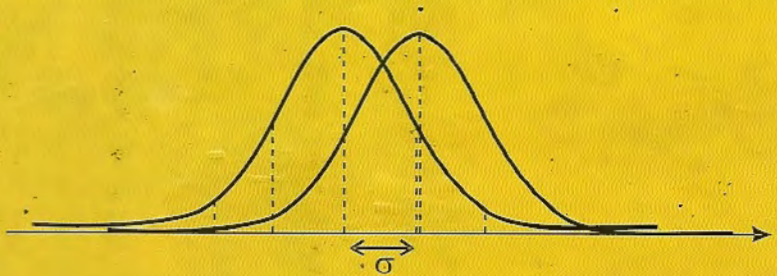
COLEÇÃO DIDÁTICA

ESTATÍSTICA

APLICADA ÀS CIÊNCIAS SOCIAIS

8ª edição revista



PEDRO ALBERTO BARBETTA

editora ufsc

Esta obra surgiu de vários anos de experiência com a atividade de ministrar aulas de Estatística para cursos das áreas de Ciências Sociais e Humanas. Um novo enfoque é aqui desenvolvido, diferenciando este de outros livros didáticos, ao motivar o aprendizado de técnicas estatísticas a partir de situações práticas e desenvolver a capacidade criativa dos alunos com diversos exemplos e exercícios que já apresentam a análise estatística pronta, deixando ao aluno a tarefa de interpretar os resultados. Tudo isso é feito com centenas de figuras, proporcionando um aprendizado mais rápido e agradável.

Pedro Alberto Barbetta

ESTATÍSTICA

APLICADA ÀS CIÊNCIAS SOCIAIS

8ª edição revista

editora ufsc

Direção editorial:
Paulo Roberto da Silva

Capa:
Paulo Roberto da Silva

Revisão:
Maria Geralda Soprana Dias

Ficha Catalográfica
(Catalogação na fonte pela Biblioteca Universitária da
Universidade Federal de Santa Catarina)

G643a Barbetta, Pedro Alberto
Estatística aplicada às Ciências Sociais / Pedro Alberto Barbetta. 8. ed. rev. – Florianópolis : Ed. da UFSC, 2012.

318p. : il. (Coleção Didática)

Inclui bibliografia

1. Estatística. 2. Ciências Sociais. I. Título.

CDU: 31:3
CDD: 300:21

ISBN: 978-85-328-0604-8

Sumário

Glossário de símbolos 9

Prefácio 13

Capítulo 1 – Introdução 15

Parte I – O planejamento da coleta dos dados

Capítulo 2 – Pesquisas e dados 23

2.1 O planejamento de uma pesquisa 24
2.2 Dados e variáveis 29
2.3 Elaboração de um questionário 32
2.4 Uma aplicação 35
2.5 Codificação dos dados 37
Anexo 39

Capítulo 3 – Técnicas de amostragem 41

3.1 Amostragem aleatória simples 45
3.2 Outros tipos de amostragens aleatórias 47
3.3 Amostragens não aleatórias 54
3.4 Tamanho de uma amostra aleatória simples 57
3.5 Fontes de erros nos levantamentos por amostragem 61

Parte II – Descrição e exploração de dados

Capítulo 4 – Dados categorizados 65

4.1 Classificação simples 65
4.2 Representações gráficas 68

4.3 Dupla classificação 71
Anexo 77

Capítulo 5 – Dados quantitativos 79
5.1 Variáveis discretas 79
5.2 Variáveis contínuas 82
5.3 Ramo-e-folhas 88

Capítulo 6 – Medidas descritivas 91
6.1 Média e desvio padrão 91
6.2 Fórmulas para o cálculo de $\overline{X}$ e S 95
6.3 Medidas baseadas na ordenação dos dados 99
6.4 Orientação para análise exploratória de dados 109

Parte III – Modelos de probabilidade

Capítulo 7 – Modelos probabilísticos 115
7.1 Definições básicas 116
7.2 O modelo binomial: caracterização e uso da tabela 126
7.3 O modelo binomial: formulação matemática 129

Capítulo 8 – Distribuições contínuas e modelo normal 133
8.1 Distribuições normais 136
8.2 Tabela da distribuição normal padrão 139
8.3 Dados observados e modelo normal 143
8.4 Aproximação normal à binomial 145

Parte IV – Inferência estatística

Capítulo 9 – Estimação de parâmetros 153
9.1 Distribuição amostral 157
9.2 Estimação de uma proporção 160
9.3 Estimação de uma média 165
9.4 Correções para tamanho de população conhecido 170
9.5 Tamanho mínimo de uma amostra aleatória simples 172

Capítulo 10 – Testes estatísticos de hipóteses 179
10.1 As hipóteses de um teste estatístico 180
10.2 Conceitos básicos 182
10.3 Testes unilaterais e bilaterais 187
10.4 Uso de distribuições aproximadas 189
10.5 Aplicação de testes estatísticos na pesquisa 191

Capítulo 11 – Testes de comparação entre duas amostras 195
11.1 Testes de significância e delineamentos de pesquisa 195
11.2 O teste dos sinais 198
11.3 O Teste t para dados pareados 201
11.4 O teste t para amostras independentes 209
11.5 Tamanho das amostras 217
11.6 Comentários adicionais 219

Parte V – Relação entre variáveis

Capítulo 12 – Análise de dados categorizados 227
12.1 O teste de associação qui-quadrado 228
12.2 Medidas de associação 241

Capítulo 13 – Correlação e regressão 251
13.1 Diagramas de dispersão 252
13.2 O coeficiente de correlação linear de Pearson 254
13.3 Correlação por postos 263
13.4 Regressão linear simples 266
13.5 Análise dos resíduos e transformações 277
13.6 Introdução à regressão múltipla 283
Anexo 288

Referências 289

Apêndice 291

Respostas de alguns exercícios 305

Glossário de símbolos

Letras do alfabeto português

Símbolo	Significado	Seções
a	Estimativa do coeficiente escalar (intercepto) de uma reta de regressão	13.4
b	Estimativa do coeficiente angular de uma reta de regressão	13.4
C	Coeficiente de contingência	12.2
C^*	Coeficiente de contingência modificado	12.2
d_Q	Desvio entre quartis	6.4
E_I	Extremo inferior	6.4
E_S	Extremo superior	6.4
E	Margem de erro/Frequência esperada	9.2, 9.3 / 12.1
E_0	Margem de erro tolerada	3.4, 9.5
gl	Graus de liberdade	9.3, 11.3, 11.4, 12.1
H_0	Hipótese nula	10.1
H_1	Hipótese alternativa	10.2
Md	Mediana	6.3
N	Tamanho (número de elementos) da população	3.1, 3.4
n	Tamanho (número de elementos) da amostra	3.1, 3.4, 7.2
n_0	Valor preliminar no cálculo do tamanho da amostra	3.4, 9.5
O	Frequência observada	12.1
p	Valor p ou probabilidade de significância	10.2
$P(A)$	Probabilidade de ocorrer o evento A	7.1
$p(x)$	Probabilidade de ocorrer o valor x	7.2
Q_I	Quartil inferior	6.4

Símbolo	Significado	Seções
Q_S	Quartil superior	6.4
r	Coeficiente de correlação de Pearson	13.2
r_s	Coeficiente de correlação de Spearman	13.3
R^2	Coeficiente de determinação da equação de regressão	13.4
S	Desvio padrão dos dados / Desvio padrão amostral	6.1, 6.2 / 9.3
S^2	Variância dos dados/ Variância amostral	6.1 / 9.3
S_P	Estimativa do erro padrão da proporção amostral	9.2, 9.4
$S_{\overline{x}}$	Estimativa do erro padrão da média amostral	9.3, 9.4
S_D	Desvio padrão de diferenças	11.3
S_a	Desvio padrão agregado	11.4
S_a^2	Variância agregada	11.4
S_e	Desvio padrão dos resíduos da regressão	13.4
t	Valor da distribuição t de Student / Valor da estatística t	9.3 / 11.3, 11.4
V	Coeficiente de associação de Cramér	12.2
$\overline{X}$	Média aritmética dos valores de X / Média amostral	6.1, 6.2 / 9.3
$\hat{y}$	Valor predito por uma equação de regressão	13.4
z	Valor padronizado / Valor da distribuição normal padrão	8.1/ 9.2

Símbolos matemáticos e letras do alfabeto grego

Símbolo	Significado	Seções
≈	Aproximadamente igual	
±	Mais ou menos	
>	Maior	
≥	Maior ou igual	
<	Menor	
≤	Menor ou igual	
α	Nível de significância de um teste estatístico / Coeficiente escalar (intercepto) de uma reta de regressão	10.2 / 13.4
β	Probabilidade do erro tipo II / Coeficiente angular de uma reta de regressão	10.2 / 13.4
γ	Coeficiente de correlação gama	12.2
ϕ	Coeficiente de associação *phi*	12.2

Símbolo	Significado	Seções
ε	Erro aleatório	13.4
μ	Média do modelo normal / Média populacional	7.1 / 9.3
μ_P	Valor esperado de uma proporção amostral	9.1
Ω	Espaço amostral	7.1
π	Probabilidade (parâmetro de um modelo) / Proporção populacional	7.1, 7.2 / 9.1
χ^2	Estatística qui-quadrado	12.1
σ	Desvio padrão do modelo normal	7.1
σ^2		
σ_P	Erro padrão da proporção amostral	9.1
$\sigma_{\overline{X}}$	Erro padrão da média amostral	9.3
$\sum X$	Soma dos elementos da variável X	6.1
$\sum X^2$	Soma dos quadrados dos elementos da variável X	6.1

Prefácio

Estatística aplicada às Ciências Sociais foi escrito com o objetivo de ser um livro-texto em disciplinas de Estatística para cursos de Ciências Sociais e Humanas. A motivação para escrever este texto surgiu quando aproximamos o ensino da Estatística a problemas práticos nas áreas sociais, inserindo os alunos em pequenos projetos de pesquisa e mostrando-lhes a necessidade do uso de técnicas estatísticas. A motivação e o aproveitamento dos alunos cresceram tanto que resolvemos desenvolver esta abordagem em forma de livro-texto.

Este texto apresenta uma introdução à estatística, acompanhada de uma orientação de como planejar e conduzir uma pesquisa quantitativa. Ao invés de apresentarmos a Estatística com um raciocínio tipicamente matemático, como é usual nos livros-texto de Estatística, optamos por apresentar os conceitos e técnicas dentro de um processo de pesquisa em Ciências Sociais e Humanas. Em geral, os capítulos iniciam com problemas práticos que motivam e justificam a introdução de técnicas estatísticas.

O livro inicia com uma visão geral das técnicas estatísticas e apresenta algumas ideias básicas sobre o planejamento de uma pesquisa social (Capítulos 2 e 3). Os Capítulos 4 a 6 trazem alguns dos principais elementos da Estatística Descritiva e da Análise Exploratória de Dados, incluindo algumas aplicações em pesquisas de campo desenvolvidas em nossa Universidade. Alguns modelos de probabilidades, que serão necessários para o entendimento de capítulos posteriores, são apresentados nos Capítulos 7 e 8. O Capítulo 9 coloca o problema de generalizar resultados da amostra para a população, através de intervalos de confiança, e é aplicado especialmente em pesquisas de levantamento

por amostragem, como nas pesquisas eleitorais. O Capítulo 10, embora enfoque também a questão de generalizar resultados da amostra para a população, o faz através de testes de hipóteses. Os conceitos de testes de hipóteses geralmente são de difícil entendimento, mas, neste livro, apresentamo-los de uma forma que os alunos não costumam ter maiores dificuldades. Os Capítulos 11 e 12 abordam testes de hipóteses e análises estatísticas bastante usados nas Ciências Sociais e Humanas. Finalmente, o Capítulo 13 apresenta procedimentos estatísticos para avaliar a relação entre duas variáveis, assim como desenvolve técnicas para construir modelos voltados para alguns tipos de relações.

Ao longo das várias edições, fomos corrigindo erros, aperfeiçoando o texto, introduzindo novos exemplos e exercícios, além de incluir saídas de pacotes computacionais estatísticos e de planilhas eletrônicas, fazendo com que o presente material sirva também como livro-texto para disciplinas que usam o computador. Nesta sexta edição, reescrevemos o texto com uma linguagem mais direta, procurando melhorar aspectos didáticos e dando maior destaque aos principais conceitos, além de aprimorar a qualidade da apresentação. Também estamos criando uma página na Internet com *slides* baseados no livro, orientações para uso de alguns pacotes computacionais e arquivos de dados para exercícios e trabalhos acadêmicos, além de outras facilidades. Ver www.inf.ufsc.br/~barbetta/livro1.htm.

Finalmente, gostaríamos de agradecer aos colegas professores e alunos que tanto contribuíram para o desenvolvimento deste texto, em especial à Profª Sívia Modesto Nassar, que teve a paciência de ler criteriosamente todo o texto e oferecer contribuições significativas nos Capítulos 2 e 3.

Pedro Alberto Barbetta

Capítulo 1

Introdução

Neste primeiro capítulo, tentaremos oferecer ao leitor uma ideia preliminar do que é *estatística* e como ela pode ser usada em pesquisas, nas áreas das ciências sociais e humanas.

Quem está estudando estatística pela primeira vez deve imaginá-la associada a números, tabelas e gráficos que serão usados no momento de organizar e apresentar os dados de uma pesquisa. Mas, como tentaremos mostrar neste livro, isto não é bem assim! A estatística pode estar presente nas diversas etapas de uma pesquisa social, desde o seu planejamento até a interpretação de seus resultados, podendo, ainda, influenciar na condução do processo da pesquisa. Tomemos o seguinte exemplo para facilitar a nossa discussão.

Exemplo 1.1 Com o objetivo de levantar conhecimentos sobre o *nível de instrução do chefe da casa*, nas famílias residentes no bairro Saco Grande II, Florianópolis – SC, decidiu-se pesquisar algumas destas famílias.[1]

Temos, no Exemplo 1.1, um problema típico de estatística aplicada: conhecer certas características de uma população, com base numa amostra.

População é o conjunto de elementos para os quais desejamos que as nossas conclusões sejam válidas – o *universo* de nosso estudo. Uma parte desses elementos é dita uma **amostra**.

[1] Este problema faz parte de uma pesquisa realizada pela UFSC, 1988. O anexo do Capítulo 4 apresenta parte dos dados coletados.

Coleta de dados

Para conhecermos certas características dos elementos de uma população (ou de uma amostra), precisamos coletar dados desses elementos. É uma fase da pesquisa que precisa ser cuidadosamente planejada para que dos dados a serem levantados se tenham informações que atendam aos objetivos da pesquisa. É no planejamento da obtenção dos dados que devemos também planejar *o que fazer* com eles. Voltaremos a essa discussão nos Capítulos 2 e 3.

No problema apresentado no Exemplo 1.1, os dados foram coletados através de entrevistas, aplicadas numa amostra de 120 famílias. Ao observar o nível de instrução do chefe da casa, o entrevistador classificava a resposta do entrevistado numa das três seguintes categorias: (1) *sem instrução*, (2) *fundamental (primeiro grau)* e (3) *médio (segundo grau)*. Ao coletar os dados desta forma, já se tinha em mente os procedimentos estatísticos que seriam usados na futura análise desses dados.

Descrição e exploração de dados

Depois de observada uma amostra de famílias (Exemplo 1.1), temos à disposição um conjunto de dados acerca da variável *nível de instrução do chefe da casa*. Esses dados devem ser organizados para que possam evidenciar informações relevantes, em termos dos objetivos da pesquisa. Esta etapa é usualmente chamada de *descrição de dados*. Um conceito importante nesta fase do trabalho é o de *distribuição de frequências*.

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

Uma distribuição de frequências do nível de instrução, por exemplo, deve informar *quantas* pessoas (ou a *percentagem* de pessoas) se enquadram em cada categoria preestabelecida. A Figura 1.1 mostra, sob forma gráfica, uma distribuição de frequências.[2] Temos, nesta figura, a informação da percentagem de chefes da casa que estão em cada nível de instrução. Em outras palavras, a Figura 1.1 fornece uma visualização do *perfil do nível educacional dos chefes das casas*, na amostra em estudo.

[2] A construção de distribuições de frequências, assim como suas representações em tabelas e gráficos, serão vistas nos Capítulos 4 e 5.

Figura 1.1 Distribuição de frequências do nível de instrução do chefe da casa. Amostra de 120 famílias do bairro Saco Grande II, Florianópolis-SC, 1988.

A região em estudo é dividida em três localidades: Conjunto Residencial Monte Verde, Conjunto Residencial Parque da Figueira e Encosta do Morro. Considerando que haja interesse em comparar essas três localidades, construímos a Figura 1.2, que apresenta três distribuições de frequências, sendo uma para cada localidade.

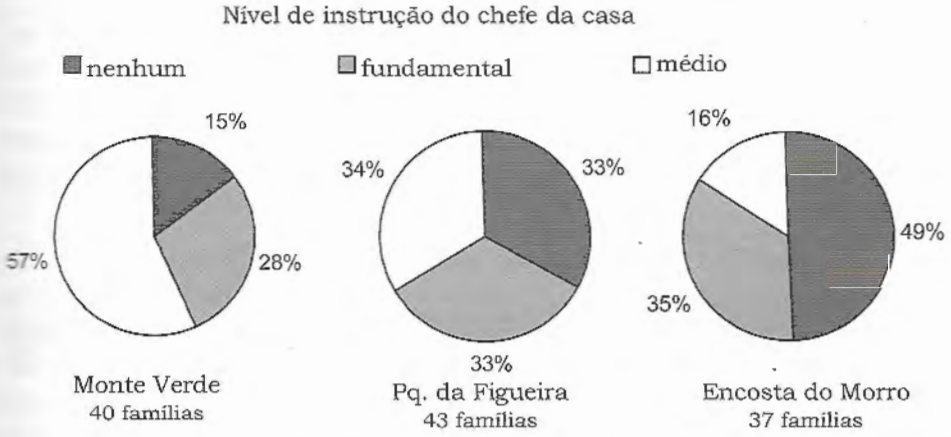


Figura 1.2 Distribuição de frequências do nível de instrução do chefe da casa, por localidade. Amostra de 120 famílias do Bairro Saco Grande II, Florianópolis – SC, 1988.

Ao descrever os dados, começamos a *explorar* como deve ser a população de onde eles foram extraídos. A Figura 1.2, por exemplo, parece sugerir que, na população em estudo, o perfil do nível de instrução do chefe da casa é melhor no Conjunto Residencial Monte Verde e pior na

Encosta do Morro, ficando o Conjunto Residencial Parque da Figueira numa situação intermediária. Este tipo de análise pode ser caracterizado como uma *análise exploratória de dados*, ou seja, uma tentativa de captar a essência das informações contidas nos dados, através da construção de tabelas e gráficos. Em termos mais técnicos, uma *análise exploratória de dados* consiste na busca de um padrão ou modelo que possa nos orientar em análises posteriores.

Inferência estatística

Ao analisar os dados de uma amostra, devemos estar atentos ao fato de que algumas diferenças podem ser meramente *casuais*, ocasionadas por características próprias da amostra, não representando, necessariamente, propriedades da população que gostaríamos de conhecer. Neste contexto, é importante estudarmos os chamados modelos probabilísticos (Capítulos 7 e 8), que são uma forma de mensurar a incerteza. Esses modelos constituem-se na base da metodologia estatística de generalizar resultados de uma amostra para a população de onde ela foi extraída, que pode ser sob a forma de *estimação de parâmetros* ou de *teste de hipóteses*.

Um **parâmetro** é uma medida que descreve certa característica dos elementos da população.

Por exemplo, na população descrita no Exemplo 1.1, *a percentagem de famílias em que o chefe da casa possui nível médio de instrução* é um parâmetro.

Na Figura 1.1, verificamos que, na amostra, a *percentagem de famílias em que o chefe da casa possui o nível médio* é de 36%. Mas este não é o valor exato do parâmetro que descrevemos, pois não pesquisamos toda a população, mas somente uma amostra. No Capítulo 9, estudaremos uma metodologia capaz de avaliar, de forma aproximada, o valor de determinado parâmetro, considerando apenas os resultados de uma amostra, ou seja, estudaremos o chamado processo de *estimação de parâmetros*.

O ato de generalizar resultados da *parte* (amostra) para o *todo* (população) é conhecido como *inferência estatística*. A estimação de parâmetros é uma forma de inferência estatística. Outra forma surge quando temos alguma hipótese sobre a população em estudo e queremos

verificar a sua validade, com base em uma amostra. São os chamados *testes estatísticos de hipóteses* ou *testes de significância*. Levin (1985, p. 1) descreve:

> O cientista tem ideias sobre a natureza da realidade (ideias que ele denomina hipóteses) e frequentemente testa suas ideias através de pesquisa sistemática.

No problema do Exemplo 1.1, poderíamos ter interesse em testar a seguinte hipótese: *a distribuição do nível de instrução do chefe da casa deve variar conforme a localidade.* Os dados da amostra, como vimos na Figura 1.2, apontam para diferentes distribuições de frequências nas três localidades. Por exemplo, enquanto no Monte Verde temos 57% de famílias com o chefe da casa possuindo o nível médio, na Encosta do Morro, este percentual cai para 16%. Mas estas diferenças nos resultados da amostra são suficientes para afirmarmos que também existem diferenças na população?

Para inferirmos se as diferenças observadas na amostra também existem em toda a população, precisamos saber se elas não poderiam ocorrer meramente pelo *acaso*. O estudo dos testes estatísticos de hipóteses (Capítulo 10) facilitará a solução desse tipo de problema.

Em pesquisas empíricas, é fundamental testar as hipóteses formuladas, pois estas, quando comprovadas estatisticamente, passam a servir de suporte para outras pesquisas, construindo-se, assim, um encadeamento de conhecimentos, levando-nos a novas fronteiras do saber (veja a Figura 1.3).

Figura 1.3 O processo iterativo da evolução do conhecimento.

Parte I

O planejamento da coleta dos dados

Como planejar adequadamente a coleta dos dados

Como alguns conceitos básicos da Estatística podem auxiliar no planejamento da pesquisa

Capítulo 2

Pesquisas e dados[1]

Em nossas decisões do dia a dia, estamos direta ou indiretamente nos baseando em dados. Ao decidir, por exemplo, pela compra de determinado bem, procuramos verificar se ele satisfaz as nossas necessidades, se o seu preço é compatível com nosso orçamento, além de outras características. Posteriormente, comparamos os dados desse bem com eventuais alternativas e, através de uma análise processada internamente em nossa mente, tomamos a decisão de comprá-lo ou não.

Nas pesquisas científicas, também precisamos coletar dados que possam fornecer informações capazes de responder às nossas indagações. Mas para que os resultados da pesquisa sejam confiáveis, tanto a coleta dos dados quanto a sua análise devem ser feitas de forma criteriosa e objetiva. A Figura 2.1 ilustra as principais etapas de uma pesquisa que envolve levantamento e análise de dados.

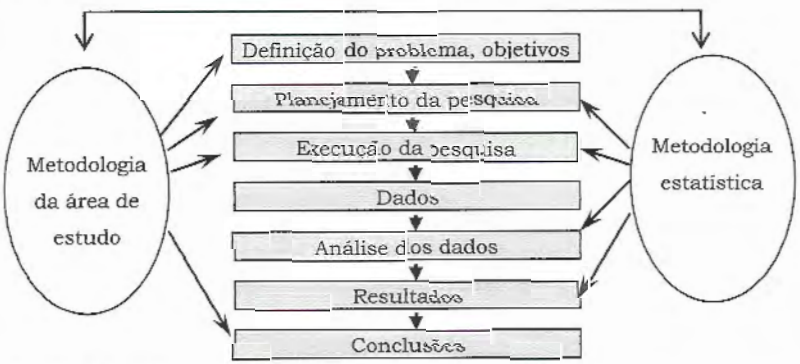


Figura 2.1 Etapas usuais de uma pesquisa quantitativa.

[1] Este capítulo teve a participação da professora Sílvia Modesto Nassar, doutora em Engenharia Biomédica e Professora Titular do Departamento de Informática e Estatística da UFSC.

Embora a aplicação de técnicas estatísticas seja feita basicamente na etapa de análise dos dados, a metodologia estatística deve ser aplicada nas diversas etapas da pesquisa, interagindo com a metodologia da área em estudo. Não é possível obter boas informações de dados que foram coletados de forma inadequada. A qualidade da informação depende da qualidade dos dados! Do mesmo modo, para que a utilização dos resultados estatísticos seja feita de forma correta, torna-se necessário que o pesquisador conheça os princípios básicos das técnicas usadas.

Neste capítulo faremos uma breve explanação sobre as linhas gerais do planejamento de uma pesquisa, dando ênfase ao planejamento da coleta de dados.

