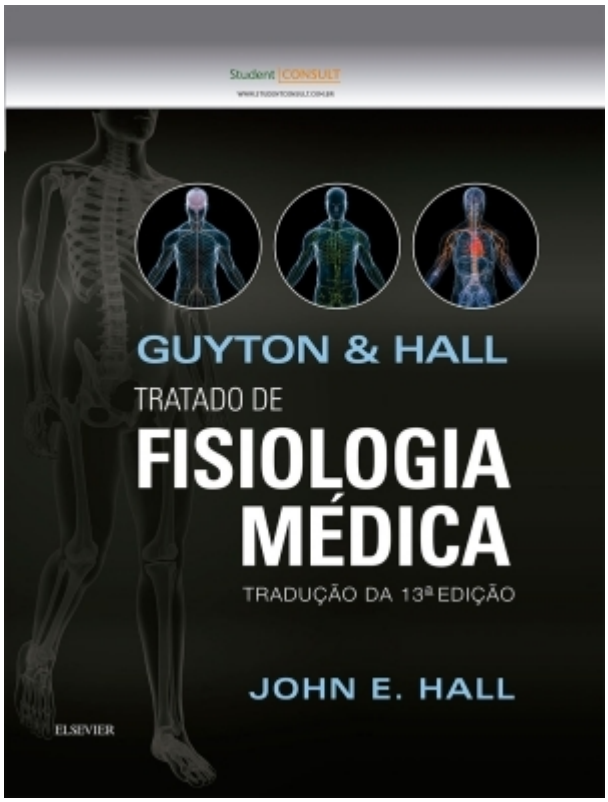
308 páginas

O livro Bioestatística – Tópicos Avançados é mais uma obra indispensável de Sonia Vieira, que leva o leitor a dominar os conceitos progressivamente, rever as próprias ideias e aperfeiçoar a aprendizagem, sempre de modo agradável. A competência e a capacidade da autora de transmitir ideias ficam demonstradas, neste livro, pela disposição dos temas, pela sequência das ideias, pelo didatismo sem prejuízo da profundidade na escolha dos exemplos e dos exercícios. Longe da aridez que se atribui (sem razão, aliás) à Bioestatística, esta obra é capaz de atingir tanto o iniciante como o expert na área. E este livro, como

os outros da autora, caracteriza- se pela precisão de linguagem, como convém ao cientista, elegância de forma, como convém ao professor e conteúdo instigante, como convém ao pesquisador. O livro Bioestatística – Tópicos Avançados explica como interpretar testes de hipóteses e como interpretar os intervalos de confiança. Apresenta os testes não paramétricos, muito usados em artigos da área de saúde. E é dada não apenas a maneira de proceder a tais testes, mas também a lógica deles. A análise e a interpretação de dados apresentados nas tabelas de contingência são tratadas de maneira clara e didática. O livro apresenta, ainda, coeficientes de correlação, coeficientes de associação e coeficiente de concordância e trata a análise de exames para diagnóstico. É, portanto, leitura obrigatória para quem se inicia em pesquisa, para quem já é pesquisador e para quem lê resultados das pesquisas. E mais importante: essa "obrigação se revela um prazer intelectual, pois é uma dessas publicações cuja leitura desperta, ao final, a ansiedade agradável da espera por outro livro da autora.Este livro deve ser visto como complemento

de outro, de nome Introdução à Bioestatística, da mesma autora. Então, tanto os estudantes que se iniciam em Estatística como aqueles que já se profissionalizaram na área verão que este livro é útil, como texto e como material de referência. Escrito para não estatísticos que já tenham tido algum curso dessa matéria, é didático, fácil de ler e explora o uso efetivo de técnicas estatísticas na solução de problemas, usando exemplos publicados na área de saúde em geral, mas especialmente em Medicina e em Odontologia. O livro reflete os muitos anos de ensino e assessoria da autora na área de Estatística. Os numerosos exemplos do texto fazem o estudante trabalhar com dados retirados de uma grande variedade de situações da vida real. Mas o livro busca desenvolver a capacidade de julgamento e não apenas ensinar o aluno a aplicar testes, mecanicamente. Para isso, explica a teoria, depois ensina a resolver um problema e apresenta vários exemplos. No final de cada capítulo, são dados exercícios, todos com respostas. De início, o livro trata os muitos tipos de dados que podem ser coletados na área da saúde. Explica como

interpretar testes de hipóteses e como interpretar os intervalos de confiança. Depois, apresenta as tabelas de contingência e os diversos testes envolvidos na análise e interpretação de tais dados. Explica, então, os testes não paramétricos, atualmente muito usados em artigos especializados. Ainda, apresenta coeficientes de correlação, coeficientes de associação e coeficiente de concordância e trata a análise de exames para diagnóstico.


Student|CONSULT
GUYTON & HALL
TRATADO DE
FISIOLOGIA
MÉDICA
TRADUÇÃO DA 13ª EDIÇÃO
JOHN E. HALL
ELSEVIER

Guyton E Hall Tratado De Fisiologia Médica

Hall, John E.