2.1 O planejamento de uma pesquisa

O problema de pesquisa

Para se iniciar qualquer processo de pesquisa, deve-se ter bem definido o problema a ser pesquisado. Isto normalmente envolve uma boa revisão da literatura sobre o tema em questão.

Formulação dos objetivos

Os objetivos de uma pesquisa devem ser elaborados de forma bastante clara, já que as demais etapas da pesquisa tomam como base esses objetivos.

Exemplo 2.1 *Objetivo geral*: conhecer o perfil de trabalho dos funcionários de determinada empresa para orientar políticas de recursos humanos.

Para podermos dar sequência a esta pesquisa, precisamos especificar melhor o que queremos conhecer da população de funcionários, ou seja, os *objetivos específicos*. Alguns destes objetivos específicos poderiam ser:

a) Conhecer o tempo médio de serviço dos funcionários na empresa.
b) Conhecer a distribuição do nível de instrução dos funcionários.
c) Verificar o interesse dos funcionários em participar de programas de treinamento.
d) Avaliar o nível de satisfação dos funcionários com o trabalho que exercem na empresa.
e) Verificar se existe associação entre o nível de satisfação do funcionário com a sua produtividade.

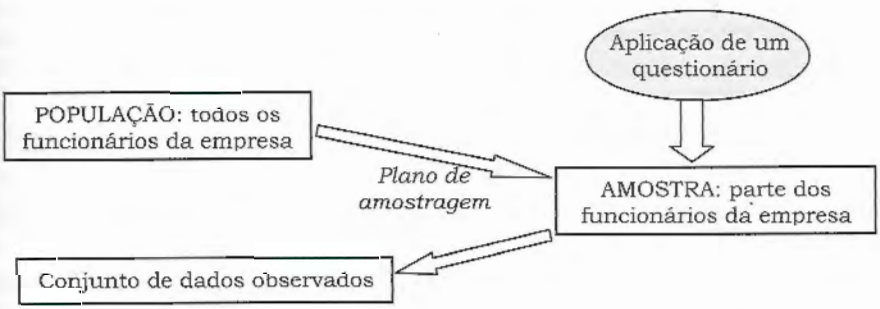
Os objetivos de (a) a (d) podem ser alcançados por uma pesquisa que descreva as características pertinentes da população. Por outro lado, o objetivo (e) é mais analítico, pois nele está embutida a hipótese de que exista associação entre satisfação e produtividade, hipótese que deverá ser colocada à prova no decorrer da pesquisa.

A elaboração dos objetivos específicos deve ser feita de tal forma que forneça uma primeira indicação das características que precisamos observar ou medir nos indivíduos a serem pesquisados. Por exemplo, para atingir aos objetivos do problema em questão, precisamos levantar as seguintes características de cada funcionário da empresa: *tempo de serviço, nível de instrução, interesse em participar de programas de treinamento, nível de satisfação com o trabalho* e *produtividade*.

Tipos de pesquisa

Depois de os objetivos estarem explicitamente traçados, devemos decidir sobre as linhas básicas da condução da pesquisa, ou seja, o delineamento da pesquisa. O Exemplo 2.1 mostra uma pesquisa de levantamento ou *survey* e o Exemplo 2.2 uma pesquisa experimental.

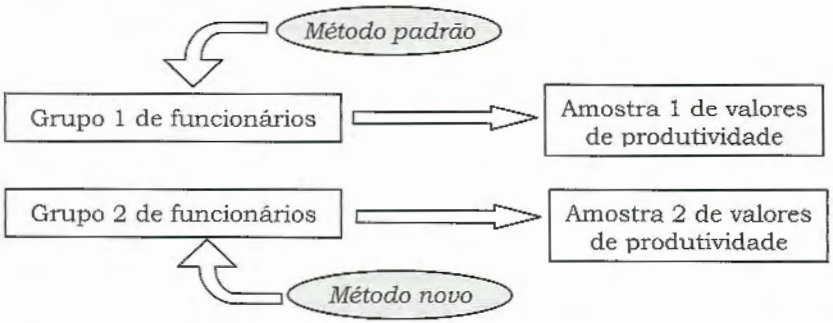
Exemplo 2.1 (continuação) *Delineamento da pesquisa*: um levantamento de dados a partir da aplicação de um questionário em uma amostra de funcionários. *Dados*: resultados de diversos atributos e medidas relativas ao sistema de trabalho dos funcionários respondentes. Esquematicamente:



Na **pesquisa de levantamento** ou ***survey*** observam-se diversas características dos elementos de uma certa população ou amostra, utilizando-se questionários ou entrevistas. A observação é feita naturalmente e sem interferência do pesquisador.

A pesquisa de levantamento é bastante comum nas Ciências Sociais e costuma gerar grandes conjuntos de dados. Na sequência deste livro, daremos mais destaque a esse tipo de pesquisa.

EXEMPLO 2.2 *Objetivo geral*: comparação de dois métodos de treinamento de funcionários, sendo um deles usualmente aplicado, e o outro, novo. Especificamente, queremos decidir qual é o método mais adequado, no sentido de aumentar a produtividade dos funcionários de determinada empresa. *Delineamento da pesquisa*: são formados dois grupos de funcionários, sendo cada grupo treinado por um dos métodos. *Dados*: uma medida de produtividade de cada operário, resultando em dois conjuntos (amostras) de valores de produtividade, relativos a cada método de treinamento. Esquematicamente:



O Exemplo 2.2 enfoca um delineamento de *pesquisa experimental.*

Na **pesquisa experimental** o pesquisador exerce controle sobre o tratamento que vai ser aplicado a cada elemento da(s) amostra(s). Há, portanto, interferência do pesquisador.

Esse tipo de pesquisa é usado para resolver problemas bem específicos, geralmente formulados sob forma de *hipóteses de causa e efeito*. No exemplo em questão, tem-se implicitamente a hipótese de que a produtividade de um funcionário é influenciada pelo método de treinamento. Geralmente a quantidade de dados gerada por uma pesquisa experimental é pequena, mas os dados são suficientemente estruturados (devido ao controle do pesquisador) para que se possa decidir, através de uma análise estatística apropriada, se uma hipótese previamente formulada pode ser aceita ou rejeitada.

Há situações em que conhecemos muito pouco sobre o universo a ser estudado. Nesses casos, podemos realizar uma *pesquisa qualitativa*, observando detalhadamente um pequeno número de elementos, sem uma

formulação criteriosa das características a serem levantadas. Na pesquisa qualitativa não se costuma aplicar métodos estatísticos e, por isto, não a abordaremos neste livro.

População e amostra

Um passo importante no delineamento da pesquisa consiste na decisão de *quem* se vai pesquisar.

População-alvo é o conjunto de elementos que queremos abranger em nosso estudo. São os elementos para os quais desejamos que as conclusões oriundas da pesquisa sejam válidas.

No exemplo sobre o perfil de trabalho dos funcionários de uma empresa, a população-alvo pode ser definida como o conjunto de todos os funcionários da empresa, numa determinada época. Contudo, se a coleta de dados for feita no próprio local de trabalho e no período de uma semana, os funcionários que neste período estão de férias ou de licença ficam inacessíveis de serem observados. Assim, as conclusões baseadas nesses dados não valem, necessariamente, para todo o conjunto de funcionários.

População acessível, ou simplesmente **população**, é o conjunto de elementos que queremos abranger em nosso estudo e que são passíveis de serem observados, com respeito às características (variáveis) que pretendemos levantar.

Quando houver diferença razoável entre a população-alvo e a população acessível, pode haver viés ao generalizar os resultados da análise para toda a população-alvo. Assim, é recomendável citar no relatório da pesquisa a limitação de que seus resultados valem especificamente para a população definida como acessível, evitando que os resultados da pesquisa sejam usados de maneira inadequada.

Nem sempre os elementos que definem a população ficam claramente definidos na formulação dos objetivos. Por exemplo, num levantamento sobre as condições socioeconômicas de um bairro, a população pode ser definida como o conjunto de *famílias* residentes no bairro, o conjunto de *indivíduos* moradores do bairro ou, ainda, como o conjunto *indivíduos com mais de dezoito anos* do bairro. A definição da população depende basicamente dos objetivos da pesquisa, das características a serem levantadas e dos recursos disponíveis. Em alguns casos, podemos trabalhar com mais de uma população.

Em grandes populações é interessante a realização de uma **amostragem**, ou seja, a seleção de uma parte da população para ser observada. Para um leigo em estatística, é surpreendente como uma amostra de 3.000 eleitores fornece um perfil bastante preciso sobre a preferência de todo o eleitorado, na véspera de uma eleição presidencial. Mas isto só é verdade se esta amostra for extraída sob um rigoroso plano de amostragem capaz de garantir a sua representatividade.[2]

A coleta de dados

Depois de definirmos os objetivos e a população a ser estudada, precisamos pensar *como* será a coleta de dados. Em muitas situações não precisamos ir até aos elementos da população para obter os dados, porque eles já existem em alguma publicação ou arquivo. É o que chamamos de *dados secundários*. No Exemplo 2.1, os dados sobre o *tempo de serviço* e *nível de instrução dos funcionários* talvez possam ser obtidos no departamento de pessoal da empresa. Outras características, tais como *interesse em participar de programas de treinamento* e *satisfação com o trabalho*, necessitam ser levantadas, observando diretamente cada funcionário. São os *dados primários*.

Nesta fase da pesquisa, devemos verificar exaustivamente o que já existe de dados sobre o assunto em estudo, pois a utilização de dados secundários pode reduzir drasticamente os custos de uma pesquisa.

Quando os dados forem levantados diretamente dos elementos da população, é necessário construir um instrumento para que sua coleta seja feita de forma organizada. Chamaremos este instrumento de **questionário**, cuja elaboração e formas de aplicação discutiremos na Seção 2.3.

Exercícios

1) Seja uma pesquisa eleitoral, a ser realizada a poucos dias de uma eleição municipal, com o objetivo de verificar a intenção de votos para cada candidato à prefeitura. Defina a população-alvo e a população acessível.

2) A pesquisa descrita no Exercício 1 é experimental ou de levantamento? Justifique.

[2] Algumas técnicas de amostragem serão estudadas no Capítulo 3.

2.2 Dados e variáveis

As variáveis surgem quando perguntamos *o que* vamos observar ou medir nos elementos de uma população ou amostra. A observação (ou medida) de uma variável num elemento da população deve gerar *um e apenas um* resultado.

As **variáveis** são as características que podem ser observadas (ou medidas) em cada elemento da população, sob as mesmas condições.

Como definir uma variável na prática?

Na população de funcionários de uma empresa, podemos definir variáveis, tais como: *tempo de serviço*, *estado civil*, etc. Podemos observá-las com perguntas do tipo:

> Há quanto tempo o Sr. (ou Sra.) trabalha nesta empresa? _____.
> Qual é o seu estado civil? _____.

Contudo, essas perguntas não estão identificando bem as variáveis de interesse, pois os funcionários podem interpretá-las de diferentes formas. Na primeira pergunta, podem ocorrer respostas como: *há pouco mais de 12 anos*, *há 7 meses*, *há muito tempo* e assim por diante, não caracterizando propriamente observações da variável *tempo de serviço*, por não estarem sendo observadas de forma homogênea.

Para que as observações do *tempo de serviço* sejam feitas sob as mesmas condições, precisamos estabelecer a sua unidade de medida, por exemplo, *anos completos de trabalho na empresa*. E a pergunta poderia ser:

> Há quanto tempo o Sr. (ou Sra.) trabalha nesta empresa? _____ anos completos.

Quanto à variável *estado civil*, as possíveis respostas são atributos. Para evitar alguma resposta estranha, podemos estabelecer previamente as possíveis alternativas de resposta. E a pergunta poderia ser:

> Qual o seu estado civil? ( ) solteiro ( ) casado ( ) viúvo ( ) desquitado ( ) divorciado

Ao efetuar estas perguntas a um funcionário da empresa, teremos, para cada pergunta, apenas uma resposta. Cada pergunta está, então, associada a uma variável.

Variáveis qualitativas e quantitativas

Quando os possíveis resultados de uma variável são números de uma certa escala, dizemos que esta variável é quantitativa. Quando os possíveis resultados são atributos ou qualidades, a variável é dita qualitativa (veja a Figura 2.2).

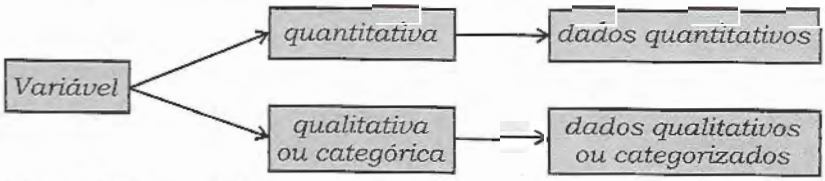


Figura 2.2 Classificação das variáveis e dos dados em termos do nível de mensuração.

No exemplo precedente, o *tempo de serviço (em anos completos)* é uma variável quantitativa, enquanto *estado civil* é qualitativa.

Na descrição das variáveis envolvidas na pesquisa, devemos incluir a escala (ou unidade) em que serão mensuradas as variáveis quantitativas e as categorias (possíveis respostas) das variáveis qualitativas. Sempre que uma característica puder ser adequadamente medida sob forma quantitativa, devemos usar este tipo de mensuração, porque as medidas quantitativas são, em geral, mais informativas do que as qualitativas. Por exemplo, dizer que um funcionário trabalha *há 30 anos* na empresa é mais informativo do que dizer que ele trabalha *há muito tempo* na empresa.

Exemplo de mensuração de uma variável

Muitas características podem ser mensuradas de várias formas e nem sempre fica evidente qual delas é a mais apropriada. Os dois itens abaixo, por exemplo, procuram levantar o nível de satisfação de um funcionário com a política de trabalho na empresa.

(a) Em termos do trabalho que você exerce na empresa, você se sente:
( ) muito satisfeito ( ) pouco satisfeito ( ) insatisfeito

(b) Dê uma nota de 0 (zero) a 10 (dez), relativa ao seu nível de satisfação com o trabalho que você exerce na empresa. Nota:_____.

No primeiro caso, o item do questionário está associado a uma *variável qualitativa*, pois o respondente deve atribuir uma resposta dentre as três categorias apresentadas. Como existe uma ordenação do nível de satisfação nas três opções, dizemos que a variável é **qualitativa ordinal**.

No segundo caso, tenta-se mensurar a característica *satisfação* quantitativamente, pois o respondente vai atribuir um valor, que ele julga ser o seu nível de satisfação, tomando-se como base uma escala de 0 a 10. Cabe observar que, apesar da mensuração quantitativa ser mais informativa, na presente situação ela pode causar algumas distorções, pois, um 7 (sete) para um respondente pode não significar exatamente um 7 (sete) para outro, já que a escala de 0 (zero) a 10 (dez) pode ser entendida de forma diferenciada entre os indivíduos.

A decisão de *como medir* determinada característica depende de vários aspectos, mas é sempre recomendável verificar se a mensuração proposta leva aos objetivos da pesquisa e, além disso, se ela é viável de ser aplicada.

Variáveis e itens de um questionário

Nem sempre há uma relação direta entre um item de um questionário e uma variável. Veja o exemplo a seguir.

> Assinale os esportes que você costuma praticar regularmente:
> ( ) futebol ( ) basquetebol ( ) voleibol
> ( ) outros. Especificar: _____.

Este item não está associado diretamente a uma única variável *esportes*, pois um respondente pode praticar mais de um esporte, violando a suposição básica da variável assumir *um e apenas um* resultado, por respondente. Podemos, por outro lado, associar várias variáveis a este item, tais como: (1) *quantidade de esportes que pratica regularmente*, (2) *futebol (pratica ou não)*, (3) *basquetebol (pratica ou não)*, e assim por diante.[3]

A especificação do esporte na categoria outros pode ser analisada posteriormente, podendo ser incluídas novas variáveis indicadoras do tipo *pratica* ou *não pratica*.

[3] Uma outra possibilidade seria definir a variável e*sportes que pratica*, tendo como possíveis respostas todas as combinações de modalidades de esportes. Mas a análise destas respostas seria difícil, dado o grande número de possíveis alternativas.

EXERCÍCIOS

3) Defina que variáveis precisam ser levantadas para cada um dos objetivos específicos do Exemplo 2.1. Considerando as suas definições, verificar quais são qualitativas e quais são quantitativas.

4) Considerando a população das crianças em creches municipais de Florianópolis, completar as definições das seguintes variáveis e verificar quais são qualitativas e quais são quantitativas.

a) altura b) peso c) idade d) sexo e) cor
f) nacionalidade do pai e g) local do nascimento.

2.3 ELABORAÇÃO DE UM QUESTIONÁRIO

Na condução de uma pesquisa, a construção de um questionário é uma etapa longa que deve ser executada com muita cautela. Tendo em mãos os objetivos da pesquisa claramente definidos, bem como a população a ser estudada, chamamos a atenção de alguns procedimentos para a construção de um questionário.

a) Separar as características (variáveis) a serem levantadas.
Para ilustrar, retomemos o Exemplo 2.1, com os seguintes objetivos específicos:
- conhecer o tempo médio de serviço dos funcionários na empresa;
- conhecer a distribuição do nível de instrução dos funcionários e
- avaliar o nível de satisfação dos funcionários com o trabalho que exercem na empresa.

Temos, então, as seguintes características a serem levantadas dentre os funcionários da empresa: *tempo de serviço*, *nível de instrução* e *nível de satisfação com o trabalho*.

b) Fazer uma revisão bibliográfica para verificar formas de mensurar as variáveis em estudo.

No exemplo precedente precisamos avaliar o nível de satisfação dos funcionários. Podemos procurar referências que nos orientem em como *medir* a satisfação. Em levantamentos de dados socioeconômicos, podemos consultar os modelos de questionários utilizados pelo IBGE, os quais já foram bastante estudados e testados.[4]

[4] IBGE é a sigla do *Instituto Brasileiro de Geografia e Estatística*, órgão responsável por diversos levantamentos no Brasil, como os censos demográficos, censos agropecuários, censos industriais e anuários estatísticos.

c) Estabelecer a forma de mensuração das variáveis a serem levantadas.

Para as variáveis quantitativas, devem estar bem definidas as unidades de medida (meses, metros, kg etc.) que devem acompanhar as respostas. Nas variáveis qualitativas deve haver uma lista completa de alternativas, mesmo que seja necessário incluir categorias como: *outros*, *não tem opinião* etc. Por exemplo, o *tempo de serviço* pode ser observado quantitativamente, *em anos completos de serviço na empresa*; e o *nível de instrução*, em categorias mutuamente exclusivas, como: *nenhum*, *fundamental*, *médio* e *superior*. O *nível de satisfação com o trabalho* pode ser avaliado por uma escala de cinco pontos, sendo 1 – *completamente insatisfeito*, 2 – *insatisfeito*, 3 – *mais ou menos satisfeito*, 4 – *satisfeito* e 5 – *completamente satisfeito*.

d) Elaborar uma ou mais perguntas para cada variável a ser observada.

A variável *nível de satisfação com o trabalho* pode ser avaliada sob vários enfoques, como a satisfação com o salário que recebe, com a segurança no emprego, com a autonomia de trabalho que a empresa oferece, etc. Estes itens podem ser avaliados isoladamente, num mesmo tipo de escala, como a escala de cinco pontos sugerida em (c). E o *nível de satisfação* ser mensurado como a soma das respostas destes itens.

e) Verificar se a pergunta está suficientemente clara.

As perguntas devem ser formuladas numa linguagem que seja compreensível para todos os elementos da população e, além disso, não devem deixar dúvidas de interpretação.

f) Verificar se a forma da pergunta não está induzindo alguma resposta.

Não se deve, por exemplo, ao tentar avaliar a satisfação de um funcionário com o trabalho que exerce, citar aspectos positivos ou negativos do trabalho. Isto pode induzir a resposta.

g) Verificar se a resposta da pergunta não é óbvia.

Dependendo da forma como se pergunta sobre a *satisfação com o valor do salário recebido*, a resposta será sempre *não*, independentemente da real satisfação que o funcionário tenha com respeito a esse item. Isto deve ocorrer, por exemplo, quando só existem dois níveis de respostas: *sim* e *não*. Usando uma escala de cinco pontos, como sugerida anteriormente, podemos detectar melhor algumas diferenças entre os respondentes.

Um aspecto fundamental nesta fase da pesquisa é o planejamento de como usar as respostas dos diversos itens para responder às indagações de nossa pesquisa. O questionário também deve ser feito de forma a

facilitar a análise dos dados. O questionário deve ser completo, no sentido de abranger as características necessárias para atingir os objetivos da pesquisa; ao mesmo tempo, não deve conter perguntas que fujam desses objetivos, pois, *quanto mais longo o questionário, menor tende a ser a confiabilidade das respostas.*

Formas de aplicação de um instrumento de pesquisa

Nesta fase, também devemos decidir sobre a forma de aplicação de nosso questionário, ou, mais genericamente, do *instrumento de pesquisa.*

Um *questionário* propriamente dito é respondido pelo próprio elemento da população, sem que algum encarregado da pesquisa observe o respondente no momento do preenchimento. Numa *entrevista estruturada,* o entrevistado responde verbalmente as perguntas e o entrevistador as transcreve para uma ficha. Nesta segunda situação, o entrevistador pode ou não interferir, sob forma de esclarecimento de algum item, anotando aspectos que julgar relevantes, mas nunca influenciando na resposta do entrevistado.

Em pesquisas que envolvem aspectos íntimos dos respondentes, deve-se dar preferência a um questionário anônimo, com o cuidado de que o respondente preencha o questionário individualmente e à vontade. Por outro lado, numa pesquisa a ser realizada numa população que tenha pessoas não alfabetizadas, uma entrevista estruturada é mais adequada.

Deve sempre haver homogeneidade na forma de aplicação dos questionários. Em pesquisas que envolvem vários entrevistadores, torna-se necessário um prévio treinamento para garantir a homogeneidade na aplicação.

Pré-testagem

Antes de iniciar a coleta de dados através de um questionário, precisamos verificar se o instrumento está bom. Nesse contexto, torna-se fundamental a realização de um *pré-teste,* aplicando o questionário em alguns indivíduos com características similares aos indivíduos da população em estudo. Somente pela aplicação efetiva do questionário é que podemos detectar algumas falhas que tenham passado despercebidas em sua elaboração, tais como: ambiguidade de alguma pergunta, resposta que não havia sido prevista, não variabilidade de respostas em alguma pergunta, etc. O pré-teste também pode ser usado para estimar o tempo de aplicação do questionário.

Exercícios

5) Elaborar um esboço de questionário para o problema descrito no Exemplo 2.1.

6) Ao longo deste capítulo escrevemos: *quanto mais longo for o questionário menor deve ser a confiabilidade das respostas.* Explique por que isto geralmente ocorre.

7) Com respeito ao Exercício 1, sobre uma pesquisa eleitoral, complemente com alguns objetivos específicos e proponha um questionário para a obtenção dos dados. Discuta sobre a forma de aplicação que você julga ser a mais adequada para a presente situação.

2.4 Uma aplicação

Nesta seção apresentaremos um exemplo de um projeto de pesquisa relativamente simples, desenvolvido com a participação dos alunos da disciplina de Estatística do curso de Ciências Sociais da UFSC, semestre 1991/1, com finalidades puramente acadêmicas.

O problema de pesquisa: A relação do aluno universitário com o curso.

Objetivo geral: Conhecer melhor a relação entre o aluno e o seu curso (curso de Ciências da Computação da UFSC), para servir de subsídio nas políticas de melhoria do curso.

Objetivos específicos:

1) Avaliar o nível de satisfação do aluno com o curso que está realizando.
2) Verificar se existe associação entre o nível de satisfação do aluno com o seu desempenho no curso.
3) Levantar os aspectos positivos e negativos do curso, na visão do aluno.

População: Estudantes que estavam cursando as três últimas fases do curso de Ciências da Computação da UFSC, semestre 1991/1.

Amostra: Alunos presentes no dia de aplicação dos questionários, realizada em salas de aula de três disciplinas obrigatórias das últimas fases do curso.[5]

[5] Como veremos no próximo capítulo, essa forma de seleção da amostra pode causar viés, pois os alunos que costumam faltar às aulas ficam quase que inacessíveis. E alguns desses alunos podem estar faltando sistematicamente por estarem insatisfeitos com o curso.

Forma de mensuração das variáveis

Satisfação com o curso: avaliação numérica, numa escala de 1 (um) a 5 (cinco), de acordo com a percepção do aluno. Além de uma medida de satisfação geral, complementa-se com avaliações de aspectos específicos do curso, como corpo docente, recursos materiais e conteúdo curricular.

Desempenho do aluno: Índice de Aproveitamento Acumulado, calculado pela instituição, em função dos conceitos (ou notas) obtidos pelo aluno nas disciplinas cursadas.

Aspectos positivos e *negativos do curso*: 1) avaliações numéricas, numa escala de 1 (um) a 5 (cinco), de acordo com o nível que o aluno julgar que melhor se adapte à sua concordância com alguns aspectos do curso; 2) avaliações qualitativas, em que o aluno descreve livremente o principal aspecto positivo e negativo do curso. Na segunda avaliação, as categorias de cada variável serão criadas depois de uma análise das respostas dos questionários, onde as respostas similares serão agrupadas numa única categoria.

Questionário

Este questionário faz parte de um trabalho acadêmico. Os questionários são anônimos, portanto não coloque seu nome. Solicitamos sua colaboração respondendo correta e francamente os diversos itens, agradecendo-lhe antecipadamente. Os resultados da pesquisa ficarão disponíveis para a comunidade acadêmica.

1) Qual é o curso que você está realizando na UFSC? _____.
2) Qual é a fase predominante em que você se encontra? _____.
3) Dê uma nota de 1 (um) a 5 (cinco), sendo 1 o nível mínimo e 5 o nível máximo, para as seguintes características relacionadas com você e seu curso.
   a) Didática dos professores de seu curso ..... (1 2 3 4 5)
   b) Nível de conhecimento dos professores ..... (1 2 3 4 5)
   c) Bibliografia disponível ..... (1 2 3 4 5)
   d) Laboratórios e outros recursos materiais ..... (1 2 3 4 5)
   e) Conteúdo dos programas das disciplinas oferecidas ..... (1 2 3 4 5)
   f) Encadeamento das disciplinas ..... (1 2 3 4 5)
   g) Satisfação com o curso, num sentido geral ..... (1 2 3 4 5)
4) Apresente o principal ponto positivo e negativo de seu curso.
   POSITIVO: _____.
   NEGATIVO: _____.
5) Anote o seu Índice de Aproveitamento Acumulado _____ (ver tabela com o aplicador).

Comentários sobre os itens do questionário

Os itens 1 e 2 são de controle, para verificar se o respondente realmente pertence à população em estudo. Estes itens não serão usados na análise dos dados.

No item 3 estamos tentando quantificar algumas características do curso, na percepção do aluno, numa escala de 1 (um) a 5 (cinco). Este item está associado com os três objetivos da pesquisa. Os subitens de (a) a (f) procuram atingir o objetivo 3, já que as respostas do subitem (g) serão usadas com vistas aos objetivos 1 e 2.

O item 4 procura complementar a informação do item 3, através de uma *pergunta aberta*.

O item 5 é uma medida de desempenho do aluno no curso, calculada pela instituição e usada para estabelecer prioridades na matrícula. Como, em geral, os alunos não sabem de cor o seu índice, o aplicador do questionário levou uma relação contendo os índices de aproveitamento de toda a turma, para que o aluno pudesse localizar o seu, transcrevendo-o na folha do questionário. As respostas deste item, juntamente com o item 3(g), serão usadas para atingir o objetivo 2.[6]

2.5 Codificação dos dados

Depois de os dados terem sido coletados, precisamos organizá-los, para facilitar a realização da análise. Tomemos o primeiro questionário respondido.

Respostas de um questionário

1) Qual o curso que você está realizando na UFSC? *Computação.*
2) Qual a fase predominante em que você se encontra? *Oitava.*
3) Dê uma nota de 1 (um) a 5 (cinco), sendo 1 o nível mínimo e 5 o nível máximo, para as seguintes características relacionadas com você e seu curso.
   a) Didática dos professores de seu curso ..... (1 **X** 3 4 5)
   b) Nível de conhecimento dos professores ..... (1 2 3 **X** 5)
   c) Bibliografia disponível ..... (1 **X** 3 4 5)

[6] A inclusão deste dado no próprio questionário era importante para podermos associá-lo com outras respostas do aluno. Como o questionário era anônimo, não seria possível incluí-lo depois da coleta dos dados.

d) Laboratórios e outros recursos materiais ..... (X 2 3 4 5)
e) Conteúdo dos programas das disciplinas oferecidas ..... (1 X 3 4 5)
f) Encadeamento das disciplinas ..... (1 X 3 4 5)
g) Satisfação com o curso, num sentido geral ..... (1 X 3 4 5)

4) Apresente o principal ponto positivo e negativo de seu curso.
POSITIVO: *Professores razoáveis.*
NEGATIVO: *Falta e má conservação de laboratórios.*

5) Anote o seu Índice de Aproveitamento Acumulado? *1,95* (ver tabela com o aplicador).

Os dados normalmente são armazenados numa matriz (ou quadro), onde cada coluna se refere a uma variável e cada linha a um respondente.[7] A Tabela 2.1 mostra os dados armazenados dos cinco primeiros respondentes. Os dados do questionário respondido acima estão na primeira linha da tabela.

Tabela 2.1 Armazenamento dos dados de cinco respondentes

nº do quest.	Item do questionário									
	3(a) didat.	3(b) conhec.	3(c) bibl.	3(d) labor.	3(e) disc.	3(f) curric.	3(g) satisf.	4(a) posit.	4(b) negat.	5 desemp.
1	2	4	2	1	2	2	2	1	2	1,95
2	2	3	2	1	2	3	3	9	1	1,72
3	3	2	1	1	3	2	3	3	3	2,39
4	2	2	3	1	4	4	3	3	5	2,57
5	3	3	4	3	3	4	2	3	1	2,51

As categorias relativas aos itens 4(a) e 4(b) foram criadas a partir de uma análise das respostas dos questionários, agrupando respostas similares. Para o item 4(a), **ponto positivo**, as categorias e correspondentes códigos foram: 1 – *Professores*, 2 – *Atualização*, 3 – *Abrangência*, 4 – *Aplicações práticas*, 5 – *Currículo e disciplinas* e 9 – *Outros*. Para o item 4(b), **ponto negativo**, foram: 1 – *Professores*, 2 – *Laboratórios e recursos materiais*, 3 – *Currículo e disciplinas*, 4 – *Aplicações*, 5 – *Atualização* e 9 – *Outros*.

No Anexo, final deste capítulo, apresentamos os dados dos 60 respondentes desta pesquisa. A análise desses dados será feita ao longo dos exercícios dos próximos capítulos.

[7] Em linguagem computacional, a matriz de dados corresponde a um *arquivo*, as variáveis são os *campos* e os dados de um respondente são os *registros* do arquivo.

Anexo

Dados da pesquisa descrita na Seção 2.4. Respostas de 60 questionários.

nº do quest.	3(a) didat.	3(b) conhec.	3(c) bibl.	3(d) labor.	3(e) disc.	3(f) curric.	3(g) satisf.	4(a) posit.	4(b) negat.	5 desemp.
1	2	4	2	1	2	2	2	1	2	1,95
2	2	3	2	1	2	3	3	9	1	1,72
3	3	2	1	1	3	2	3	3	3	2,39
4	2	2	3	1	4	4	3	3	5	2,57
5	3	3	4	3	3	4	2	3	1	2,51
6	2	2	2	1	3	1	3	9	2	2,04
7	4	3	1	1	4	2	5	1	9	1,99
8	2	3	2	2	2	3	3	.	1	2,69
9	3	3	2	3	4	4	4	5	2	2,57
10	3	4	2	1	3	4	4	1	1	2,10
11	3	3	2	2	3	3	3	2	2	3,61
12	4	4	2	3	4	3	4	1	2	2,37
13	2	3	3	4	4	3	4	3	1	1,62
14	2	2	3	2	3	3	3	1	2	1,87
15	2	3	3	2	4	3	3	.	.	2,47
16	3	3	1	2	3	4	3	2	1	2,61
17	2	4	3	4	4	2	3	3	1	2,73
18	4	4	1	1	4	4	5	9	2	2,50
19	3	4	2	1	4	3	3	1	4	3,12
20	2	2	1	1	3	3	3	9	1	3,19
21	2	3	2	1	3	4	3	2	2	3,65
22	3	4	4	3	4	4	5	1	2	3,01
23	2	3	2	3	4	3	3	1	1	2,13
24	3	4	4	4	4	3	3	9	9	1,25
25	3	4	2	3	4	5	4	1	9	2,34
26	3	3	2	2	3	4	3	2	5	2,69
27	3	4	2	3	3	3	4	9	3	2,59
28	3	3	2	4	3	4	2	9	1	2,27
29	2	2	1	3	2	1	2	1	3	1,30
30	3	3	1	3	4	4	4	9	1	3,18
31	3	4	2	3	3	4	4	3	1	2,54
32	2	3	1	1	3	3	3	2	5	2,07
33	3	3	2	1	4	2	4	1	1	2,26
34	2	4	4	3	4	5	4	9	1	2,02
35	3	2	2	4	3	2	3	.	4	2,19
36	3	4	2	2	3	4	4	4	2	3,48
37	3	3	3	4	3	4	2	4	1	3,29
38	3	3	3	4	3	3	3	.	1	2,94
39	2	3	1	3	3	4	3	9	1	2,92
40	4	4	1	3	4	4	3	.	1	2,10
41	3	3	3	3	4	2	3	3	4	2,37
42	2	3	2	3	3	3	3	.	1	2,43
43	3	4	2	2	3	4	4	4	3	2,00

nº do quest.	3(a) didat.	3(b) conhec.	3(c) bibl.	3(d) labor.	3(e) disc.	3(f) curric.	3(g) satisf.	4(a) posit.	4(b) negat.	5 desemp.
44	2	2	2	1	3	3	3	4	1	1,83
45	3	3	2	3	4	5	4	9	1	2,93
46	2	3	1	2	4	3	3	9	2	2,50
47	3	4	3	3	4	4	5	2	1	3,00
48	3	3	3	4	3	4	3	9	1	2,06
49	3	3	2	1	3	3	3	9	1	1,56
50	3	4	2	1	3	3	3	.	2	2,27
51	3	3	1	1	2	3	3	.	2	2,14
52	4	4	2	2	4	3	4	9	9	2,42
53	3	4	1	2	3	3	4	1	2	3,56
54	3	3	3	2	5	4	3	5	2	3,52
55	3	4	3	2	4	4	4	.	.	3,22
56	4	3	5	3	4	4	4	5	1	3,63
57	3	4	3	2	3	4	3	1	2	3,53
58	2	3	3	3	4	4	2	5	1	2,13
59	3	4	3	3	5	5	3	5	1	2,31
60	3	3	1	1	3	3	3	.	.	3,62

NOTA: O ponto (.) representa não resposta.

Capítulo 3

Técnicas de amostragem[1]

A amostragem é naturalmente usada em nossa vida diária. Por exemplo, para verificar o tempero de um alimento em preparação, podemos provar (observar) uma pequena porção. Estamos fazendo uma *amostragem*, ou seja, extraindo do todo (*população*) uma parte (*amostra*), com o propósito de termos uma ideia (*inferirmos*) sobre a qualidade do tempero de todo o alimento.

Nas pesquisas científicas, em que se deseja conhecer algumas características (*parâmetros*) de uma população, também podemos observar apenas uma amostra de seus elementos e, com base nos resultados da amostra, obter valores aproximados, ou *estimativas*, para os parâmetros de interesse. Esse tipo de pesquisa é usualmente chamado de *levantamento por amostragem*. Contudo, a seleção dos elementos que serão efetivamente observados deve ser feita sob uma metodologia adequada, de tal forma que os resultados da amostra sejam suficientemente informativos para se inferir sobre os parâmetros populacionais. E o objetivo do presente capítulo é estudar esta metodologia, ou seja, o *processo de amostragem*.

Alguns conceitos e exemplos

Como definimos no capítulo anterior,

> **População** é o conjunto de elementos para os quais desejamos que as conclusões da pesquisa sejam válidas, com a restrição de que esses elementos possam ser observados ou mensurados sob as mesmas condições.

[1] Este capítulo teve a participação da professora Sílvia Modesto Nassar, doutora em Engenharia Biomédica e Professora Titular do Departamento de Informática e Estatística da UFSC.

A população pode ser formada por pessoas, famílias, estabelecimentos industriais, ou qualquer outro tipo de elementos, dependendo basicamente dos objetivos da pesquisa. Mas, em geral, o interesse se resume em alguns *parâmetros*.

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

EXEMPLO 3.1 Numa pesquisa epidemiológica, a população pode ser definida como todas as pessoas da região em estudo, no momento da pesquisa. O principal parâmetro a ser avaliado deve ser *a percentagem de pessoas contaminadas*.

EXEMPLO 3.2 Numa pesquisa eleitoral, a três dias de uma eleição municipal, a população são os eleitores que vão votar no município (*população-alvo*), mas, para viabilizar a pesquisa, é comum definir a população como o conjunto dos eleitores que residem no município. Os principais parâmetros são *as percentagens de votos de cada candidato*, no momento da pesquisa.

EXEMPLO 3.3 Para planejar políticas de recursos humanos numa empresa, com milhares de funcionários, pode ser realizada uma pesquisa para avaliar alguns parâmetros da população de funcionários, tais como: *tempo médio de serviço, percentagem de funcionários com nível de instrução superior, percentagem de funcionários com interesse num certo programa de treinamento*, etc.

Nos três exemplos, o leitor pode perceber a dificuldade em pesquisar toda a população. São situações em que se recomenda usar amostragem. Veja a Figura 3.1.

Figura 3.1 Pesquisa eleitoral: um caso típico de levantamento por amostragem.

O termo **inferência estatística** refere-se ao uso apropriado dos dados de uma amostra para se ter conhecimento sobre parâmetros da população de onde foi extraída a amostra. Os valores calculados, com base na amostra e com o objetivo de avaliar parâmetros desconhecidos, são chamados *estimativas* desses parâmetros. Numa pesquisa eleitoral, por exemplo, as percentagens dos candidatos, divulgadas antes da eleição, são *estimativas* das verdadeiras percentagens, relativas a toda a população de eleitores.

Amostra: parte dos elementos de uma população.

Amostragem: o processo de seleção da amostra.

Estimativa: valor calculado com base na amostra e usado com a finalidade de avaliar aproximadamente um parâmetro.

Exemplo 3.3 (continuação) Se uma amostra de 200 funcionários da empresa acusar 60% de favoráveis a um certo programa de treinamento, podemos dizer que o valor 60% é uma *estimativa* da percentagem de funcionários da empresa favoráveis a esse programa de treinamento.

Por que amostragem?

1) *Economia.* Em geral, torna-se bem mais econômico o levantamento de somente uma parte da população.
2) *Tempo.* Numa pesquisa eleitoral, a três dias de uma eleição presidencial, não haveria tempo suficiente para pesquisar toda a população de eleitores do país, mesmo que houvesse recursos financeiros em abundância.
3) *Confiabilidade dos dados.* Quando se pesquisa um número reduzido de elementos, pode-se dar mais atenção aos casos individuais, evitando erros nas respostas.
4) *Operacionalidade.* É mais fácil realizar operações de pequena escala. Um dos problemas típicos nos grandes censos (pesquisas de toda a população) é o controle dos entrevistadores.

Quando o uso de amostragem não é interessante?

1) *População pequena.* Imagine que se queira saber a percentagem de mulheres numa sala de aula com dez alunos, antes de conhecer a turma. É intuitiva a necessidade de observar quase todos os estudantes da sala para se ter uma estimativa razoável. Em especial, quando a amostragem é obtida sorteando elementos da população (*amostragem aleatória*), mais vale o tamanho absoluto da amostra do que a percentagem que ela representa na população.

2) *Característica de fácil mensuração.* Talvez a população não seja tão pequena, mas a variável que se quer observar é de tão fácil mensuração que não compensa investir num plano de amostragem. Por exemplo, para verificar a percentagem de funcionários favoráveis à mudança no horário de um turno de trabalho, podemos entrevistar toda a população no próprio local de trabalho. Esta atitude pode também ser politicamente mais recomendável.

3) *Necessidade de alta precisão.* A cada dez anos o IBGE realiza um censo demográfico *para* estudar diversas características da população brasileira. Dentre essas características, tem-se o parâmetro *número de habitantes residentes no país.* É um parâmetro que precisa ser avaliado com grande precisão; por isso, pesquisa-se toda a população.

Plano de amostragem

Para elaborar um plano de amostragem, devemos ter bem definidos os objetivos da pesquisa, a população a ser amostrada, bem como os parâmetros que precisamos estimar para atingir aos objetivos da pesquisa. Num plano de amostragem deve constar a definição da unidade de amostragem, a forma de seleção dos elementos da população e o tamanho da amostra.

A **unidade de amostragem** é a unidade a ser selecionada para se chegar aos elementos da população. As unidades de amostragem podem ser os próprios elementos da população, ou outras unidades que sejam mais fáceis de serem selecionadas, mas que tenham correspondência com os elementos da população. Por exemplo, numa população de famílias moradoras de uma certa cidade, podemos planejar a seleção de domicílios residenciais da cidade. Chegando ao domicílio (*unidade de amostragem*), podemos chegar à família moradora deste domicílio (*elemento da população*).

A seleção dos elementos que farão parte da amostra pode ser feita sob alguma forma de *sorteio.* São as chamadas **amostragens aleatórias**, que são particularmente interessantes por permitirem a utilização das técnicas clássicas de inferência estatística, facilitando a análise dos dados e fornecendo maior segurança ao generalizar resultados da amostra para a população. Estudaremos, inicialmente, alguns tipos de amostragem, em especial as aleatórias. Posteriormente, discutiremos a questão do tamanho da amostra.

3.1 Amostragem aleatória simples

Para selecionar uma amostra aleatória simples, precisamos ter uma *lista* completa dos elementos da população (ou de unidades de amostragem apropriadas). Este tipo de amostragem consiste em selecionar a amostra através de um sorteio, sem restrição.

Seja uma população com N elementos. Uma forma de extrair uma amostra aleatória simples de tamanho n, sendo $n < N$, é identificar os elementos da população em pequenos pedaços de papel e retirar, ao acaso, n pedaços. Consideraremos, neste livro, que o sorteio seja feito *sem* reposição, ou seja, cada elemento da população não pode ser sorteado mais que uma vez.

A amostragem aleatória simples tem a seguinte propriedade: *qualquer subconjunto da população, com o mesmo número de elementos, tem a mesma probabilidade de fazer parte da amostra.* Em particular, temos que *cada elemento da população tem a mesma probabilidade* (dada por n/N) *de pertencer à amostra.*

Números aleatórios

As tabelas de números aleatórios facilitam o processo de seleção de uma amostra aleatória. São formadas por números resultantes de sucessivos sorteios independentes de {0, 1, 2,..., 9}. A seguir, são apresentados alguns números aleatórios (as duas primeiras linhas da Tabela 1 do apêndice). Os espaços colocados a cada dois algarismos servem, apenas, para facilitar a visualização da tabela, não interferindo na sua utilização.

Números aleatórios

59 58 48 36 47	92 85 05 08 65	47 49 10 41 05	10 75 59 75 99	17 28 97 99 75
53 26 21 50 21	37 93 85 52 86	86 22 75 34 37	69 85 25 03 78	50 26 18 25 10

Exemplo 3.4 Com o objetivo de estudar algumas características dos funcionários de uma certa empresa, vamos extrair uma amostra aleatória simples de tamanho cinco. A listagem dos funcionários da empresa é apresentada a seguir.[2]

[2] Para facilitar a exemplificação das técnicas de amostragem, usaremos populações pequenas. Contudo, como já discutimos, não se costuma usar amostragem aleatória em população muito pequena.

POPULAÇÃO: funcionários da empresa

Aristóteles	Anastácia	Arnaldo	Bartolomeu	Bernardino
Cardoso	Carlito	Cláudio	Ermílio	Hercílio
Ernestino	Endevaldo	Francisco	Felício	Fabrício
Geraldo	Gabriel	Getúlio	Hiraldo	João da Silva
Joana	Joaquim	Joaquina	José da Silva	José de Souza
Josefa	Josefina	Maria José	Maria Cristina	Mauro
Paula	Paulo César			

Para utilizar uma tabela de números aleatórios, precisamos associar cada elemento da população a um número. Por simplicidade, consideraremos números inteiros sucessivos, com a mesma quantidade de algarismos, iniciando-se por 1 (um).

Numeração dos elementos da população

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

Para extrairmos uma amostra aleatória simples de tamanho $n = 5$, basta tomar cinco números aleatórios do conjunto {01, 02,..., 32}. Os funcionários associados aos números selecionados formarão a amostra. Não existe forma específica para extrair os números da tabela. Iniciaremos, neste exemplo, pela primeira linha, desprezando os valores que estiverem fora do conjunto {01, 02,..., 32} e os valores que se repetirem.

Números aleatórios extraídos da tabela: 05, 08, 10, 17 e 28.
Amostra: {Bernardino, Cláudio, Hercílio, Gabriel e Maria José}

Na prática, estamos interessados na observação de certas variáveis associadas aos elementos da amostra. No exemplo em questão, poderíamos estar interessados na variável *tempo de serviço na empresa, em anos completos*. Denominaremos esta variável de X. Para cada funcionário da amostra, temos um valor para a variável X. O conjunto desses valores é chamado *amostra aleatória simples da variável X*, conforme ilustrado a seguir:

Amostra de funcionários:

{Bernardino, Claúdio, Hercílio, Gabriel e Maria José}

Amostra da variável X: $\{X_1, X_2, X_3, X_4, X_5\}$,

onde X_1 é o tempo de serviço do Bernardino, X_2 é o tempo de serviço do Cláudio, etc.

Exercícios

1) Considerando a população do Exemplo 3.4, extraia uma amostra aleatória simples de $n = 10$ funcionários. Inicie pela segunda linha da tabela de números aleatórios (Tabela 1 do apêndice).

2) Ainda com respeito ao Exemplo 3.4, suponha que o tempo de serviço destes funcionários, em anos completos, são os valores seguintes:

Aristóteles	2	Anastácia	5	Arnaldo	2	Bartolomeu	1	Bernardino	11
Cardoso	16	Carlito	3	Cláudio	1	Ermílio	13	Hercílio	10
Ernestino	7	Endevaldo	2	Francisco	0	Felício	10	Fabrício	5
Geraldo	8	Gabriel	8	Getúlio	2	Hiraldo	9	João da Silva	4
Joana	2	Joaquim	22	Joaquina	3	José da Silva	4	José de Souza	2
Josefa	1	Josefina	5	Maria José	3	Maria Cristina	3	Mauro	11
Paula	4	Paulo César	2						

Apresente a amostra da variável *tempo de serviço* associada à amostra de funcionários obtida no Exercício 1.

3) Usando a primeira coluna da tabela de números aleatórios, extraia uma amostra aleatória simples de 4 (quatro) letras do alfabeto da língua portuguesa.

4) Os elementos de uma certa população estão dispostos numa lista, cuja numeração vai de 1.650 a 8.840. Descreva como você usaria uma tabela de números aleatórios para obter uma amostra de 100 elementos. Seria necessário efetuar nova numeração?

5) Seja um conjunto de 20 crianças numeradas de 1 a 20. Usando uma tabela de números aleatórios, divida aleatoriamente essas crianças em dois grupos de 10 crianças.

3.2 Outros tipos de amostragens aleatórias

Amostragem sistemática

Muitas vezes, é possível obter uma amostra de características parecidas com a aleatória simples, por um processo bem mais rápido do que o apresentado na seção anterior. Por exemplo, para tirar uma amostra de 1.000 fichas, dentre uma população de 5.000 fichas, podemos tirar, sistematicamente, uma ficha a cada cinco. Para garantir que cada ficha da população tenha a mesma probabilidade de pertencer à amostra, devemos sortear a primeira ficha dentre as cinco primeiras.

Uma amostra sistemática poderá ser tratada como uma amostra aleatória simples se os elementos da população estiverem ordenados aleatoriamente. A relação N/n é chamada **intervalo de seleção**. No exemplo das fichas, o intervalo de seleção é $5.000/1.000 = 5$.

Exemplo 3.5 Usaremos, como exemplo, a população dos $N = 32$ funcionários do Exemplo 3.4. Vamos realizar uma amostragem sistemática para obtermos uma amostra de tamanho $n = 5$. Calculemos, inicialmente, o intervalo de seleção: $N/n = 32/5 \approx 6$.