9788535285543

1176 páginas

A 13ª edição do Guyton & Hall Tratado de Fisiologia Médica mantém a longa tradição deste best-seller como o melhor livro-texto de Fisiologia Médica do mundo. Diferentemente de outros livros, este guia claro e de fácil compreensão tem voz autoral única e consistente e ressalta o conteúdo mais relevante para os estudantes clínicos e pré-clínicos. O texto detalhado, porém esclarecedor, é complementado por ilustrações didáticas que resumem conceitos-chave em fisiologia e fisiopatologia. • O texto com fonte maior enfatiza a informação essencial sobre como o corpo deve manter a homeostasia de modo

a permanecer saudável, ao mesmo tempo em que as informações de apoio e os exemplos são detalhados com tamanho de fonte menor e destacados em lilás. • As figuras e tabelas de resumo transmitem de maneira facilitada os processos chave apresentados no texto. • Contém a nova tabela de referência rápida de valores laboratoriais padrão no final do livro. • Acréscimo do número de figuras, correlações clínicas e mecanismos moleculares e celulares importantes para a medicina clínica. • Inclui o conteúdo online em português do Student Consult, que oferece uma experiência digital aprimorada: banco de imagens, referências, perguntas e respostas e animações. Junto com a nova edição da consagrada referência mundial da fisiologia, Guyton & Hall, você também tem acesso à forma mais inovadora, simples, visual e objetiva de aprender fisiologia, o Homem Virtual, a maneira inteligente de estudar fisiologia em 3D.


TRATADO DE
GINECOLOGIA
FEBRASGO
EDITORES
César Eduardo Fernandes • Marcos Felipe Silva de Sá
COORDENADORES
Agnaldo Lopes da Silva Filho • Luciano de Melo Pompei
Rogério Bonassi Machado • Sérgio Podgaec
ELSEVIER
febrasgo

Tratado de ginecologia Febrasgo

Fernandes, César Eduardo

9788535292145

1024 páginas

Obra referência para as provas da especialidade, certificação e recertificação na área de Ginecologia e Obstetrícia. Chancela Febrasgo. Obra referência para as provas da especialidade.

Compre agora e leia


TRATADO DE
OBSTETRÍCIA
FEBRASGO
EDITORES
César Eduardo Fernandes • Marcos Felipe Silva de Sá
COORDENADORES
Corintio Mariani Neto • Eduardo Cordioli
Olímpio Barbosa de Moraes Filho
ELSEVIER
febrasgo

Tratado de obstetrícia

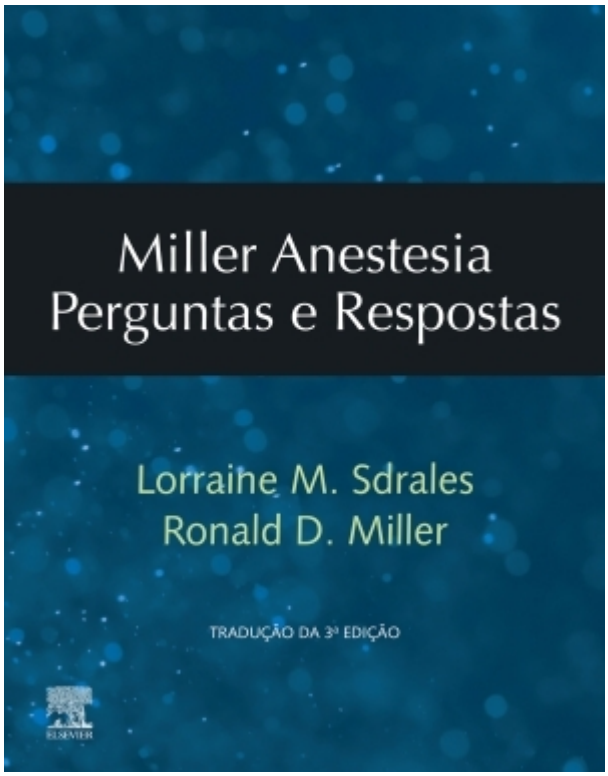
Febrasgo

9788535292213

1024 páginas

Domine o conteúdo da ginecologia e obstetricia e passe nas provas da sociedade com o novo tratado da Febrasgo, um texto de referência para esta importante área. Chancela Febrasgo Referência para as provas da especialidade, certificação e recertificação.

Compre agora e leia


Miller Anestesia
Perguntas e Respostas
Lorraine M. Sdrales
Ronald D. Miller
TRADUÇÃO DA 3ª EDIÇÃO

Miller – Anestesia Perguntas e Respostas

Sdrales, Lorraine M

9788535291537

544 páginas

Miller's Anesthesia Review é um guia de estudo que permite avaliar seus conhecimentos para se preparar para a prova de título, possui mais de 3800 perguntas e respostas comentadas sobre os diversos temas. - Aborda de diversas formas a distribuição da anestesia em vários contextos de acordo com o estado do paciente da doença, praticamente em quase todos os capítulos os autores do livro Bases da Anestesia são os mesmos para o perguntas e respostas, possui mais de 3800 perguntas e respostas comentadas sobre os diversos temas. - Serve para facilitar a

aprendizagem e a retenção de conceitos fundamentais de anestesia que são necessários para uma sólida base de conhecimento e competência clínica