População: funcionários da empresa

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

Devemos sortear um elemento dentre os seis primeiros, podendo, para isso, tomar um número da tabela de números aleatórios. Tomando, por exemplo, o primeiro número de um algarismo da segunda linha (53 26...), temos que o primeiro funcionário da amostra é o quinto elemento, portanto o Bernardino. E a amostra sistemática:[3]

5 → Bernardino
5 + 6 = 11 → Ernestino
11 + 6 = 17 → Gabriel
17 + 6 = 23 → Joaquina
23 + 6 = 29 → Maria Cristina

Amostragem estratificada

A técnica da amostragem estratificada consiste em dividir a população em subgrupos, que denominaremos **estratos**. Os estratos devem ser internamente mais homogêneos do que a população toda, com respeito às principais variáveis em estudo. Por exemplo, para estudar o interesse dos funcionários, de uma grande empresa, em realizar um programa de treinamento, podemos estratificar a população por nível de instrução, pelo nível hierárquico ou por setor de trabalho. Devemos escolher um critério de estratificação que forneça estratos bem homogêneos, com respeito ao que se está estudando. Assim, é fundamental um prévio conhecimento sobre a população em estudo.

[3] Devido ao arredondamento no cálculo do intervalo de seleção, o número n de elementos da amostra pode ficar diferente do número planejado. Se o intervalo de seleção for grande (digamos, maior que 10) a diferença será desprezível.

Sobre os diversos estratos da população, são realizadas seleções aleatórias, de forma independente. A amostra é obtida através da agregação das amostras de cada estrato (veja a Figura 3.2).

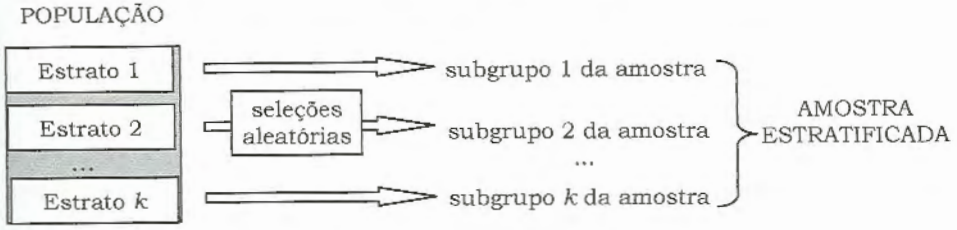


Figura 3.2 Esquema da seleção de uma amostragem estratificada.

Amostragem estratificada proporcional: neste caso particular de amostragem estratificada, a proporcionalidade do tamanho de cada estrato da população é mantida na amostra. Por exemplo, se um estrato corresponde a 20% do tamanho da população, ele também deve corresponder a 20% da amostra. Veja a Figura 3.3.

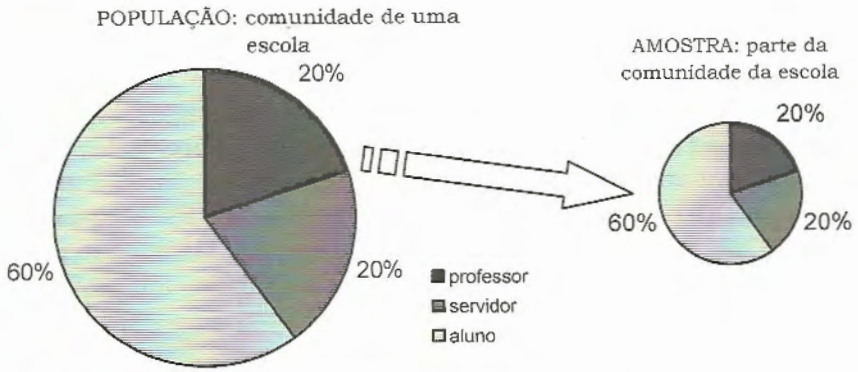


Figura 3.3 Ilustração de uma amostragem estratificada proporcional.

A amostragem estratificada proporcional garante que cada elemento da população tenha a mesma probabilidade de pertencer à amostra.

Exemplo 3.6 Com o objetivo de estudar o estilo de liderança preferido pela comunidade de uma escola, vamos realizar um levantamento por amostragem. A população é composta por 10 professores, 10 servidores técnico-administrativos e 30 alunos, que identificaremos da seguinte maneira:

POPULAÇÃO

Professores:	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Servidores:	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Alunos:	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
	A11	A12	A13	A14	A15	A16	A17	A18	A19	A20
	A21	A22	A23	A24	A25	A26	A27	A28	A29	A30

Supondo que a preferência, quanto ao estilo de liderança, possa ser relativamente homogênea dentro de cada categoria, vamos realizar uma amostragem estratificada proporcional por categoria, para obter uma amostra global de tamanho $n = 10$. A tabela seguinte mostra as relações de proporcionalidade.

Tabela 3.1 Cálculo do tamanho da amostra em cada estrato.

ESTRATO	Proporção na população	Tamanho do subgrupo na amostra
Professores	10/50 = 0,20 (ou 20%)	$n_p = (0,20) \cdot 10 = 2$
Servidores	10/50 = 0,20 (ou 20%)	$n_s = (0,20) \cdot 10 = 2$
Alunos	30/50 = 0,60 (ou 60%)	$n_a = (0,60) \cdot 10 = 6$

Para selecionar aleatoriamente dois professores, usaremos a numeração já existente na população, substituindo o 10 por 0, o que permite usar a Tabela 1 do apêndice com apenas um algarismo. Usando a primeira linha (59 58...), temos os seguintes professores selecionados: {P5, P9}. Para os servidores, usando a segunda linha (53 26...), com o mesmo processo de numeração, temos: {S5, S3}. Para os alunos, precisamos extrair números de dois algarismos. Usando a própria numeração da população e a terceira linha da tabela, temos: {A7, A2, A16, A5, A24, A22}.

A amostra {P5, P9, S5, S3, A7, A2, A16, A5, A24, A22} é uma amostra estratificada proporcional da comunidade da escola. Cada indivíduo desta amostra deverá ser pesquisado para se levantar a característica de interesse, ou seja, o estilo de liderança por ele preferido.

■

Desde que, no problema em estudo, os estratos formam subgrupos mais homogêneos do que a população como um todo, uma amostra estratificada proporcional tende a gerar resultados mais próximos dos parâmetros populacionais, quando comparada com uma amostra aleatória simples de mesmo tamanho.

Amostragem estratificada uniforme: seleciona-se a mesma quantidade de elementos em cada estrato. No exemplo precedente, para se obter uma amostra estratificada uniforme de $n = 12$ indivíduos, devemos selecionar 4 indivíduos de cada categoria.

A amostragem estratificada uniforme costuma ser usada em situações em que o maior interesse é obter estimativas separadas para cada estrato, ou quando se deseja comparar os diversos estratos.

É importante observar que na fase de análise dos dados deve-se levar em conta o planejamento amostral utilizado. Por exemplo, se os dados proveem de uma amostragem estratificada não proporcional, os cálculos de médias e proporções devem ser feitos em cada estrato. Caso se queira uma média ou proporção global, devemos agregar os resultados de cada estrato por uma média aritmética ponderada, tomando como pesos as proporções de cada estrato na população.

Amostragem de conglomerados

Chamamos de *conglomerado* a um agrupamento de elementos da população. Por exemplo, numa população de domicílios residenciais de uma cidade, os quarteirões formam conglomerados de domicílios. Num primeiro estágio, são selecionados alguns conglomerados. Depois, ou se observam todos os elementos dos conglomerados selecionados no primeiro estágio (*amostragem de conglomerados em um estágio*), ou, como é mais comum, faz-se nova seleção, tomando amostras de elementos dos conglomerados extraídos no primeiro estágio (*amostragem de conglomerados em dois estágios*). Todas as seleções devem ser aleatórias (ver Figura 3.4).

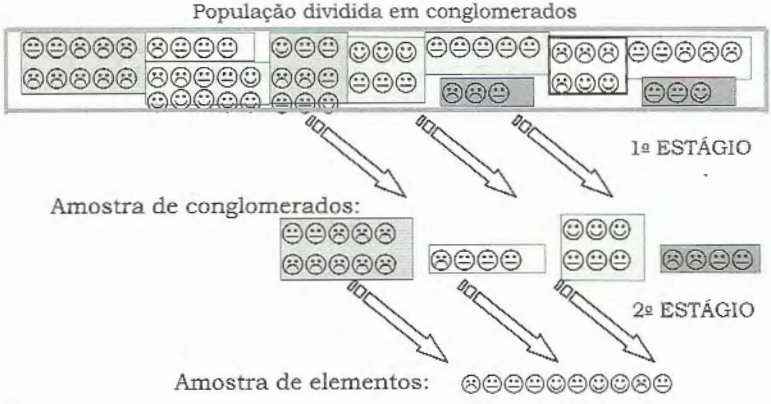


Figura 3.4 Ilustração do processo de amostragem de conglomerados em dois estágios.

Em pesquisas de grande escala, a amostragem pode ser feita em mais estágios. Por exemplo, para selecionar uma amostra de domicílios do estado de Santa Catarina, podemos selecionar municípios (primeiro estágio); dos municípios selecionados, selecionar setores censitários (segundo estágio);[4] e dos setores censitários selecionados, selecionar domicílios (terceiro estágio).

Chamamos de **fração de amostragem** à relação n/N, ou seja, a proporção da população que será efetivamente observada. Se a fração de amostragem for constante para todos os conglomerados selecionados, então todos elementos da população têm a mesma probabilidade de pertencer à amostra.

EXEMPLO 7.7 Seja o problema de selecionar uma amostra de domicílios de uma cidade. Podemos tomar as ruas como conglomerados, como indicado no quadro a seguir, onde *A1* representa o primeiro domicílio da Rua *A*, *A2* o segundo, e assim por diante.

Ruas	Domicílios
A	*A1 A2 A3 A4 A5 A6*
B	*B1 B2 B3 B4 B5 B6 B7 B8 B9 B10 B11 B12 B13 B14*
C	*C1 C2 C3 C4 C5 C6 C7 C8 C9 10*
D	*D1 D2 D3 D4*
E	*E1 E2 E3 E4 E5 E6 E7 E8*

Vamos realizar uma amostragem de conglomerados, selecionando três ruas (primeiro estágio) e, nas ruas selecionadas, uma fração de amostragem de 50% de domicílios (segundo estágio). Então:

1º ESTÁGIO. Seja a seguinte numeração das ruas (unidades de amostragem neste estágio): $1 \rightarrow A$, $2 \rightarrow B$, $3 \rightarrow C$, $4 \rightarrow D$ e $5 \rightarrow E$. Tomemos, por exemplo, os números com um algarismo da sexta linha da tabela de números aleatórios (24 26 56...), que leva à amostra de conglomerados (ruas) *B*, *D* e *E*, pois: $2 \rightarrow B$, $4 \rightarrow D$ e $5 \rightarrow E$.

2º ESTÁGIO. Para satisfazer a fração de amostragem de 50% em cada conglomerado, precisamos selecionar 7 domicílios da Rua *B*, 2 da *D* e 4 da *E*.

Rua B. Tomando números de dois algarismos, a partir da sétima linha da tabela de números aleatórios, e usando a própria numeração de identificação dos domicílios, chegamos a *B9, B2, B1, B11, B12, B3* e *B4*.

[4] Setores censitários são pequenas áreas contíguas, com aproximadamente o mesmo número de domicílios. Essas áreas são determinadas pelo IBGE e usadas em suas pesquisas.

Rua D. Tomando números com um algarismo na décima primeira linha, selecionamos os domicílios *D4* e *D3*.

Rua E. Usando a décima segunda linha, selecionamos *E5, E3, E6* e *E4*.

Amostra selecionada: {*B9, B2, B1, B11, B12, B3, B4, D4, D3, E5, E3, E6, E4*}.

■

O leitor deve observar que, ao contrário dos planos discutidos anteriormente, a amostragem de conglomerados não exige uma lista de todos os elementos da população. Basta, no primeiro estágio, uma lista de conglomerados e, no segundo estágio, uma lista de elementos, mas somente para os conglomerados previamente selecionados.

Ao contrário da amostragem estratificada, as estimativas de uma amostra de conglomerados tendem a gerar resultados mais distantes dos parâmetros populacionais, quando comparada com uma amostra aleatória simples de mesmo tamanho. Contudo, seu custo financeiro tende a ser bem menor.

Exercícios

6) Selecione uma amostra estratificada uniforme, de tamanho $n = 12$, da população do Exemplo 3.6.

7) Considerando a população de funcionários do Exemplo 3.4, faça uma amostragem estratificada proporcional de tamanho $n = 8$, usando a variável *sexo* para a formação dos estratos.

8) O mapa seguinte simboliza os domicílios de um bairro. Os quadros grandes correspondem aos quarteirões, divididos em duas localidades (estratos) do bairro. Os números dentro dos quadradinhos (domicílios) correspondem ao número de cômodos do domicílio, que é a variável a ser levantada na pesquisa.

4 5 2 9 4 7 1 2 6 4	1 4 4 6 4 5 2 3 2 3	7 2 2 4 6 8 2 4 5 6
8 5 2 3 8 5 2 4 5 9	4 1 6 3 4 2 5 6 4 3	2 3 5 4 4 3 4 5 4 2

Estrato A

9 8 18 22 8 9 7 7 9 9	8 7 9 6 14 9 9 8 7 12	14 8 9 8 8 15 8 9 8 8

Estrato B

a) Selecione uma amostra estratificada proporcional de 9 domicílios. Anote o número de cômodos dos domicílios selecionados na amostra.

b) *Faça* uma amostragem de conglomerados em dois estágios. No primeiro estágio, selecione 3 quarteirões e, no segundo estágio, 3 domicílios em cada conglomerado selecionado. Anote o número de cômodos dos domicílios amostrados.

3.3 Amostragens não aleatórias

Existem situações práticas em que a seleção de uma amostra aleatória é muito difícil, ou até mesmo impossível. Geralmente a maior dificuldade está na obtenção de uma lista dos elementos da população. Algumas vezes este problema é contornável pela amostragem aleatória de conglomerados, que exige, inicialmente, apenas uma lista de conglomerados. Em outras vezes, quando nem isso é possível, passamos a pensar em procedimentos não aleatórios para seleção da amostra. Veremos, também, algumas situações em que uma amostragem não aleatória pode ser mais adequada do que uma amostragem aleatória.

Em geral, as técnicas de amostragens não aleatórias procuram gerar amostras que, de alguma forma, representem razoavelmente bem a população de onde foram extraídas. Discutiremos, em particular, a amostragem por cotas e a amostragem por julgamento.

Amostragem por cotas

A amostragem por cotas assemelha-se com a amostragem estratificada proporcional. A população é vista de forma segregada, dividida em diversos subgrupos. Seleciona-se uma cota de cada subgrupo, proporcional ao seu tamanho. Ao contrário da amostragem estratificada, a seleção não precisa ser aleatória. Para compensar a falta de aleatoriedade na seleção, costuma-se dividir a população num grande número de subgrupos. Numa pesquisa socioeconômica, a população pode ser dividida por localidade, por nível de instrução, por faixas de renda, etc.

Amostragem por julgamento

Os elementos escolhidos são aqueles julgados como típicos da população que se deseja estudar. Por exemplo, num estudo sobre a produção

científica dos departamentos de ensino de uma universidade, um estudioso sobre o assunto pode escolher os departamentos que ele considera serem aqueles que melhor representam a universidade em estudo.

No exemplo precedente, a utilização de uma amostragem aleatória pode não ser recomendável, já que temos uma população pequena.[5] Por outro lado, dependendo do que se pretenda estudar sobre produção científica, um levantamento de todos os departamentos pode gastar muito tempo. Então, o uso de uma amostragem por julgamento pode ser uma boa alternativa, mesmo com a limitação de que os resultados desta pesquisa não necessariamente valham para todos os departamentos da universidade.

Estudos comparativos

Os exemplos que vimos neste capítulo tinham como objetivo a descrição de certas características da população. Em muitos casos, o principal objetivo é comparar certas características em duas ou mais populações. Por exemplo, para se comparar o hábito de fumar entre a população de indivíduos com câncer no pulmão e a população de indivíduos sadios, podemos usar duas amostras de indivíduos, sendo uma composta de pessoas com câncer no pulmão, e outra de pessoas sadias.

Por razões práticas, uma amostra de pessoas com câncer no pulmão é geralmente obtida num hospital, tomando-se todas as pessoas em tratamento dessa doença. Obviamente essa amostra não é uma amostra aleatória de toda a população de pessoas com câncer no pulmão. Mas, em estudos comparativos, normalmente o principal objetivo não é a generalidade, mas sim, a busca das verdadeiras diferenças entre as amostras que estão em análise.

Neste contexto, a principal preocupação no plano de amostragem é obter amostras comparáveis, ou seja, que se diferenciem somente com respeito ao fator de comparação. No presente exemplo, o fator de comparação é o atributo de *ter câncer no pulmão*. Assim, as duas amostras devem ser o mais similar possível, a não ser o fato de que uma delas é formada por pessoas *com câncer no pulmão* e a outra não. Nessas duas amostras se estudaria e compararia o *hábito de fumar*.

---

[5] A maioria das universidades brasileiras tem menos de cinquenta departamentos de ensino. Como veremos posteriormente, para grande parte dos estudos de levantamento, uma amostra aleatória razoável deve conter centenas de observações, ou atingir um número de observações próximo ao tamanho de toda a população.

Num estudo experimental, em que é possível controlar os elementos que vão pertencer a cada um dos grupos, a comparabilidade dos grupos (amostras) pode ser obtida por uma *divisão aleatória* dos elementos entre os grupos. Para comparar dois métodos de ensinar matemática para crianças, podemos sortear uma parte das crianças escolhidas para o estudo, alocando-as no grupo de ensino do primeiro método. As outras crianças ficariam no grupo de ensino do outro método. No final do experimento, os dois métodos seriam comparados com respeito ao aprendizado de matemática.

Exercícios

9) Comente sobre os seguintes planos de amostragens, apontando suas incoerências, quando for o caso.
   a) Com a finalidade de estudar o perfil dos consumidores de um supermercado, observaram-se os consumidores que compareceram ao supermercado no primeiro sábado do mês.
   b) Com a finalidade de estudar o perfil dos consumidores de um supermercado, fez-se a coleta de dados durante um mês, tomando a cada dia um consumidor da fila de cada caixa do supermercado, variando sistematicamente o horário da coleta dos dados.
   c) Para avaliar a qualidade dos itens que saem de uma linha de produção, observaram-se todos os itens das 14:00 às 14:30 horas.
   d) Para avaliar a qualidade dos itens que saem de uma linha de produção, observou-se um item a cada meia hora, durante todo o dia.
   e) Para estimar a percentagem de empresas que investiram em novas tecnologias no último ano, enviou-se um questionário a todas as empresas. A amostra foi formada pelas empresas que responderam ao questionário.

10) Num estudo sobre o estado nutricional dos estudantes da rede escolar de uma cidade, decidiu-se complementar os dados antropométricos com alguns exames laboratoriais. Como não se podia exigir que o estudante fizesse esses exames, decidiu-se estratificar a população por nível escolar (fundamental e médio) e por tipo de escola (pública e privada), selecionando voluntários em cada estrato, até completar as cotas. Com base nos dados da tabela abaixo, qual deve ser a cota a ser amostrada em cada estrato, considerando que se deseja uma amostra de 200 estudantes?

Distribuição dos estudantes da rede escolar, segundo o nível e o tipo de escola

	Tipo de escola	
Nível escolar	pública	privada
fundamental	48%	14%
médio	26%	12%

3.4 Tamanho de uma amostra aleatória simples

O cálculo do tamanho da amostra é um problema complexo e, neste livro, ficaremos restritos ao caso da amostragem aleatória simples. Também não abordaremos aspectos financeiros, mesmo sabendo que muitas vezes o tamanho da amostra fica restrito aos recursos disponíveis.

A heterogeneidade da população e os tipos de parâmetros que se quer estimar (proporções, médias, etc.) são pontos importantes na determinação do tamanho da amostra. Esses pontos entrarão em fórmulas mais refinadas, as quais apresentaremos no Capítulo 9. Nesta seção, ficaremos restritos a uma formulação bastante genérica, usada nas pesquisas em que queremos usar a amostra para estimar diversas proporções (ou percentagens).[6]

Conceito de erro amostral

Como já definimos, *parâmetro* é uma medida que descreve certa característica dos elementos da população. De forma análoga, *estatística* é uma medida associada aos elementos da amostra. A estatística, quando usada para avaliar (ou *estimar*) um parâmetro, também é chamada de *estimador*. Por exemplo, na população dos funcionários de uma empresa, π = *percentagem de funcionários favoráveis a um programa de treinamento* é um parâmetro. Numa amostra a ser retirada, P = *percentagem de favoráveis ao programa de treinamento*, na amostra, é uma estatística. P também pode ser considerado um *estimador* do parâmetro π.

Erro amostral é a diferença entre uma estatística e o parâmetro que se quer estimar.

Para a determinação do tamanho da amostra, o pesquisador precisa especificar o **erro amostral tolerável**, ou seja, o quanto ele admite *errar* na avaliação do(s) parâmetro(s) de interesse. Por exemplo, na divulgação de pesquisas eleitorais, é comum encontrarmos no relatório algo como: *a presente pesquisa tolera um erro de 2%*. Isso quer dizer que, quando a pesquisa aponta determinado candidato com 20% de preferência do eleitorado, está afirmando, na verdade, que a preferência por esse candidato, em toda a população de eleitores, é um valor no intervalo de 18% a 22% (ou seja, 20% $\pm$ 2%).

[6] Como a abordagem que estamos apresentando é bastante genérica, ela pode fornecer um tamanho de amostra superior ao tamanho que seria necessário para uma dada situação específica.

A especificação do *erro amostral tolerável* deve ser feita sob um enfoque probabilístico, pois, por maior que seja a amostra, existe o risco de o sorteio gerar uma amostra com características bem diferentes das características da população de onde ela está sendo extraída. Na abordagem preliminar desta seção, consideraremos sempre o erro amostral sob 95% de probabilidade. Assim, se fixarmos o erro amostral tolerável em 2%, estaremos afirmando que uma estatística, calculada com base na amostra a ser selecionada, não deve diferir do parâmetro em mais que 2%, com 95% de probabilidade.

UMA FÓRMULA PARA O TAMANHO MÍNIMO DA AMOSTRA

Sejam: N tamanho (número de elementos) da população;
n tamanho (número de elementos) da amostra;
n_0 uma primeira aproximação para o tamanho da amostra e
E_0 erro amostral tolerável.

Um primeiro cálculo do tamanho da amostra pode ser feito, mesmo sem se conhecer o tamanho da população, através da seguinte expressão:[7]

$$n_0 = \frac{1}{E_0^2}$$

Se a população for muito grande (digamos, mais que vinte vezes o valor calculado n_0), então n_0 já pode ser adotado como tamanho da amostra ($n = n_0$). Caso contrário, é sugerida a seguinte correção:

$$n = \frac{N \cdot n_0}{N + n_0}$$

EXEMPLO 3.8 Planeja-se um levantamento por amostragem para avaliar diversas características (parâmetros) da população das $N = 200$ famílias moradoras de um certo bairro. Os principais parâmetros são proporções (ou percentagens), tais como: *percentagem de famílias que usam programas de alimentação popular, percentagem de famílias que moram em casas próprias*, etc. Qual deve ser o tamanho mínimo de uma amostra aleatória simples para que possamos admitir, com 95% de probabilidade, que os erros amostrais não ultrapassem 4% ($E_0 = 0{,}04$)?

Solução. Primeiramente:

[7] Lembramos que esta expressão é voltada para a estimação de *proporções*, com probabilidade aproximada de 95% do erro amostral não superar E_0. No Capítulo 9 voltaremos a esta discussão.

$$n_0 = \frac{1}{(0{,}04)^2} = 625$$

Corrigindo, em função do tamanho N da população, temos:

$$n = \frac{(200)\cdot(625)}{200+625} = \frac{125.000}{825} = 152 \text{ famílias.}$$

■

Exemplo 3.9 Considerando os objetivos e os valores fixados no exemplo anterior, qual deveria ser o tamanho da amostra se a pesquisa fosse ampliada para todo o município, que contém $N = 200.000$ famílias residentes?

Solução. O valor de n_0 continua o mesmo do caso anterior ($n_0 = 625$), mas com a correção em termos do novo valor de N, temos:

$$n = \frac{(200.000)\cdot(625)}{200.000+625} = 623 \text{ famílias.}$$

No Exemplo 3.9, praticamente não houve alteração com a correção em termos do tamanho N da população ($n_0 = 625$ e $n = 623$). Em geral, se a população for muito grande, podemos usar n_0 como o tamanho da amostra ($n = n_0$).

■

No Exemplo 3.8, para garantir o erro amostral não superior a 4%, foi necessária uma amostra abrangendo 76% da população (152 elementos extraídos de 200); enquanto no Exemplo 3.9 foi suficiente uma amostra de apenas 0,3% da população (623 de 200.000). Portanto, é errônea a ideia de que para uma amostra ser representativa ela deva abranger uma percentagem fixa da população (veja a Figura 3.5).

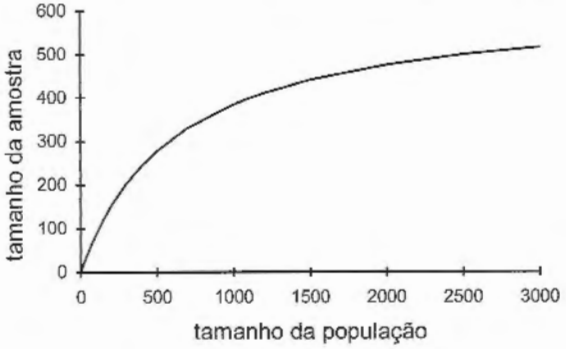


Figura 3.5 Relação entre tamanho da população e tamanho da amostra para um dado erro amostral.

Tamanho da amostra em subgrupos da população

É comum termos interesse em estudar separadamente certos subgrupos da população. Por exemplo, numa pesquisa eleitoral, podemos ter interesse em saber as preferências das mulheres e dos homens. Numa pesquisa sobre condições socioeconômicas das famílias de uma cidade, podemos querer apresentar resultados para cada bairro da cidade.

Quando precisamos efetuar estimativas sobre partes (subgrupos) da população, é necessário calcular o tamanho da amostra para cada uma dessas partes. O tamanho total da amostra vai corresponder à soma dos tamanhos das amostras dos subgrupos. Pelo exposto, o tamanho total da amostra pode ser muito grande. Por isso, o pesquisador não deve ser muito exigente na precisão das estimativas nos subgrupos, tolerando erros amostrais maiores.

Exemplo 3.10 Seja o problema do Exemplo 3.9, mas suponha que se queira fazer estimativas isoladas para os seguintes estratos: (1) centro da cidade, (2) bairros e (3) periferia, mantendo-se a mesma precisão para cada estrato (E_0 = 0,04). Seriam necessárias:

$$n = \frac{1}{E_0^2} = \frac{1}{(0{,}04)^2} = 625$$

Portanto, a amostra total deve conter: $n_{total} = 3 \cdot (625) = 1.875$ famílias. ■

Observamos que na fase de análise dos dados, os cálculos são feitos para cada estrato. Para se ter resultados de todo o município, é necessário agregar os resultados dos estratos por uma média ponderada, tomando-se como peso o tamanho relativo de cada estrato no município.

Exercícios

11) Para estudar a preferência do eleitorado a uma semana da eleição presidencial, qual deve ser o tamanho de uma amostra aleatória simples de eleitores para garantir, com 95% de probabilidade, um erro amostral não superior a 2%?

22) Numa empresa com 1.000 funcionários, deseja-se estimar a percentagem de funcionários favoráveis a um certo programa de treinamento. Qual deve ser o tamanho de uma amostra aleatória simples que garanta, com 95% de probabilidade, um erro amostral não superior a 5%?

3.5 Fontes de erros nos levantamentos por amostragem

O *erro amostral*, definido como a *diferença* entre uma *estatística* (a ser calculada com base em uma amostra de n elementos) e o verdadeiro valor do *parâmetro* (característica de uma população de N elementos), parte do princípio de que as n observações da amostra são obtidas sem erros. Havendo *erros* ou *desvios* nos dados da própria amostra, a diferença entre a estatística e o parâmetro pode ser maior que o limite tolerável, E_0. Por isso, o planejamento e a execução da pesquisa devem ser feitos com muita cautela, para evitar, ou reduzir os erros nos próprios dados da amostra, conhecidos como *erros não amostrais*. Abordaremos alguns desses erros, comuns em pesquisas de levantamentos.

População acessível diferente da população-alvo

Muitas vezes, queremos pesquisar uma certa população-alvo, mas, por conveniência, retiramos uma amostra de um conjunto incompleto de elementos (*população acessível* ou *população amostrada*). Por exemplo, numa pesquisa eleitoral para avaliar a preferência dos eleitores de um município, costuma-se tomar como base para a seleção da amostra a lista de domicílios residenciais do município, o que deixa inacessíveis os eleitores que moram em outros municípios, mas com domicílio eleitoral no município em estudo.

Devemos concentrar esforços para retirar a amostra de toda a população-alvo. Quando isso não for possível, devemos limitar a abrangência da pesquisa à população que foi efetivamente estudada.

POPULAÇÃO COM TELEFONE

Falta de resposta

É comum não conseguirmos respostas de alguns elementos selecionados na amostra, como ocorre frequentemente quando a população em estudo é a humana, pois nem todos se dispõem à responder a um questionário ou dar uma entrevista. O entrevistador, eticamente e respeitando o direito do entrevistado em não participar, deve ter capacidade de persuasão e empenhar-se para conseguir a participação do maior número possível dos indivíduos selecionados.

Uma prática muito comum, mas que pode levar a sérias distorções nos resultados, é a de substituir indivíduos que se recusam a responder ou que não são encontrados no momento da pesquisa. Para evitar esse problema, devemos efetuar vários retornos aos elementos selecionados na amostra.

RESPOSTA SOCIALMENTE ACEITÁVEL

VOCÊ É CORRUPTO?

Erros de mensuração

Nem sempre conseguimos medir exatamente aquilo que queremos. Por exemplo, numa pesquisa eleitoral, o eleitor pode, por várias razões, apontar um candidato, quando na verdade ele pretende votar em outro.

Podemos reduzir a ocorrência desse tipo de erro com a elaboração de um questionário que tenha alguns itens de controle, capazes de detectar algumas *más respostas*. Um bom treinamento dos entrevistadores também ajuda a reduzir esses erros.

Além desses três tipos de erros não amostrais, poderíamos citar muitos outros. O pesquisador, ao aplicar métodos adequados de estatística, consegue avaliar, de alguma forma, a magnitude provável dos *erros amostrais*. Mas o tratamento dos *erros não amostrais* é mais difícil e depende fundamentalmente do planejamento e execução da pesquisa.

Exercícios complementares

13) Considere a seguinte população composta de 40 crianças do sexo masculino (representados por H1, H2,..., H40) e 20 crianças do sexo feminino (representadas por M1, M2,..., M20).

H1	H2	H3	H4	H5	H6	H7	H8	H9	H10
H11	H12	H13	H14	H15	H16	H17	H18	H19	H20
H21	H22	H23	H24	H25	H26	H27	H28	H29	H30
H31	H32	H33	H34	H35	H36	H37	H38	H39	H40
M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
M11	M12	M13	M14	M15	M16	M17	M18	M19	M20

a) Retire desta população de 60 crianças, uma amostra aleatória simples de tamanho n = 10. Use a primeira coluna da tabela de números aleatórios.
b) Retire desta população uma amostra aleatória estratificada proporcional de tamanho n = 12, usando o sexo como variável estratificadora. Use a segunda coluna da tabela de números aleatórios para o estrato dos homens e a terceira coluna para o estrato das mulheres.
c) Se o estudo tem por objetivo avaliar o tipo de brincadeira preferida pelas crianças, qual é o tipo de amostra você acredita ser a mais adequada? E se for para avaliar o quociente de inteligência? Justifique suas respostas.

14) Uma empresa tem 3.414 empregados repartidos nos seguintes departamentos: Administração (914), Transporte (348), Produção (1.401) e Outros (751). Deseja-se extrair uma amostra para verificar o grau de satisfação em relação à qualidade da comida do refeitório. Apresente um plano de amostragem para esse problema.

Parte II

Descrição e exploração de dados

Como extrair informações dos dados

Como construir, apresentar e interpretar tabelas, gráficos e medidas descritivas

Capítulo 4

Dados categorizados

Nos três próximos capítulos, vamos considerar que os dados já foram efetivamente observados, sejam de uma amostra ou de uma população. E o objetivo básico consistirá em introduzir técnicas que permitam organizar, resumir e apresentar esses dados, de tal forma que possamos interpretá-los à luz dos objetivos da pesquisa. Esta parte do tratamento dos dados é chamada de *Estatística Descritiva.*

Com os dados adequadamente resumidos e apresentados em tabelas e gráficos, poderemos observar determinados aspectos relevantes e começar a delinear hipóteses a respeito da estrutura do universo em estudo. É a chamada *Análise Exploratória de Dados.*

No presente capítulo, aprenderemos a descrever e explorar dados de *variáveis qualitativas*, isto é, variáveis cujos possíveis resultados são observados na forma de categorias. É o caso de variáveis como *nível de instrução, sexo, estado civil,* etc. Por exemplo, ao observar a variável *sexo* (*gênero*) num conjunto de indivíduos, estaremos classificando cada indivíduo na categoria *masculino* ou na categoria *feminino.*

4.1 Classificação simples

Iniciaremos o tratamento de dados analisando isoladamente cada variável (*análise univariada*).

Um dos primeiros passos para entendermos o comportamento de uma variável, em termos dos elementos observados, é a construção de uma distribuição de frequências.

A **distribuição de frequências** compreende a organização dos dados de acordo com as ocorrências dos diferentes resultados observados. Ela pode ser apresentada sob forma tabular ou gráfica.

Para ilustrar a construção de uma distribuição de frequências, considere os dados de um levantamento de uma amostra de 40 famílias do Conjunto Residencial Monte Verde, com respeito à variável *nível de instrução do chefe da casa* (ver anexo deste capítulo).

Dados do *último nível de instrução* completado pelo chefe da casa (códigos: 1 – *nenhum*; 2 – *fundamental* e 3 – *médio*):

3 3 2 2 3 1 3 3 3 2 2 1 2 2 3 2 3 3 3 3
3 3 3 2 2 3 1 3 2 3 3 2 3 1 1 1 3 3 3 3

Para construir uma distribuição de frequências com dados de uma variável qualitativa, basta *contar* a quantidade de resultados observados em cada categoria (ver Tabela 4.1).[1]

Tabela 4.1 Distribuição de frequências do último nível de instrução completado pelo chefe da casa, numa amostra de 40 famílias do conjunto residencial Monte Verde, Florianópolis – SC, 1988.

Nível de Instrução	Frequência	Percentagem
nenhum	6	15,0
fundamental	11	27,5
médio	23	57,5
Total	40	100,0

A primeira coluna da Tabela 4.1 mostra todas as categorias previamente estabelecidas da variável *nível de instrução*. A segunda coluna resulta da *contagem* de quantas observações se identificam com cada categoria; são as frequências observadas. Finalmente, a terceira coluna

[1] A apresentação de tabelas num relatório é regida por normas específicas elaboradas pelo Instituto Brasileiro de Geografia e Estatística (IBGE) e adotadas pela Associação Brasileira de Normas Técnicas (ABNT). Toda tabela deve ser auto-explicativa, sendo necessário um título que informe ao leitor *o que* está sendo apresentado, *onde* e *quando* foram coletados os dados. Uma tabela tem sua estrutura formada por três linhas horizontais, sendo duas que delimitam o cabeçalho e uma que faz o fechamento. Qualquer outra linha vertical ou horizontal poderá ser traçada, desde que venha contribuir para melhor leitura dos dados da tabela, mas ela não deve ser fechada nas verticais. Alguma explicação complementar pode ser colocada no rodapé da tabela, em particular, a *fonte*, quando se trata de dados secundários. A inserção de uma tabela num relatório somente deve ser feita após ela ser referenciada no texto.

apresenta uma medida relativa da frequência de cada categoria. As *percentagens* são obtidas dividindo-se a frequência de cada categoria pelo número total de observações e, em seguida, multiplicando-se por 100 (cem). As medidas relativas (percentagens) são particularmente importantes para comparar distribuições de frequências.

A Tabela 4.2 mostra três distribuições de frequências. A primeira corresponde à distribuição da Tabela 4.1, e as outras duas às distribuições do nível de instrução do chefe da casa em outras duas localidades.[2]

Tabela 4.2 Distribuição de frequências do último nível de instrução completado pelo chefe da casa, numa amostra de 120 famílias, dividida segundo as localidades do bairro Saco Grande II, Florianópolis – SC, 1988.

Nível de instrução	Localidade		
	Monte Verde	Pq. da Figueira	Encosta do Morro
nenhum	6 (15,0)	14 (32,6)	18 (48,7)
fundamental	11 (27,5)	14 (32,6)	13 (35,1)
médio	23 (57,5)	15 (34,8)	6 (16,2)
Total	40 (100,0)	43 (100,0)	37 (100,0)

NOTA: Os números entre parênteses correspondem às percentagens em relação ao total de famílias observadas em cada localidade.

Interpretação da Tabela 4.2 – As famílias pesquisadas no Conjunto Residencial Monte Verde apresentam, relativamente, os chefes da casa com os melhores níveis de instrução. Por outro lado, temos nas famílias pesquisadas na Encosta do Morro o pior perfil, em termos de grau de instrução do chefe da casa, com quase 50% deles não tendo concluído nem o fundamental.[3]

O leitor deve notar que, ao organizar e resumir os dados numa distribuição de frequências, não é dada a informação de quais elementos pertencem a cada categoria (por exemplo, quais indivíduos não têm nem o nível de instrução fundamental não aparece na distribuição de frequências do nível de instrução). Contudo, para entender o comportamento geral de uma variável, essa informação normalmente não é relevante.

[2] Uma tabela do tipo Tabela 4.2, pelo seu formato, é conhecida como *tabela de dupla entrada* ou *tabela de contingência*.

[3] Note que a análise é feita especificamente com respeito às famílias pesquisadas. Inferências para a população serão discutidas a partir do Capítulo 9.

EXERCÍCIOS

1) Com base nos dados do anexo deste capítulo, construa uma tabela de frequências para a variável PAP (uso, ou não, de programas de alimentação popular), considerando, apenas, as famílias residentes no Conjunto Residencial Monte Verde.
2) Construa uma distribuição de frequências para a variável PAP (ver anexo), para cada localidade em estudo. Apresente essas distribuições numa tabela de dupla entrada e interprete.
3) Sejam os resultados da pesquisa descrita na Seção 2.4, cujos dados estão no anexo do Capítulo 2. Faça uma distribuição de frequências para o *principal ponto positivo do Curso de Ciências da Computação da UFSC, na visão do aluno.* Interprete.

4.2 REPRESENTAÇÕES GRÁFICAS

As representações gráficas fornecem, em geral, uma visualização mais sugestiva do que as tabelas. Portanto, constituem-se numa forma alternativa de apresentação de distribuições de frequências. Nesta seção, apresentaremos o gráfico de barras e o gráfico de setores, que são particularmente importantes na representação de distribuições de frequências de dados categorizados.

GRÁFICO DE BARRAS

A Figura 4.1 representa a distribuição de frequências da Tabela 4.1 por um gráfico de barras. Cada categoria é representada por uma barra de comprimento proporcional à sua frequência (número de famílias), conforme identificação do eixo horizontal.[4]

Opcionalmente, pode-se apresentar as categorias no eixo horizontal e a frequência no eixo vertical. É o chamado *gráfico de colunas.*

[4] Da mesma forma que as tabelas, as figuras devem conter um título, contendo as informações do seu conteúdo e colocado abaixo dela.

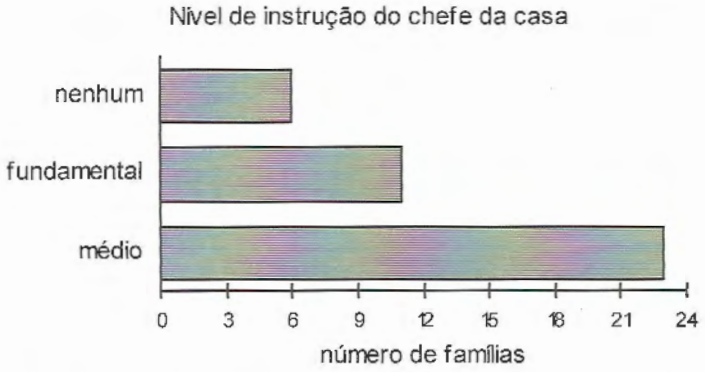


Figura 4.1 Distribuição de frequências do último nível de instrução completado pelo chefe da casa, numa amostra de quarenta famílias do Conjunto Residencial Monte Verde, Florianópolis – SC, 1988.

Gráfico de setores

Para construir um gráfico de setores, basta fazer uma relação entre um ângulo, em graus, e a frequência observada em cada categoria, lembrando que um círculo tem 360°. O esquema, a seguir, mostra esta relação para a categoria *nenhum*:

Relação entre o tamanho do setor (α_1) e o círculo todo (360°).	=	Relação entre a frequência da categoria (6) e o total observado (40).

$$\frac{\alpha_1}{360^\circ} = \frac{6}{40}$$

Donde: $\alpha_1 = \frac{6}{40}(360) = 54^\circ$

Repetindo a "regra de três" para as outras categorias, temos:

categoria 1 (*nenhum*): setor de tamanho $\alpha_1 = 54^\circ$;
categoria 2 (*fundamental*): setor de tamanho $\alpha_2 = 99^\circ$;
categoria 3 (*médio*): setor de tamanho $\alpha_3 = 207^\circ$.

Com a ajuda de um transferidor, podemos construir o gráfico indicado na Figura 4.2.

Figura 4.2 Distribuição de frequências do último nível de instrução completado pelo chefe da casa, numa amostra de quarenta famílias do Conjunto Residencial Monte Verde, Florianópolis – SC, 1988.

Em se tratando da descrição de dados de variáveis ordinais, como no presente caso, recomendamos os gráficos de barras ou de colunas, que permitem enfatizar a ordem das categorias.

Gráfico de barras múltiplas

Para efetuar uma análise comparativa de várias distribuições, podemos construir vários gráficos de setores, ou um gráfico de barras múltiplas, como na Figura 4.3, que representa graficamente as distribuições de frequências da Tabela 4.2. No eixo horizontal, optamos por colocar as frequências relativas, em forma de percentagens, para facilitar a comparação.

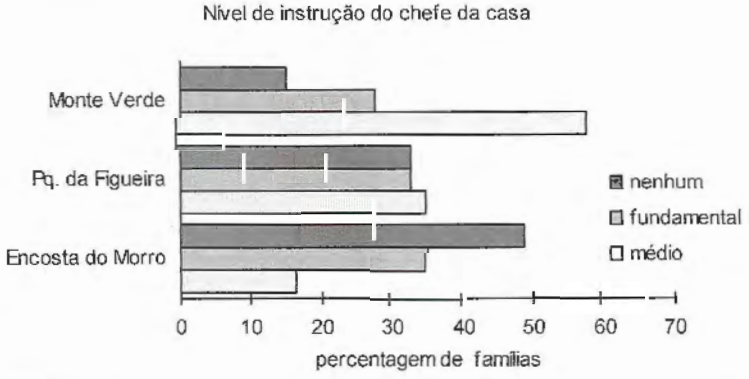


Figura 4.3 Distribuição de frequências do último nível de instrução completado pelo chefe da casa, numa amostra de 120 famílias, dividida segundo as localidades do bairro Saco Grande II, Florianópolis – SC, 1988.

Que tipo de gráfico usar?

Para representar distribuições de frequências de variáveis qualitativas nominais com poucas categorias, o gráfico de setores tem sido muito usado, principalmente devido a sua visualização, possibilidade de apresentação em três dimensões e possibilidade de destacar alguma categoria através de um leve afastamento do setor.

Quando a variável é ordinal, gráficos de barras ou de colunas são mais indicados, pois permitem manter a ordem das categorias. Esses gráficos também são mais adequados quando se têm muitas categorias ou quando se quer dar mais destaque às categorias mais frequentes. Neste último caso, podemos ordenar as categorias pelas frequências.

Gráficos de barras (ou de colunas) múltiplas são usados para representar mais de uma distribuição de frequências, ou distribuições de frequências conjuntas de duas variáveis qualitativas, como as que serão vistas na próxima seção.

Distribuições de frequências de variáveis quantitativas têm gráficos próprios, como os histogramas, que serão estudados no Capítulo 5. Já no Capítulo 13 serão apresentados os diagramas de dispersão, que permitem analisar possíveis relações entre duas variáveis quantitativas.

Exercícios

4) Faça um gráfico de barras e um gráfico de setores para representar a distribuição de frequências do Exercício 1.

5) Faça um gráfico de barras múltiplas para representar as distribuições de frequências do Exercício 2.

4.3 Dupla classificação

Este tópico focaliza uma análise conjunta de duas variáveis qualitativas (*análise bivariada*).

Nas Ciências Sociais e Humanas, é comum o interesse em verificar se duas variáveis apresentam-se associadas num certo conjunto de elementos. Por exemplo, pode-se ter interesse em verificar se o percentual de *usuários de programas de alimentação popular* varia de acordo com a *faixa de renda*, o que caracteriza uma *associação* entre o *uso de programas de alimentação popular* e *a faixa de renda* nas famílias pesquisadas. Esse

tipo de análise passa pelas distribuições conjuntas de frequências, que geralmente são apresentadas nas chamadas *tabelas de contingência* ou *tabelas de dupla entrada*, como veremos a seguir.

Para construirmos uma distribuição conjunta de frequências, devemos observar simultaneamente as duas variáveis nos elementos em estudo. A Figura 4.4 mostra a construção de uma distribuição conjunta, com as variáveis *nível de instrução do chefe da casa* e *uso de programas de alimentação popular*.

As cinco primeiras observações das variáveis *nível de instrução do chefe da casa* e *uso de programas de alimentação popular* (anexo deste capítulo).
Códigos do nível de instrução: 1 – *nenhum*; 2 – *primeiro grau* e 3 – *segundo grau*.
Códigos do uso de programas: 1 – *sim* e 0 – *não*.

Dados		
família	nível de instrução	uso de programas
1	3	0
2	3	0
3	2	1
4	2	0
5	3	1
...	...	...

Construção da tabela

Uso de programas	Nível de instrução		
	1	2	3
1		|	|
0			||

Figura 4.4 Esquema de como fazer a contagem para uma distribuição conjunta.

Para a construção da distribuição conjunta de frequências, cada elemento (família) deve pertencer a uma e apenas uma célula da tabela.[5] Fazendo a classificação de todas as famílias observadas e contando as frequências em cada célula, chegamos à Tabela 4.3. O leitor deve notar que os totais das colunas formam a distribuição de frequências da variável *nível de instrução do chefe da casa*, quando observada isoladamente; enquanto os totais das linhas constituem a distribuição da variável *uso de programas de alimentação popular*.

Tabela 4.3 Distribuição conjunta de frequências do nível de instrução do chefe da casa e uso de programas de alimentação popular.

Uso de programas	Nível de instrução do chefe da casa			Total
	nenhum	fundamental	médio	
sim	31	22	25	78
não	7	16	19	42
Total	38	38	44	120

[5] Chamamos de *célula* ao cruzamento de uma linha com uma coluna.

Para facilitar a análise de uma tabela de contingência, podemos incluir as frequências relativas (percentagens), que podem ser calculadas em relação aos totais das linhas ou colunas, dependendo do objetivo. Na Tabela 4.4 são incluídas as percentagens em relação aos totais das colunas. Esta tabela evidencia os perfis do uso de programas de alimentação popular, considerando as famílias separadas por nível de instrução do chefe da casa (*perfis coluna*).

Tabela 4.4 Distribuição do uso de programas de alimentação popular, por nível de instrução do chefe da casa.

Uso de programas	Nível de instrução do chefe da casa			Total
	nenhum	fundamental	médio	
sim	31 (81,6)	22 (57,9)	25 (56,8)	78 (65,0)
não	7 (18,4)	16 (42,1)	19 (43,2)	42 (35,0)
Total	38 (100,0)	38 (100,0)	44 (100,0)	120 (100,0)

NOTA: Os números entre parênteses são percentagens em relação aos totais das colunas.

Interpretação da Tabela 4.4 – Nos dados observados, verifica-se uma associação entre o uso de programas de alimentação popular e o nível de instrução do chefe da casa, pois, enquanto no nível de instrução mais baixo, a grande maioria das famílias pesquisadas usam os programas (81,6%), no nível de instrução mais alto, pouco mais da metade usam esses programas (56,8%).[6]

A Tabela 4.5 mostra a Tabela 4.3 acrescida de percentagens em relação ao total das linhas. Esta tabela evidencia os perfis do nível de instrução do chefe da casa, considerando a amostra dividida em famílias que usam e famílias que não usam os programas (*perfis linha*). A interpretação da Tabela 4.5 é deixada para o leitor.

Tabela 4.5 Distribuição do nível de instrução do chefe da casa, segundo o uso de programas de alimentação popular.

Uso de programas	Nível de instrução do chefe da casa			Total
	nenhum	fundamental	médio	
sim	31 (39,7)	22 (28,2)	25 (32,1)	78 (100,0)
não	7 (16,7)	16 (38,1)	19 (45,2)	42 (100,0)
Total	38 (31,7)	38 (31,7)	44 (36,7)	120 (100,0)

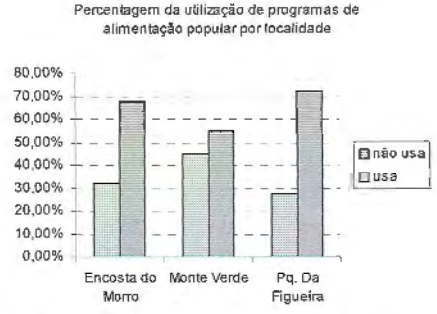
NOTA: Os números entre parênteses são percentagens em relação aos totais das linhas.

[6] Uma análise estatística mais elaborada, como veremos no Capítulo 12, poderá detectar se essa associação é realmente válida para toda a população de famílias do bairro em estudo.

Na Seção 4.1, quando discutíamos classificação simples, juntamos três distribuições de frequências da variável *nível de instrução do chefe da casa*, correspondentes a três localidades diferentes (Tabela 4.2). Observamos, agora, que esse tipo de tabela também pode ser analisado como uma tabela de contingência, como apresentado nesta seção, mesmo que na sua construção não tenhamos observado simultaneamente as duas variáveis, pois as *localidades* já estavam previamente estabelecidas – constituem *estratos* da população.

Uso do computador

Com o uso de programas computacionais de estatística, ou mesmo com planilhas eletrônicas, as tabelas e gráficos podem ser feitos com relativa facilidade. A Figura 5.5 mostra uma tabela e um gráfico feitos com o auxílio do *Microsoft Excel®*, utilizando os dados sobre localidade e uso de programas de alimentação popular do anexo.[7] Deixamos a interpretação da saída computacional como exercício para o leitor.



Contagem de p.a.p	Local			
p.a.p	Encosta do Morro	Monte Verde	Pq. da Figueira	Total Global
não usa	32,43%	45,00%	27,91%	35,00%
usa	67,57%	55,00%	72,09%	65,00%
Total Global	100,00%	100,00%	100,00%	100,00%

Figura 4.4 Saída computacional do relatório de tabela e gráficos dinâmicos do Excel®.

Exercícios

6) Considerando os dados do anexo deste capítulo, classifique as famílias com renda mensal de até 5 salários mínimos, como de *renda baixa*; famílias com rendimentos mensais acima de 5 salários mínimos, como de *renda alta*. A

[7] Em www.inf.ufsc.br/~barbetta/livro1.htm você pode obter algumas orientações sobre o uso do Excel para análise exploratória de dados.

amostra sugere alguma associação entre *renda familiar* e *uso de programas de alimentação popular*? Justifique através da construção e interpretação de uma tabela de contingência.

7) As tabelas a seguir baseiam-se numa amostra de adolescentes de Santa Catarina (Fundação Promover – SC, 1990). Calcule os perfis de percentagens que julgar mais convenientes e interprete.

Tabela 1 Relação entre *participação religiosa* e *uso de bebidas alcoólicas*.

Participação religiosa	Uso de bebidas alcoólicas	
	sim	não
frequentemente	82	460
às vezes	323	921
não participa	86	126

Tabela 2 Relação entre *alegria* e *satisfação sexual*.

Sentimento do respondente	Satisfação sexual	
	satisfeito	frustrado
alegre	525	69
triste	34	19

8) Ao estudar, numa certa população, a possível associação entre *nível de instrução* e *uso de programas de alimentação popular*, suspeita-se que a variável *renda familiar* esteja induzindo esta associação. A Tabela 1 apresenta os elementos classificados segundo o nível de instrução (baixo ou alto) e quanto ao uso de programas de alimentação popular (sim ou não). A Tabela 2 faz a mesma classificação, mas separando os indivíduos em termos da renda familiar (baixa ou alta).

Tabela 1 Elementos classificados segundo o *nível de instrução* e *uso de programas de alimentação popular*.

Nível de instrução	Uso de programas	
	sim	não
baixo	350	150
alto	200	300

Tabela 2 Elementos classificados segundo a renda familiar, *nível de instrução* e *uso de programas de alimentação popular*.

Renda familiar	Nível de instrução	Uso de programas	
		sim	não
baixa	baixo	320	80
	alto	80	20
alta	baixo	30	70
	alto	120	280

a) Qual é a sua conclusão sobre a associação entre o *nível de instrução* e *uso de programas de alimentação popular*, sem levar em conta a *renda familiar* (Tabela 1)?

b) Analisando a Tabela 2, isto é, considerando também a *renda familiar*, o que você conclui?

Exercícios complementares

9) Com o objetivo de verificar se existe associação entre a carreira escolhida (Economia, Administração ou Ciências Contábeis) e tabagismo (fumante ou não fumante), numa determinada faculdade, fez-se uma enquete onde se verificaram os seguintes dados: dos 620 alunos do Curso de Economia, 157 eram fumantes; dos 880 alunos do Curso de Administração, 218 eram fumantes; e dos 310 alunos das Ciências Contábeis, 77 eram fumantes. Apresente estes dados numa tabela de contingência (ou tabela de dupla entrada), calcule percentagens que facilitem visualizar uma possível associação e discuta se os dados sugerem uma associação.

10) Os dados a seguir referem-se à participação em programas de treinamento (1 = *sim* e 0 = *não*) e desempenho no trabalho (1 = *ruim*, 2 = *regular*, 3 = *bom*) dos 30 funcionários de uma empresa.

Ind.	partic.	desemp.	Ind.	partic.	desemp.	Ind.	partic.	desemp.
1	1	2	11	0	2	21	1	2
2	1	3	12	0	1	22	0	2
3	1	3	13	0	2	23	0	1
4	0	2	14	0	1	24	0	1
5	0	1	15	1	2	25	1	3
6	1	1	16	1	3	26	0	1
7	0	1	17	0	1	27	0	2
8	1	3	18	1	2	28	1	3
9	1	3	19	0	1	29	0	3
10	0	1	20	0	2	30	1	3

a) Construa uma distribuição de frequências para cada variável, apresentando-as em forma gráfica.

b) Construa a distribuição de frequências conjunta. Apresente esta distribuição numa tabela de dupla entrada, calculando percentagens que enfatizam o desempenho dos funcionários em cada grupo (participantes e não participantes).

11) Os alunos do Curso de Psicologia da UFSC (turma 302, sem. 99/2) realizaram uma pesquisa com moradores de Florianópolis, à respeito da coleta seletiva de lixo. Uma das tabelas é apresentada a seguir:

Sistema de coleta seletiva de lixo

Nível de instrução do respondente	conhece		colabora	
	sim	não	sim	não
nenhum	12	9	9	10
fundamental	23	3	16	15
médio	43	3	30	22
superior incompleto	25	1	13	19
superior completo	50	1	26	27

Calcule percentagens que facilitem a interpretação da tabela e descreva as principais informações.

Anexo

Este anexo contém parte dos dados de entrevistas realizadas em famílias residentes no Saco Grande II, Florianópolis – SC, 1988.[8] A pesquisa foi realizada pela UFSC e tinha como objetivo principal avaliar os efeitos políticos dos programas de alimentação popular. Transcrevemos, a seguir, algumas das variáveis levantadas, numa amostra de 120 famílias.

Variáveis e códigos

Local (localidade da moradia):

1 = Conjunto Residencial Monte Verde;
2 = Conjunto Residencial Parque da Figueira;
3 = Encosta do Morro.

P.a.p. (uso de algum programa de alimentação popular):

0 = não;
1 = sim.

Instr. (último nível de instrução completado pelo chefe da casa):

1 = nenhum;
2 = fundamental;
3 = médio.

Tam. (número de pessoas residentes no domicílio).

Renda (renda familiar mensal, em quantidade de salários mínimos).

Dados de 120 famílias

Nº	Local	P.a.p.	Instr.	Tam.	Renda	Nº	Local	P.a.p.	Instr.	Tam.	Renda
1	1	0	3	4	10,3	17	1	1	3	3	8,9
2	1	0	3	4	15,4	18	1	0	3	4	12,9
3	1	1	2	4	9,6	19	1	0	3	4	5,1
4	1	0	2	5	5,5	20	1	1	3	4	12,2
5	1	1	3	4	9,0	21	1	1	3	5	5,8
6	1	1	1	1	2,4	22	1	1	3	5	12,9
7	1	0	3	2	4,1	23	1	0	3	5	7,7
8	1	1	3	3	8,4	24	1	0	2	4	1,1
9	1	1	3	6	10,3	25	1	0	2	8	7,5
10	1	1	2	4	4,6	26	1	1	3	4	5,8
11	1	0	2	6	18,6	27	1	1	1	5	7,2
12	1	1	1	4	7,1	28	1	0	3	3	8,6
13	1	0	2	4	12,9	29	1	1	2	4	5,1
14	1	0	2	6	8,4	30	1	0	3	5	2,6
15	1	0	3	3	19,3	31	1	1	3	5	7,7
16	1	0	2	5	10,4	32	1	1	2	2	2,4

[8] Hoje a região pesquisada compreende os bairros Saco Grande e Monte Verde.

Nº	Local	P.a.p.	Instr.	Tam.	Renda	Nº	Local	P.a.p.	Instr.	Tam.	Renda
33	1	1	3	5	4,8	77	2	1	3	4	2,7
34	1	1	1	2	2,1	78	2	0	2	4	2,4
35	1	1	1	6	4,0	79	2	0	2	4	3,6
36	1	1	1	8	12,5	80	2	0	3	5	6,4
37	1	1	3	3	6,8	81	2	0	3	2	11,3
38	1	1	3	5	3,9	82	2	1	1	5	3,8
39	1	0	3	5	9,0	83	2	1	2	3	4,1
40	1	0	3	3	10,9	84	3	1	1	5	1,8
41	2	1	2	5	5,4	85	3	1	3	5	7,1
42	2	1	1	3	6,4	86	3	0	1	3	13,9
43	2	1	1	6	4,4	87	3	1	2	6	4,0
44	2	1	1	5	2,5	88	3	1	1	6	2,9
45	2	0	1	6	5,5	89	3	1	2	9	3,9
46	2	1	1	8	.	90	3	1	1	4	2,2
47	2	1	3	4	14,0	91	3	0	2	3	5,8
48	2	1	2	4	8,5	92	3	0	2	5	2,8
49	2	1	1	5	7,7	93	3	1	2	5	4,5
50	2	0	2	3	5,8	94	3	0	2	4	5,8
51	2	1	3	5	5,0	95	3	0	3	8	3,9
52	2	0	1	3	4,8	96	3	0	2	7	2,8
53	2	1	2	2	2,8	97	3	1	1	3	1,3
54	2	1	2	4	4,2	98	3	1	3	5	3,9
55	2	1	3	3	10,2	99	3	1	3	5	5,0
56	2	1	2	4	7,4	100	3	1	1	5	0,1
57	2	1	2	5	5,0	101	3	0	2	3	4,6
58	2	0	3	2	6,4	102	3	1	2	4	2,6
59	2	0	3	4	5,7	103	3	0	1	6	2,3
60	2	1	2	4	10,8	104	3	1	2	5	4,9
61	2	0	3	1	2,3	105	3	1	1	5	2,3
62	2	1	1	7	6,1	106	3	1	1	3	3,9
63	2	1	1	3	5,5	107	3	1	1	4	2,1
64	2	1	1	7	3,5	108	3	1	1	4	2,7
65	2	1	3	3	9,0	109	3	1	2	5	11,1
66	2	1	3	6	5,8	110	3	1	1	6	6,4
67	2	0	1	6	4,2	111	3	0	3	7	25,7
68	2	1	3	3	6,8	112	3	1	1	4	0,9
69	2	1	2	5	4,8	113	3	1	3	5	3,9
70	2	1	3	5	6,0	114	3	1	1	5	5,1
71	2	1	2	7	9,0	115	3	1	2	6	4,2
72	2	1	1	4	5,3	116	3	1	1	6	4,4
73	2	1	3	4	3,1	117	3	1	1	7	7,9
74	2	0	3	1	6,4	118	3	0	1	4	4,2
75	2	1	1	3	3,9	119	3	0	1	4	3,5
76	2	1	2	3	6,4	120	3	0	2	6	11,4

NOTA: O ponto (.) representa falta de resposta e "Nº" representa o número de ordem da família pesquisada.

Capítulo 5

Dados quantitativos

Quando a variável em estudo for mensurada numericamente, temos grande ganho em termos de técnicas de análise exploratória de dados. Este capítulo trata da construção de distribuições de frequências de variáveis quantitativas, bem como das interpretações que podemos fazer sobre essas distribuições.

Uma variável quantitativa é dita **discreta** quando seus possíveis valores puderem ser listados.

O *número de filhos de um casal* e o *número de cômodos de uma casa* são exemplos de variáveis discretas, pois a primeira só pode assumir valores no conjunto {0, 1, 2,...}, enquanto a segunda no conjunto {1, 2, 3,...}. As variáveis discretas geralmente resultam de alguma contagem.

Uma variável quantitativa é dita **contínua** quando puder assumir qualquer valor num intervalo.

O *peso de um indivíduo* é uma variável contínua, pois pode assumir qualquer valor no intervalo, digamos, de 0 a 300 kg. As variáveis contínuas costumam ser geradas por um instrumento de mensuração.

5.1 Variáveis discretas

A construção de distribuições de frequências de dados de variável discreta pode ser feita da mesma forma que uma distribuição de frequências de dados categorizados, desde que não haja grande quantidade

de diferentes valores observados.[1] Como exemplo, usaremos os dados da variável *número de pessoas residentes no domicílio*, considerando uma amostra de quarenta residências do Conjunto Residencial Monte Verde (anexo do Capítulo 4).

Dados

4	4	4	5	4	1	2	3	6	4	6	4	4	6	3	5	3	4	4	4
5	5	5	4	8	4	5	3	4	5	5	2	5	2	6	8	3	5	5	3

A Tabela 5.1 apresenta a distribuição de frequências desses dados, construída através da contagem das repetições de cada valor.

Tabela 5.1 Distribuição de frequências do número de pessoas residentes no domicílio, numa amostra de quarenta residências do Conjunto Residencial Monte Verde, Florianópolis – SC, 1988.

Número de pessoas	Frequência de residências	Percentagem de residências
1	1	2,5
2	3	7,5
3	6	15,0
4	13	32,5
5	11	27,5
6	4	10,0
7	0	0,0
8	2	5,0

Para representar graficamente a distribuição de frequências de uma variável quantitativa, devemos construir um par de eixos cartesianos. Na abscissa (eixo horizontal) construímos uma escala para representar os valores da variável em estudo, enquanto na ordenada (eixo vertical), representamos a frequência de cada valor.

A Figura 5.1 mostra duas formas alternativas de representação gráfica da distribuição de frequências da Tabela 5.1. A primeira consiste em traçar hastes verticais sobre os valores efetivamente observados (Figura 5.1a). A altura de cada haste deve ser proporcional à frequência do correspondente valor. Na segunda representação, substituímos os riscos por retângulos (Figura 5.1b). Esses retângulos devem ter a mesma largura e podem ser justapostos. O eixo vertical (das frequências) deve sempre iniciar no zero; o eixo horizontal (dos valores da variável) pode iniciar próximo ao menor valor da variável.

[1] Quando a variável apresenta grande número de diferentes valores, podemos usar os artifícios que descreveremos para variáveis contínuas (Seção 5.2).

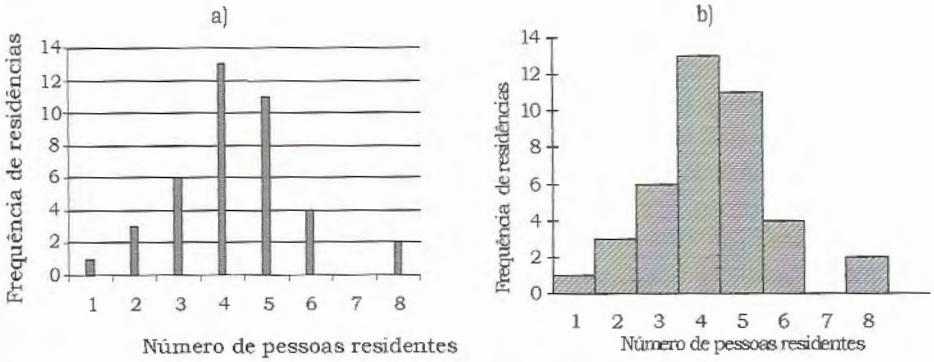


Figura 5.1 Representações gráficas da distribuição de frequências da Tabela 5.1.

Exercícios

1) Observando a Figura 5.1, descreva qual a quantidade típica (ou *faixa típica)* de moradores por domicílio. Existe algum domicílio muito diferente dos demais, em termos do número de moradores?

2) Considerando os dados do anexo do Capítulo 2, faça os seguintes itens:
   a) construa uma tabela de distribuição de frequências para o *nível de satisfação do aluno com o curso* (item 3.g do questionário);
   b) apresente essa distribuição sob forma gráfica e
   c) interprete.

3) As duas tabelas de frequências seguintes referem-se às distribuições do número de filhos dos pais e dos avós maternos de uma amostra de 212 alunos da UFSC, pesquisada pelos alunos do Curso de Ciências Sociais, primeiro semestre de 1990.

Distribuição do número de filhos dos pais dos respondentes

Nº de filhos	1	2	3	4	5	6	7	8	9	10	11	12
Frequência	10	45	32	50	23	23	9	7	6	2	3	2

Distribuição do número de filhos dos avós maternos dos respondentes

Nº de filhos	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Frequência	2	17	32	17	29	23	20	22	21	14	8	6	2	4	0	1	0	1

Apresente essas duas distribuições em gráficos e faça uma descrição comparativa entre elas.

5.2 VARIÁVEIS CONTÍNUAS

Para as variáveis contínuas, não faz muito sentido contar as repetições de cada valor, pois, considerando que dificilmente os valores se repetem, não chegaríamos a um resumo apropriado.

DIAGRAMA DE PONTOS

Quando temos um conjunto com poucos dados, podemos analisá-lo através de um diagrama de pontos, isto é, representando cada resultado (valor) por um ponto na reta de números reais (veja a Figura 5.2).

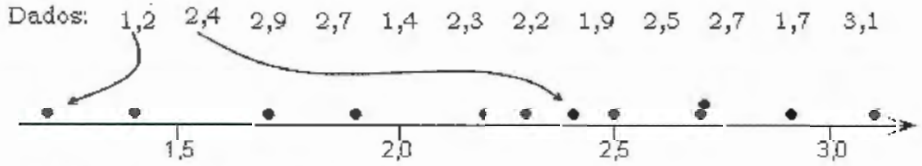


Figura 5.2 Construção de um diagrama de pontos.

É possível colocar duas ou mais distribuições num mesmo gráfico; basta identificar os pontos com símbolos diferentes, ou colocá-los em níveis diferentes, como ilustra a Figura 5.3.

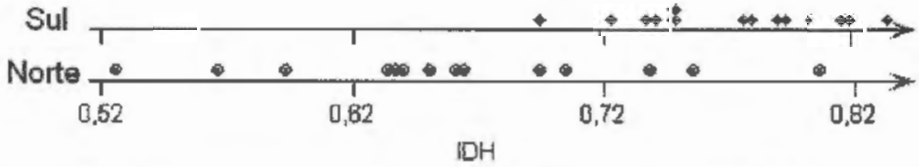


Figura 5.3 Diagrama de pontos do Índice de Desenvolvimento Humano (IDH) de duas amostras aleatórias de quatorze municípios: uma da Região Sul e outra da Região Norte.[2]

Interpretação da Figura 5.3 – Os municípios da amostra da Região Sul apresentam, em geral, IDH maiores do que os municípios da amostra da Região Norte. Também observamos que as duas amostras de municípios diferenciam-se quanto à dispersão dos valores. Enquanto na amostra da Região Sul os municípios apresentam IDH relativamente próximos (maior homogeneidade), na amostra da Região Norte os valores variam bastante de município para município (maior heterogeneidade).

[2] Dados extraídos do Atlas do Desenvolvimento Humano (www.pnud.org.br/atlas). O IDH, calculado para cada município, foi construído com base nos dados do Censo Demográfico de 2000. Observe que neste exemplo os elementos das amostras são municípios.

Tabela de frequências

Nas Ciências Sociais, geralmente trabalhamos com conjuntos de centenas ou milhares de observações, fazendo com que o diagrama de pontos fique impraticável. Podemos construir distribuições de frequências, agrupando resultados em *classes* preestabelecidas. As *classes* são pequenos intervalos mutuamente exclusivos, tais que, quando reunidos, abrangem todo o conjunto de dados. Em outras palavras, as classes devem ser construídas de tal forma que todo valor observado pertença a *uma e apenas uma* classe. Por simplicidade, e para facilitar a interpretação, consideraremos todas as classes com a mesma amplitude.

Como exemplo, usaremos as *taxas de alfabetização* de uma amostra aleatória de quarenta municípios brasileiros.[3]

Dados:

57,25	76,85	92,90	89,07	75,49	84,33	65,28	94,59	71,20	82,30
72,81	66,01	90,52	87,94	58,88	86,34	45,37	81,15	94,83	81,42
54,70	67,95	69,91	95,02	77,62	57,14	91,22	64,65	85,70	81,34
59,07	68,04	73,22	95,34	88,40	83,52	64,19	64,17	95,34	84,66

Observe que todos os valores estão no intervalo de 40 a 100 (o menor valor é 45,37 e o maior é 95,34). Devemos definir um conjunto de classes mutuamente exclusivas, tais que, quando reunidas, contenham todos os valores. Uma possível escolha seria construir 6 (seis) classes com amplitude aproximada de 10 (dez), como segue:

de 40,00 a 49,99; de 50,00 a 59,99; ...; de 90,00 a 99,99

Para simplificar a notação, representaremos essas classes por:

40,00 |— 50,00; 50,00 |— 60,00; ...; 90,00 |— 100,00

sendo que o símbolo "|—" representa o intervalo entre os dois valores, incluindo o valor do lado esquerdo e excluindo o valor do lado direito.

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

classes	contagem	frequência
40 |— 50	|	1
50 |— 60	||||	5
60 |— 70	|||| |||	8
70 |— 80	|||| |	6
80 |— 90	|||| |||| ||	12
90 |— 100	|||| |||	8

[3] Dados do Censo Demográfico, 2000 (www.ibge.gov.br).

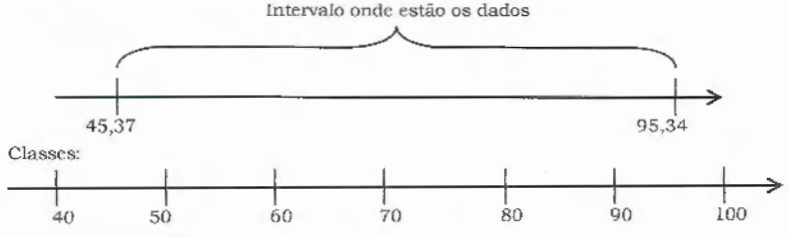
Na apresentação de uma tabela de frequências, é comum colocar também os *pontos médios* das classes, isto é, para cada classe, a média dos seus limites. Por exemplo, na classe 40 |— 50 o ponto médio é 5. O ponto médio representa o *valor típico* da classe. A Tabela 5.2 apresenta a distribuição de frequências dos dados em discussão.

Tabela 5.2 Tabela de frequências de valores da taxa de alfabetização, relativos a uma amostra aleatória de municípios brasileiros, ano 2000.

Classes da taxa de alfabetização	Ponto médio	Frequência de municípios	Percentagem de municípios
40 |— 50	45	1	2,5
50 |— 60	55	5	12,5
60 |— 70	65	8	20,0
70 |— 80	75	6	15,0
80 |— 90	85	12	30,0
90 |— 100	95	8	20,0
Total	–	40	100,0

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

Em geral, são usadas de cinco a vinte classes, dependendo da quantidade de dados e dos objetivos. Dentro desta faixa, uma sugestão é usar, aproximadamente, $\sqrt{n}$ classes, onde n é a quantidade de valores.[4] Em nosso exemplo: $n = 40$, resultando em $\sqrt{n} = 6{,}32$, o que sugere seis ou sete classes; adotamos 6 classes. Como os dados extremos são 45,37 (o menor) e 95,34 (o maior), temos uma amplitude total de $95{,}34 - 45{,}37 \approx 50$. Assim, se as classes iniciarem pelo menor valor, cada classe deve ter amplitude: $50/6 = 8{,}33$. Mas, para facilitar a leitura da tabela de frequências, optamos por iniciar em 40,00 e usar classes com intervalos iguais a 10,00. Esquematicamente:



[4] Quando se têm valores discrepantes no conjunto de dados, recomenda-se que o número de classes seja maior.

Uma forma alternativa de apresentar distribuições de frequências de variáveis quantitativas é através de gráficos, tais como os histogramas e os polígonos de frequências, que apresentaremos a seguir.

Histograma

A Figura 5.4 mostra um histograma de frequências, construído a partir da Tabela 5.2. São retângulos justapostos, feitos sobre as classes da variável em estudo. A altura de cada retângulo é proporcional à frequência observada da correspondente classe.[5]

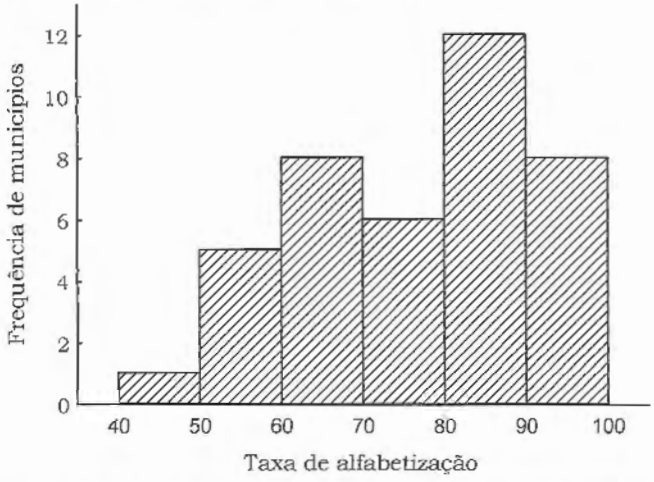


Figura 5.4 Histograma de frequências de valores da taxa de alfabetização, relativos a uma amostra aleatória de municípios brasileiros, ano 2000.

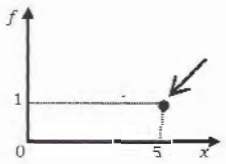
Interpretação da Figura 5.4 – Observamos um contingente razoável de municípios com taxas de alfabetização acima de 80 (dentre a população adulta, mais de 80% de alfabetizados). Mas também há muitos municípios com taxas de alfabetização muito baixa (entre 50 a 80). Uma análise similar por região demográfica poderia trazer mais informações relevantes.

Polígono de Frequências

O polígono de frequências é uma representação gráfica alternativa. Para construí-lo, toma-se o ponto médio (x) e a correspondente frequência (f) de cada classe. Colocamos os pares (x, f) como pontos num par de eixos

[5] Quando as classes não têm a mesma amplitude, é necessário fazer alguns ajustes. Veja, por exemplo, Bussab e Morettin (2002, p. 27). O histograma também poderia ser feito usando percentagens no eixo vertical, mas a sua forma não mudaria.

cartesianos. A ilustração ao lado mostra a representação do ponto (5, 1), num par de eixos cartesianos. Para completar o gráfico, devemos unir os pontos com semirretas, ligando os pontos extremos ao eixo horizontal.



A Figura 5.5 apresenta o polígono de frequências construído a partir da Tabela 5.2. O leitor deve notar que as informações fornecidas pelo polígono de frequências são equivalentes às observadas num histograma.

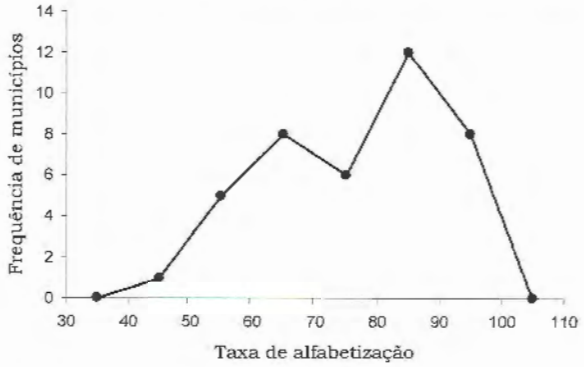


Figura 5.5 Polígono de frequências de valores da taxa de alfabetização, relativos a uma amostra aleatória de municípios brasileiros, ano 2000.

A Figura 5.6 apresenta dois polígonos de frequências num mesmo gráfico, usando dados do anexo do Capítulo 4. O uso de *percentagens* no lugar de *frequências absolutas* foi proposital, porque facilita as comparações entre as duas distribuições de renda. Deixamos para o leitor a interpretação das informações contidas neste gráfico.

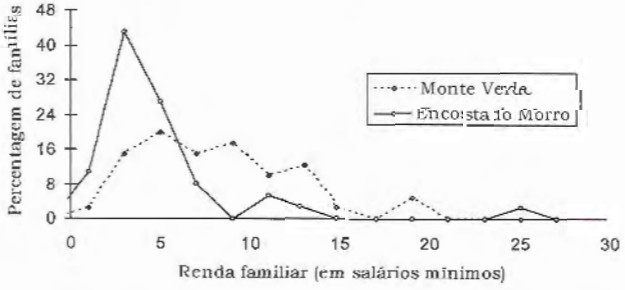


Figura 5.6 Distribuições de frequências das rendas familiares no Monte Verde (amostra de 40 famílias) e na Encosta do Morro (amostra de 37 famílias), Bairro Saco Grande II, Florianópolis-SC, 1988.

O leitor deve observar que um gráfico como o da Figura 5.6 permite explorar possíveis relações entre uma variável quantitativa (*renda*) e uma variável qualitativa (*localidade*). Ao comparar histogramas ou polígonos de frequências, devemos observar a posição no eixo horizontal (nível típico dos valores), a dispersão e a assimetria.

Dizemos que uma distribuição é **simétrica** quando um lado da distribuição é o *reflexo* do outro lado.

É comum medidas físicas terem distribuições razoavelmente simétricas. Por outro lado, distribuições de renda em geral são assimétricas, pois existem mais pessoas com baixa renda do que pessoas com alta renda (*principalmente no Brasil!*). Veja a Figura 5.7.

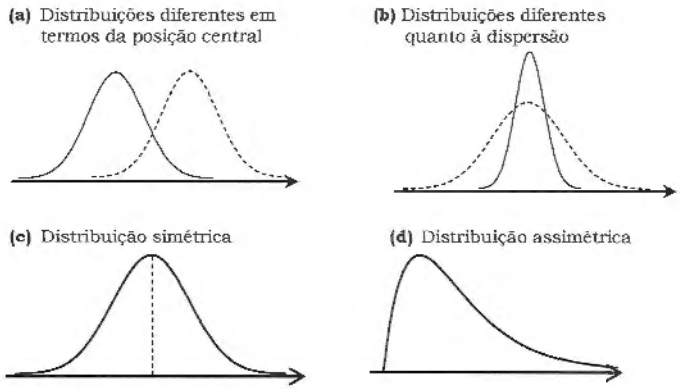


Figura 5.7 Diferentes formas de distribuições de frequências.

Exercícios

4) Os dados a seguir são medidas da *identidade social* que os professores sentem em relação ao seu departamento de ensino. Foram observadas duas amostras de 12 professores: uma no Departamento de Engenharia Mecânica e a outra no Departamento de História, ambas na UFSC. Pelo instrumento utilizado, pode-se dizer que quanto maior o valor, maior é a identificação social do professor com o departamento de ensino a que pertence.

Valores de identidade social

Departamento de Eng. Mecânica	Departamento de História
46 48 47 48 49 50	35 24 43 43 44 33
37 46 47 48 44 47	38 35 39 37 40 35

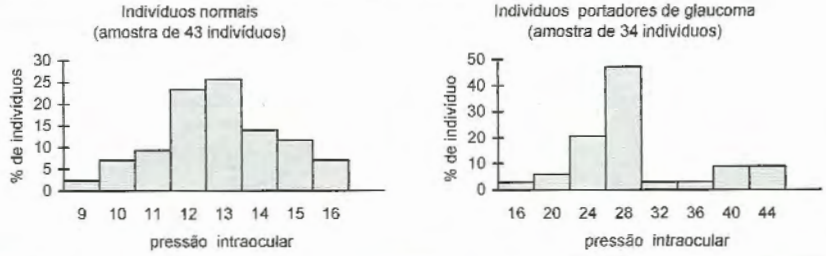
Fonte: Laboratório de Psicologia Social / UFSC, 1990.

Apresente os dois conjuntos de dados num diagrama de pontos e faça uma análise comparativa.

5) Considere os dados do anexo do Capítulo 2.
   a) Construa uma tabela de frequências para o desempenho do aluno no curso (item 5 do questionário).
   b) Faça um histograma. Interprete.
   c) Construa um polígono de frequências.

6) Considerando os dados sobre *renda familiar* do anexo do Capítulo 4, construa três histogramas, sendo um para cada localidade. Faça uma comparação descrevendo as diferenças entre as três distribuições de renda familiar.

7) Os gráficos apresentados a seguir representam distribuições de pressões intraoculares para indivíduos normais e para indivíduos portadores de glaucoma. Quais as semelhanças e diferenças que podemos observar na pressão intraocular desses dois grupos de indivíduos?



5.3 RAMO-E-FOLHAS

Quando a quantidade de dados não for muito grande (digamos, até uma centena de observações), podemos construir, com relativa facilidade, um *ramo-e-folhas*, o qual fornece a forma da distribuição de frequências e ainda preserva a magnitude aproximada dos valores. Num *ramo-e-folhas*, os dados ficam ordenados crescentemente, o que facilita a obtenção de algumas medidas descritivas, como veremos no próximo capítulo.

Voltemos a considerar as taxas de alfabetização de uma amostra de municípios brasileiros. Para facilitar a construção do *ramo-e-folhas*, vamos usar apenas os dois algarismos mais relevantes, desprezando os algarismos decimais.

Para cada valor, o primeiro algarismo é colocado do lado esquerdo do traço vertical, formando os *ramos*. O segundo algarismo é colocado do lado direito do traço, formando as *folhas*. Assim, o valor 57 fica representado por 5 | 7 (veja a segunda linha da Figura 5.8a), o 76 por 7 | 6 (quarta linha), e assim por diante. Na apresentação final de um *ramo-e-folhas*, devemos também ordenar as *folhas*, como mostra a Figura 5.8b.

Dados com os dois algarismos mais relevantes:

57	76	92	89	75	84	65	94	71	82
72	66	90	87	58	86	45	81	94	81
54	67	69	95	77	57	91	64	85	81
59	68	73	95	88	83	64	64	95	84

a)

4	5
5	78479
6	56794844
7	651273
8	942761151834
9	24045155

b)

4	5
5	47789
6	44456789
7	123567
8	111234456789
9	01244555

Figura 5.8 Construção de um *ramo-e-folhas*

O leitor deve notar que, ao observar os dados num *ramo-e-folhas*, vê-se a forma da distribuição de frequências, como se fosse um *histograma deitado* (compare o *ramo-e-folhas* da Figura 5.8b com o histograma da Figura 5.4).

No histograma, temos a liberdade de escolher a amplitude do intervalo de classe; num ramo e folhas, também podemos dividir cada *ramo* em dois ou cinco.[6] Na Figura 5.9, os algarismos (*folhas*) de 0 a 4 ficaram num *ramo* e os algarismos de 5 a 9 no outro *ramo*. A *unidade* indica como devem ser lidos os valores. Em nosso exemplo, temos a unidade igual a 1 (um), ou seja, os valores são lidos naturalmente, emendando o *ramo* com a *folha*. Por exemplo, 4 | 5 é lido como 45.

4	5
5	4
5	7789
6	444
6	56789
7	123
7	567
8	1112344
8	56789
9	01244
9	555

Unidade = 1
4 | 5 = 45

Figura 5.9 Apresentação, em *ramo-e-folhas*, dos valores da taxa de alfabetização, relativos a uma amostra aleatória de municípios brasileiros, ano 2000.

Na construção de um *ramo-e-folhas*, a escolha dos algarismos mais relevantes depende do conjunto de dados em análise. Tomemos um novo exemplo, onde trabalharemos com dois algarismos.

Dados da população residente dos municípios do Oeste Catarinense, 1986.

6.512	8.453	30.592	9.279	105.083	21.083	17.968	25.089	14.857
3.682	19.985	11.133	24.959	12.315	28.339	9.612	12.935	19.739
18.084	13.084	5.464	30.377	26.966	9.094	11.943	21.234	44.183
17.189	9.709	8.713	16.127	3.163	33.245	27.291		

Fonte: IBGE.

[6] Em cada ramo, podemos ter até dez algarismos diferentes. Então, dividindo-se por dois ou cinco, temos a mesma quantidade de algarismos possíveis em cada ramo (cinco e dois, respectivamente).

Ao construir um *ramo-e-folhas* para estes dados, optamos por desprezar os três últimos algarismos, transformando a unidade básica de *habitantes* para *mil habitantes* (veja a Figura 5.10).

```
0 | 33
0 | 56889999
1 | 112234
1 | 677899
2 | 114
2 | 5678
3 | 003
3 |
4 | 4
```

Unidade = 1.000
0|3 = 3.000
Valor discrepante: 10|5

Figura 5.10 Apresentação, em *ramo-e-folhas*, da população residente nos municípios da Microrregião Oeste catarinense, 1986.

EXERCÍCIOS

8) Considerando os dados do anexo do Capítulo 2, construa um *ramo-e-folhas* para os valores do desempenho do aluno no curso. Interprete.

9) Considerando os dados do anexo do Capítulo 4, construa um *ramo-e-folhas* para a *renda familiar*, em cada localidade. Interprete.

EXERCÍCIOS COMPLEMENTARES

10) Foram anotados os tempos decorridos entre a incidência de uma certa doença e sua cura, em 50 pacientes. Estes tempos são os seguintes, em horas:

21	44	27	323	99	90	20	66	39	16
47	96	127	74	82	92	69	43	33	12
41	84	02	61	35	74	02	83	03	13
41	10	24	24	80	87	40	14	82	58
16	35	114	120	67	37	126	31	56	04

Construa um histograma e comente sobre alguns aspectos relevantes desta distribuição.

11) A tabela seguinte apresenta os salários, em reais, dos funcionários de duas empresas.

Empresa A						Empresa B					
400	1200	300	280	700	190	230	420	110	230	330	420
350	620	340	620	550	2100	380	520	190	310	620	380
480	720	310	620	1700	3200	1100	840	210	630	160	240
1800	1320	920	780	1100	510	160	190	200	230	990	355
720	830	400	2900	830	320	3500	230	120	290	340	720
130	190	980	320	1540	920						
420	380	590	1320	2720	3000						

Faça uma descrição comparativa usando gráficos apropriados.

Capítulo 6

Medidas descritivas

Nos dois capítulos anteriores, aprendemos a organizar dados em distribuições de frequências, onde foi possível visualizar como uma variável se distribui, em termos dos elementos observados. Neste capítulo, vamos usar outra estratégia, que pode ser usada de forma alternativa ou complementar, para descrever e explorar *dados quantitativos*.

Quando a variável em estudo é *quantitativa*, podemos resumir certas informações dos dados (valores) por algumas medidas descritivas. Por exemplo, para se conhecer o *peso típico* de recém-nascidos numa comunidade, podemos calcular a *média* ou a *mediana* dos pesos dos recém-nascidos nessa comunidade. Para se ter ideia da magnitude de *variação do peso* dessas crianças, podemos calcular o chamado *desvio padrão*. Em suma, neste capítulo vamos aprender a calcular e interpretar certas medidas que descrevem informações específicas de um conjunto de valores.

Primeiramente, consideraremos a média e o desvio padrão, que são as medidas mais usadas para estudar a posição central e a dispersão. Na Seção 6.3 introduziremos algumas medidas alternativas.

6.1 Média e desvio padrão

A média aritmética

O conceito de *média aritmética*, ou simplesmente *média*, é bastante familiar. Matematicamente, podemos defini-la como a soma dos valores dividida pelo número de valores observados. Por exemplo, dada a nota

final dos oito alunos de uma turma (4, 5, 5, 6, 6, 7, 7 e 8), podemos calcular a média aritmética por:

$$\frac{4+5+5+6+6+7+7+8}{8} = 6$$

De modo geral, dado um conjunto de n valores de uma certa variável X, podemos definir a **média aritmética** por:

$$\overline{X} = \frac{\sum X}{n}$$

onde $\sum X$ representa a soma dos valores da variável X. Em geral, a média aritmética é bastante informativa. Se, por exemplo, na primeira avaliação de uma disciplina, a média das notas dos alunos foi igual a 7,0, e na segunda avaliação foi igual a 9,0, podemos dizer que, em geral, os alunos tiveram melhor aproveitamento na segunda avaliação, mesmo sem nos referirmos às notas de cada aluno individualmente. Mas devemos sempre ter em mente que a média é um resumo dos dados e, por isso, pode esconder informações relevantes.

EXEMPLO 6.1 Vamos considerar a comparação de três turmas de estudantes em termos de suas notas (veja a Tabela 6.1 e Figura 6.1).

Tabela 6.1 Notas finais de três turmas de estudantes e as respectivas médias.

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

Desvio Padrão
A 1,3

B 3,5

C 2,7

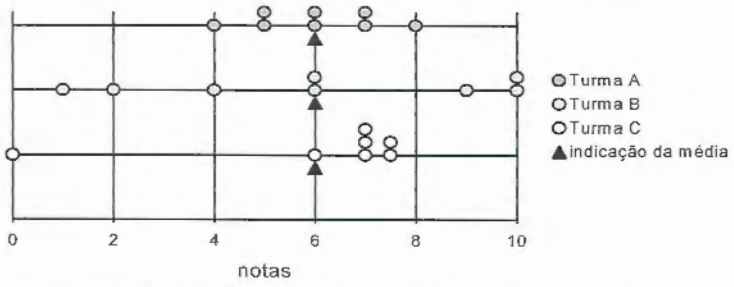


Figura 6.1 Representação das distribuições das notas de três turmas e as correspondentes posições das médias aritméticas.

Observando a Figura 6.1, percebemos que em cada diagrama de pontos a média aritmética representa, num certo sentido, a posição central dos valores. Mais especificamente, podemos dizer que a média aritmética indica o *centro* de um conjunto de valores, considerando o conceito físico de *ponto de equilíbrio* ou *centro de gravidade*. Se imaginarmos os pontos como pesos sobre uma tábua, a *média* é a posição em que um suporte equilibraria a tábua.

A média aritmética *resume* o conjunto de dados em termos de uma *posição central* ou *valor típico*, mas, em geral, não fornece informação sobre outros aspectos da distribuição.

Observamos, na Figura 6.1, que os três conjuntos de valores, apesar de estarem distribuídos sob diferentes formas, apontam para uma mesma média. Comparando as notas da Turma A com as notas da Turma B, verificamos que as notas da Turma B são bem mais *dispersas*, indicando que essa turma é mais heterogênea. Na Turma C, observamos um ponto discrepante dos demais, uma nota extremamente baixa. Com isso, a média fica abaixo da maioria das notas da turma.[1]

Para melhorar o resumo dos dados, podemos apresentar, ao lado da média aritmética, uma medida de dispersão, como a variância ou o desvio padrão.

A variância e o desvio padrão

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

Descrição	notação	resultados numéricos							
Valores (notas dos alunos)	X	4	5	5	6	6	7	7	8
Média	$\overline{X}$				6				
Desvios	$X-\overline{X}$	-2	-1	-1	0	0	1	1	2
Desvios quadráticos	$(X-\overline{X})^2$	4	1	1	0	0	1	1	4

[1] Podemos observar no diagrama de pontos referente à Turma C que a presença de um valor discrepante *arrasta* a média para o seu lado. Assim, a média deixa de representar propriamente um *valor típico* do conjunto de dados. Um tratamento mais adequado para dados que contenham valores discrepantes será visto na Seção 6.3.

Para evitar o problema dos desvios negativos, vamos trabalhar com os desvios quadráticos, $(X - \overline{X})^2$. A variância é definida como a média aritmética dos desvios quadráticos. Por conveniência, vamos calcular esta média, usando como denominador $n-1$ no lugar de n.[2] Assim, definimos a **variância** de um conjunto de valores pela expressão:

$$S^2 = \frac{\sum (X - \overline{X})^2}{n-1}$$

onde $\Sigma (X - \overline{X})$ é a soma dos desvios quadráticos. Em relação ao conjunto de notas da Turma A, a variância é

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

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

$$S = \sqrt{\frac{\sum (X - \overline{X})^2}{n-1}}$$

Em termos do conjunto de notas da Turma A, temos o seguinte desvio padrão: $S = \sqrt{1{,}71} = 1{,}31$.

Ao compararmos os desvios padrões de vários conjuntos de dados, podemos avaliar quais dados se distribuem de forma mais (ou menos) dispersa. O desvio padrão será sempre *não negativo* e será tão maior quanto mais dispersos forem os valores em análise. A Tabela 6.2 mostra o desvio padrão das notas de cada uma das três turmas de alunos, referente aos dados do Exemplo 6.1.

Tabela 6.2 Medidas descritivas das notas finais dos alunos de três turmas.

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

[2] Muitos autores costumam diferenciar a fórmula da variância quando os dados se referem a uma população ou a uma amostra. Quando os dados representam uma população de N elementos, a variância é definida com o denominador N. Quando os dados se referem a uma amostra de n elementos devemos usar o denominador $n-1$. Por simplicidade, vamos considerar sempre o segundo caso.

Ao analisarmos a Tabela 6.2, verificamos, através das médias, que os alunos das três turmas *tenderam* a ter as notas em torno de seis, mas, pelos desvios padrões, concluímos que os alunos da Turma A obtiveram notas relativamente próximas umas das outras, quando comparados aos alunos das outras turmas. Por outro lado, as notas dos alunos da Turma B foram as que se apresentaram mais heterogêneas.[3]

O desvio padrão fornece informação sobre a dispersão (variância ou heterogeneidade) dos valores.

Exercícios

1) Faça os cálculos dos desvios padrões das notas dos alunos das turmas B e C (Tabela 6.1). Verifique se os resultados conferem com os apresentados na Tabela 6.2.
2) Admita que todos os alunos de uma Turma D obtiveram notas iguais a *sete*. Qual o valor da média aritmética? E qual o valor do desvio padrão?
3) A tabela seguinte mostra os resultados dos cálculos das médias e desvios padrões das taxas de crescimento demográfico dos municípios de duas microrregiões catarinenses. Quais as conclusões que você pode tirar desta tabela?

Medidas descritivas das taxas de crescimento demográfico de duas microrregiões de Santa Catarina, 1970-80.

Microrregião	Nº de municípios	Média	Desvio padrão
Serrana	12	-0,36	0,67
Litoral de Itajaí	8	3,55	2,47

6.2 Fórmulas para o cálculo de $\overline{X}$ e S

Ao calcular o desvio padrão nos casos em que a média, $\overline{X}$, acusar um valor fracionário, os desvios, $X - \overline{X}$, acumularão erros de arredondamento, que poderão comprometer o resultado final. Para evitar este inconveniente, podemos usar a seguinte fórmula para o cálculo do

[3] Observe, pela Figura 6.1, que as notas da turma C estão mais concentradas do que as da turma A. Porém, o valor discrepante, além de deslocar a média, aumenta o desvio padrão. Se o valor discrepante fosse desconsiderado, o desvio padrão das notas da turma C seria o menor de todos – a média seria 7 e o desvio padrão 0,55.

desvio padrão, que é matematicamente equivalente àquela apresentada no tópico anterior:

$$S = \sqrt{\frac{\sum X^2 - n\overline{X}^2}{n-1}}$$

onde: $\sum X^2$ é a soma dos valores quadráticos;
$\overline{X}^2$ é a média elevada ao quadrado; e
n é o número de valores.

Ilustraremos o uso desta nova formulação com as notas obtidas pelos alunos da Turma A (Exemplo 6.1).

Valores (notas) X: 4 5 5 6 6 7 7 8 ($\sum X = 48$ e $\overline{X} = 6$)

Valores ao quadrado X^2: 16 25 25 36 36 49 49 64 ($\sum X^2 = 300$)

Assim,

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

Como era de se esperar, chegamos ao mesmo resultado encontrado anteriormente.

PONDERANDO PELAS FREQUÊNCIAS

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

$$\sum X = 4 + 5 + 5 + 6 + 6 + 7 + 7 + 8,$$

que é equivalente a: $4 \times 1 + 5 \times 2 + 6 \times 2 + 7 \times 2 + 8 \times 1 = \sum (X \cdot f)$

onde consideramos apenas os valores distintos de X e ponderamos pelas respectivas frequências, f. Analogamente, podemos calcular a soma quadrática dos valores de X por

$$\sum (X^2 \cdot f) = 4^2 + 5^2 \times 2 + 6^2 \times 2 + 7^2 \times 2 + 8^2 =$$

Com esta nova notação, as formulações de média e desvio padrão são apresentadas a seguir.

$$\overline{X} = \frac{\sum (X \cdot f)}{n} \quad \text{e} \quad S = \sqrt{\frac{\sum (X^2 \cdot f) - n \cdot \overline{X}^2}{n-1}}$$

A Tabela 6.3 mostra a sequência de cálculos para a obtenção da média e do desvio padrão, usando as notas finais dos alunos da Turma A.

Tabela 6.3 Cálculos auxiliares para a obtenção de $\overline{X}$ e S.

Nota X	Frequência *f*	*X·f*	$X^2 \cdot f$
4	1	4	16
5	2	10	50
6	2	12	72
7	2	14	98
8	1	8	64
Total	8	48	300

Assim, $\overline{X} = \frac{48}{8} = 6$ e $S = \sqrt{\frac{300 - 8 \cdot (6)^2}{7}} = 1{,}31$

Os cálculos usando as frequências facilitam bastante quando existirem muitas repetições de valores.

Dados grupados em classes

Quando os dados estão grupados em classes, os cálculos de $\overline{X}$ e S somente poderão ser feitos de forma aproximada, usando o *ponto médio* de cada classe para representar os valores que ocorreram nessa classe (veja Exemplo 6.2).[4]

Exemplo 6.2 Cálculo aproximado de $\overline{X}$ e S dos valores da taxa de alfabetização, relativos a uma amostra aleatória de municípios brasileiros, ano 2000.

Classes da taxa de alfabetização	Ponto médio *X*	Frequência de municípios *f*	*X·f*	$X^2 \cdot f$
40 |— 50	45	1	45	2.025
50 |— 60	55	5	275	15.125
60 |— 70	65	8	520	33.800
70 |— 80	75	6	450	33.750
80 |— 90	85	12	1.020	86.700
90 |— 100	95	8	760	72.200
Total	–	40	3.070	243.600

[4] Ao buscarmos dados em fontes secundárias, muitas vezes já os encontramos grupados em distribuições de frequências, donde os cálculos de $\overline{X}$ e S somente poderão ser feitos de forma aproximada.

Donde:[5]

$$\overline{X} = \frac{3.070}{40} = 76,75 \quad \text{e} \quad S = \sqrt{\frac{243.600 - (40)\cdot(76,75)^2}{n-1}} = 14,30$$

Média ponderada

O cálculo da média e do desvio padrão com ponderação pela frequência é um caso particular de média e desvio padrão ponderados. Em geral, a ponderação é feita sempre que precisamos dar mais importância a um caso do que a outro. Por exemplo, a média aritmética simples dos valores do Índice de Desenvolvimento Humano (IDH) dos municípios da Microrregião da Grande Florianópolis, embora seja um valor central do IDH desses municípios, não corresponde ao IDH da Microrregião, porque temos municípios mais importantes (mais populosos) que outros. Para se ter o IDH da Grande Florianópolis, precisamos ponderar pela população do município, como segue:

Município	População p	IDH X	$X \cdot p$
Antônio Carlos	6.434	0,83	5.320,9
Biguaçu	48.077	0,82	39.327,0
Florianópolis	342.315	0,88	299.525,6
Governador Celso Ramos	11.598	0,79	9.162,4
Palhoça	102.742	0,82	83.837,5
Paulo Lopes	5.924	0,76	4.496,3
Santo Amaro da Imperatriz	15.708	0,84	13.241,8
São José	173.559	0,85	14.7351,6
São Pedro de Alcântara	3.584	0,80	2.849,3
Soma	709.941	7,37	605.112,5

Média simples: $\overline{X} = \frac{\sum X}{n} = \frac{7,37}{9} = 0,82$

Média ponderada: $\overline{X}_p = \frac{\sum (X \cdot p)}{\sum p} = \frac{605.112,5}{709.941} = 0,85$

[5] Se tivéssemos feito os cálculos diretamente com os 40 valores da taxa de alfabetização (ver capítulo anterior), encontraríamos $\overline{X} = 76,89$ e $S = 13,41$.

Exercícios

4) Dado o seguinte conjunto de dados: {7, 8, 6, 10, 5, 9, 4, 12, 7, 8}, calcule:
   a) a média e
   b) o desvio padrão.

5) Calcule a média e o desvio padrão da seguinte distribuição de frequências:

Distribuição de frequências do tamanho da família, numa amostra de 40 famílias do Conjunto Residencial Monte Verde, Florianópolis, SC, 1988.

Tamanho da família	Frequência de famílias	Percentagem de famílias
1	1	2,5
2	3	7,5
3	6	15,0
4	13	32,5
5	11	27,5
6	4	10,0
7	0	0,0
8	2	5,0

6) Faça um histograma para a distribuição de frequências do Exemplo 6.2 e indique o valor da média aritmética no gráfico.

7) Considerando os dados do anexo do Capítulo 2, obtenha a média e o desvio padrão dos valores do índice de desempenho do aluno (item 5 do questionário), considerando:
   a) os dados do anexo do Capítulo 2 (cálculo exato);
   b) a tabela de distribuição de frequências construída no Exercício 5 do capítulo anterior (cálculo aproximado).

8) Sejam os dados do anexo do Capítulo 2.
   a) Calcule as médias e os desvios padrões das respostas dos itens 3(a) a 3(g) do questionário.
   b) Apresente os resultados numa tabela.
   c) Interprete, considerando os objetivos 1 e 3 da pesquisa (Seção 2.4, Capítulo 2).

9) Sejam os dados do anexo do Capítulo 4.
   a) Calcule a renda familiar média em cada uma das três localidades.
   b) Calcule o desvio padrão da renda familiar em cada localidade.
   c) Apresente esses resultados numa tabela.
   d) O que você pode concluir a partir desses resultados?

6.3 Medidas baseadas na ordenação dos dados

A média e o desvio padrão são as medidas mais usadas para avaliar a posição central e a dispersão de um conjunto de valores. Contudo, essas medidas são fortemente influenciadas por *valores discrepantes*. Por

exemplo, nas notas da Turma C (Exemplo 6.1), o valor discrepante 0 (zero) *puxa* a média para baixo, como ilustra a Figura 6.2. Apesar de a média aritmética ser 6 (seis), o diagrama de pontos sugere que o valor 7 (sete) seja um valor *mais típico* para representar as notas da turma, pois, além de ser o valor *mais frequente*, ele é o *valor do meio*, deixando metade das notas abaixo dele e metade acima.

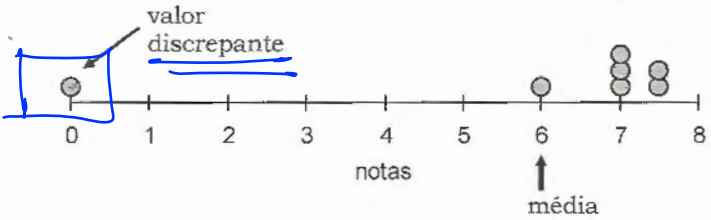


Figura 6.2 A influência de um valor discrepante no cálculo da média aritmética.

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

valores inf. valores sup.

4

A MEDIANA

A mediana avalia o centro de um conjunto de valores, sob o critério de ser o valor que divide a distribuição ao meio, deixando os 50% menores valores de um lado e os 50% maiores valores do outro lado. Por exemplo, o conjunto de valores {2, 3, 4, 5, 8} tem como mediana o valor 4 (quatro), porque a quantidade de valores com magnitude inferior a 4 é a mesma do que a quantidade de valores com magnitude superior a 4. Mais precisamente:

mediana

> Dado um conjunto de n valores, definimos ***mediana*** como o valor, M_d, que ocupa a posição $\frac{n+1}{2}$, considerando os dados ordenados crescente ou decrescentemente. Se $\frac{n+1}{2}$ for fracionário, toma-se como mediana a média dos dois valores de posições mais próximas a $\frac{n+1}{2}$.

EXEMPLOS:

a) Conjunto de notas da Turma C: {0; 6; 7; 7; 7; 7,5 7,5}

→ posição: $\frac{n+1}{2} = 4$ → $M_d = 7$.

b) {5, 3, 2, 8, 4}
Ordenando: 2, 3, 4, 5, 8 → posição: $\frac{n+1}{2} = 3$ → $M_d = 4$.

c) {3, 5, 6, 7, 10, 11} → posição: $\frac{n+1}{2} = 3{,}5$ (3ª e 4ª) → $M_d = \frac{6+7}{2} = 6{,}5$

Comparação entre média e mediana

A Figura 6.3 mostra os valores da média e da mediana num diagrama de pontos. Note que o valor discrepante 62 *puxa* mais a média do que a mediana.

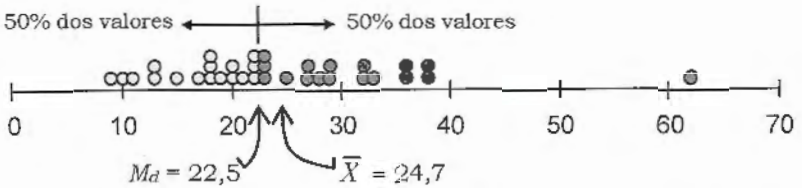


Figura 6.3 Posição da média e da mediana no diagrama de pontos das taxas de mortalidade infantil dos municípios da Microrregião Oeste de Santa Catarina, 1982.

A Figura 6.4 mostra as posições da média e da mediana em distribuições com diferentes formas: uma simétrica e outra assimétrica. No primeiro caso, a média e a mediana são iguais. Em distribuições assimétricas, a média tende a se deslocar para o lado da cauda mais longa.

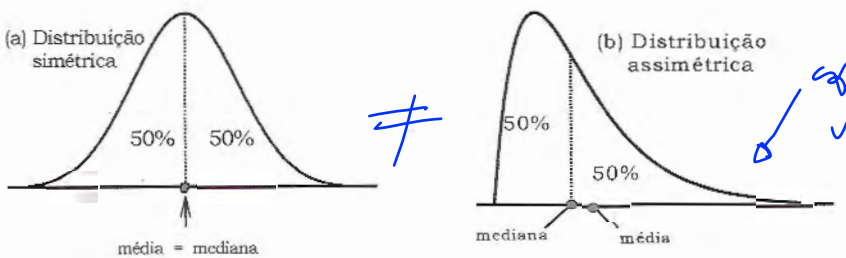


Figura 6.4 Posições da média e mediana, segundo a forma (simétrica ou assimétrica) da distribuição.

Em geral, dado um conjunto de valores, a média é a medida de posição central mais adequada, quando se supõe que estes valores tenham uma distribuição razoavelmente simétrica, enquanto a mediana surge como uma alternativa para representar a posição central em distribuições

então a mediana é para casos muito assimétricos

muito assimétricas.[6] Muitas vezes, calculam-se ambas as medidas para avaliar a posição central sob dois enfoques diferentes, como também para se ter uma primeira avaliação sobre a assimetria da distribuição.

Quartis e extremos

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

Chamamos de *extremo inferior*, E_I, ao menor valor dos dados em análise. De *extremo superior*, E_S, ao maior valor. Por exemplo, dado o conjunto de valores {5, 3, 6, 11, 7}, temos $E_I = 3$ e $E_S = 11$.

Chamamos de *primeiro quartil* ou *quartil inferior*, Q_I, ao valor que delimita os 25% menores valores. De *terceiro quartil* ou *quartil superior*, Q_S, o valor que separa os 25% maiores valores. O *segundo quartil*, ou *quartil do meio*, é a própria mediana, que separa os 50% menores dos 50% maiores valores. Veja a Figura 6.5.

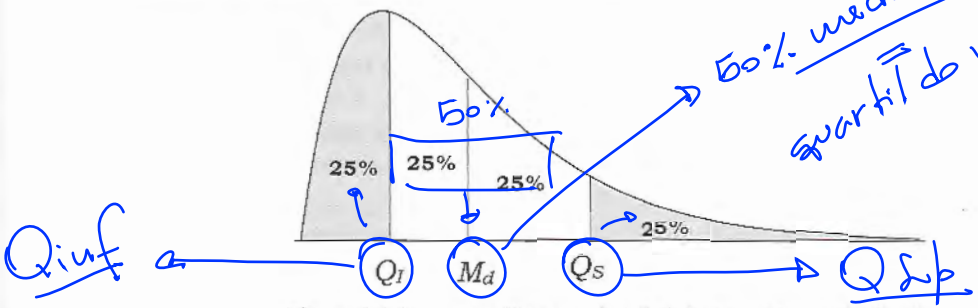


Figura 6.5 Os quartis dividem a distribuição em quatro partes iguais.

Dado um conjunto de valores ordenados, podemos obter, de forma aproximada, o quartil inferior, Q_I, como a mediana dos valores de posições menores ou iguais à posição da mediana. A mediana dos valores de posições maiores ou iguais à posição da mediana corresponde ao quartil

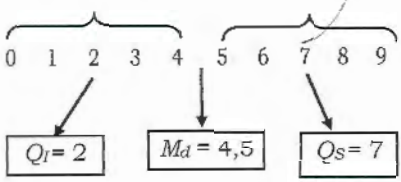
[6] Mesmo para variáveis que supostamente tenham distribuições razoavelmente simétricas, a média e a mediana podem não se igualar, porque, em geral, estamos observando apenas alguns valores (amostras) dessas variáveis. Para variáveis com distribuições razoavelmente simétricas, a média é a medida de posição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

superior, Q_S.[7] Se a mediana coincidir com um valor do conjunto de valores, vamos convencionar em considerá-la tanto no cômputo de Q_I como de Q_S.

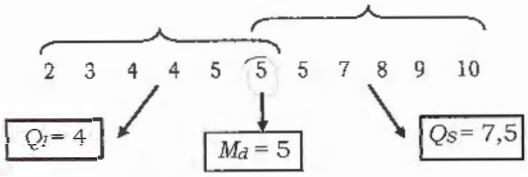
Exemplos:

a) Dados: 2, 0, 5, 7, 9, 1, 3, 4, 6, 8. → é preciso ordenar

Ordenando:



b) Dados (já ordenados):



Exemplo 6.3 Obtenção da mediana num *ramo-e-folhas*: valores referentes às taxas de alfabetização de quarenta municípios brasileiros, ano 2000.[8]

ramos folhas

(1)	4	5
(2)	5	4
(6)	5	7789
(9)	6	444
(14)	6	56789
(17)	7	123
(20)	7	567
(20)	8	1112344
(13)	8	56789
(8)	9	01244
(3)	9	555

Unidade = 1
4|5 = 45

$n = 40$ → posição: $\frac{n+1}{2} = 20,5$ (20ª e 21ª) → $M_d = \frac{77+81}{2} = 79$.

[7] Dado um conjunto de valores, nem sempre conseguimos dividi-lo exatamente em quatro partes iguais. O procedimento exposto oferece uma solução aproximada, mas bastante satisfatória quando a quantidade de valores for grande e com poucas repetições.

[8] No *ramo-e-folhas*, construído na seção 5.7, incluímos uma coluna à esquerda com as frequências acumuladas. Essas frequências foram acumuladas das extremidades até o centro (mediana) da distribuição, o que facilita a contagem das frequências para a obtenção da mediana e quartis.

Para os quartis: $n' = 20$ → posição 10,5 (10ª e 11ª). Daí:

$$Q_I = 65{,}5 \text{ e } Q_S = 87{,}5.$$

Podemos considerar o valor $M_d = 79$ como o *valor típico* das taxas de alfabetização dos quarenta municípios em estudo, pois metade dos municípios acusa taxa de alfabetização inferior a 79 e a outra metade tem níveis mais elevados de alfabetização. Com os quartis, podemos dizer que os 50% dos municípios mais típicos, em termos de alfabetização, acusam taxas variando de 65,5 a 87,5. Podemos dizer, também, que 25% desses municípios têm taxas de alfabetização não superiores a 65,5; enquanto 25% de municípios têm taxas iguais ou superiores a 87,5.

■

ESQUEMA DE CINCO NÚMEROS

O esquema de cinco números é uma forma de apresentação da mediana, quartis e extremos, como mostramos ao lado. Através desses cinco números podemos ter informações sobre a posição central, dispersão e assimetria da distribuição de frequências, como ilustra a Figura 6.6.

$n = 40$

M_d	79	
Q	65,5	87,5
E	45	95

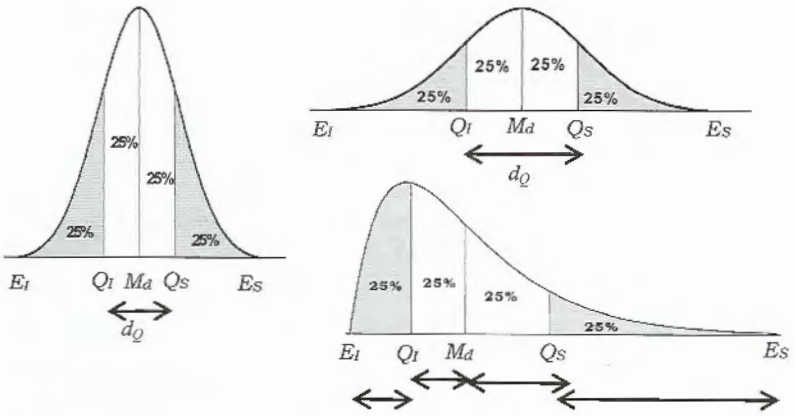


Figura 6.6 Posições da mediana, quartis e extremos em distribuições diferentes quanto à dispersão e assimetria.

O desvio entre quartis, $d_Q = Q_S - Q_I$, é muitas vezes usado como uma medida de dispersão. Veja na Figura 6.6 que, quanto mais dispersa a distribuição, maior será o valor de d_Q. Em distribuições mais dispersas, os valores dos quartis (e dos extremos) ficam mais distantes. Em distribuições simétricas, a distância entre o quartil inferior e a mediana é igual à distância entre a mediana e o quartil superior, enquanto que em distribuições assimétricas isto não acontece.

Uma regra muitas vezes usada para detectar valores discrepantes é verificar se existe algum valor do conjunto de dados que se afasta mais do que $(1,5)\cdot d_Q$ do quartil superior (ou inferior). No Exemplo 6.3, temos:

$$d_Q = Q_S - Q_I = 87,5 - 65,5 = 22$$
$$Q_I - (1,5)\cdot d_Q = 65,5 - (1,5)\cdot(22) = 32,5$$
$$Q_S + (1,5)\cdot d_Q = 87,5 + (1,5)\cdot(22) = 120,5$$

Como nenhum valor está fora do intervalo [32,5; 120,5], não temos valor suspeito de ser discrepante.

Exemplo 6.4 Com o objetivo de comparar as distribuições da renda familiar em duas localidades, construímos um *ramo-e-folhas* e um esquema de cinco números para cada localidade, como mostramos a seguir. Os dados fazem parte do anexo do Capítulo 4.

Renda familiar mensal em quantidade de salários mínimos

Conj. Res. Monte Verde

1	1
2	1446
3	9
4	168
5	11588
6	8
7	12577
8	4469
9	6
10	3349
11	
12	25999
13	
14	
15	4

Unidade = 0,1
1|1 = 1,1

Discrepantes:
18|6 e 19|3

$n = 40$

M_d	7,7	
Q	4,95	10,35
E	1,1	19,3

Encosta do Morro

0	19
1	38
2	123367889
3	599999
4	224569
5	188
6	4
7	19

Discrepantes:
11|1, 11|4, 13|9 e 25|7

$n = 37$

M_d	3,9	
Q	2,7	5,1
E	0,1	25,7

Notamos, inicialmente, que o nível de renda no Conjunto Residencial Monte Verde (mediana de 7,7 salários mínimos) é maior do que na Encosta do Morro (mediana de 3,9 salários mínimos). No Monte Verde, 50% das famílias mais típicas, em termos de renda, estão na faixa de 4,95 a 10,35 salários mínimos mensais; já na Encosta do Morro, as rendas familiares estão na faixa de 2,7 a 5,1 salários mínimos mensais.

A distribuição de renda na Encosta do Morro é mais concentrada em torno de um valor típico. Esta característica pode ser observada pelo desvio entre os quartis, d_Q, que é menor na Encosta do Morro do que no Monte Verde. O desvio entre extremos é maior na Encosta do Morro, mas tal desvio deve ser observado com cautela, pois em ambas as distribuições os extremos superiores são valores discrepantes em relação à maioria dos outros valores.

As duas distribuições são razoavelmente simétricas, quando observadas próximas de suas medianas, pois, em ambas as distribuições, as distâncias entre Q_I e M_d são próximas das distâncias entre M_d e Q_S. Contudo, fora do intervalo entre os quartis temos uma cauda mais longa do lado direito, mostrando que existem algumas poucas famílias com renda relativamente alta em relação ao típico destas localidades. O valor 0,1 salários mínimos, que aparece no extremo inferior da distribuição da Encosta do Morro, apesar de não ser um valor discrepante em termos estatísticos, é um valor estranho de renda familiar. Provavelmente tenha sido coletado erroneamente e deveria passar por uma verificação.

■

Diagrama em caixas

Uma maneira de apresentar aspectos relevantes de uma distribuição de frequências é através do chamado *diagrama em caixas* ou *desenho esquemático*. Traçamos dois retângulos: um representando o espaço entre o quartil inferior e a mediana, e o outro entre a mediana e o quartil superior. Esses retângulos, em conjunto, representam a faixa dos 50% dos valores mais típicos da distribuição. Entre os quartis e os extremos traçamos uma linha. Caso existam valores discrepantes [valores inferiores a $Q_I - 1{,}5 \cdot d_Q$ ou superiores a $Q_S + 1{,}5 \cdot d_Q$], a linha é traçada até o último valor não discrepante; e os valores discrepantes são indicados por pontos (veja a Figura 6.7).

A Figura 6.8 mostra a forma do *diagrama em caixas* para uma distribuição simétrica e para uma distribuição assimétrica. Note as diferenças e imagine como ficaria um *diagrama em caixas* se tivéssemos uma distribuição mais dispersa.

A Figura 6.9 apresenta os *diagramas em caixas* das duas distribuições de renda do Exemplo 6.4. Compare esta representação com os *ramos-e-folhas* vistos anteriormente.

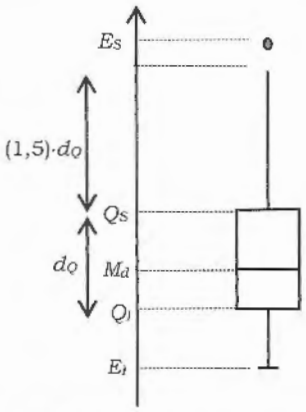


Figura 6.7 Esquema para construção de um diagrama em caixas.

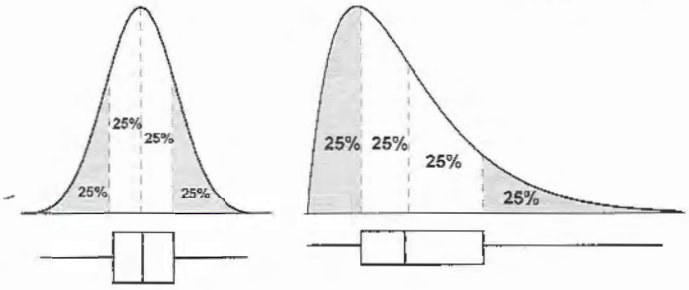


Figura 6.8 *Diagrama em caixas* e a forma da distribuição.

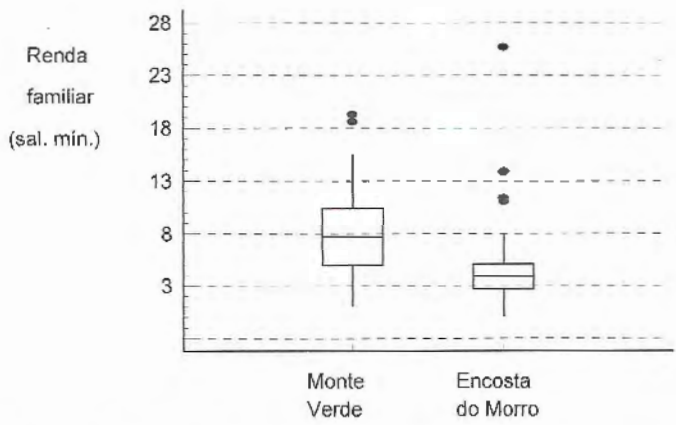


Figura 6.9 Representação em *diagramas em caixas* das distribuições de renda do Exemplo 6.4.

USO DO COMPUTADOR

Em geral, nos pacotes computacionais de estatística, ou mesmo em planilhas eletrônicas, é bastante simples obter um conjunto de medidas descritivas dos valores de uma variável quantitativa. A Figura 6.10 apresenta medidas descritivas da renda, em salários mínimos, de uma amostra de famílias de um bairro de Florianópolis (anexo do Capítulo 4). As medidas descritivas foram obtidas através da planilha eletrônica *Excel®*. Ao lado é apresentado o histograma de frequências para facilitar a interpretação. [9]

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

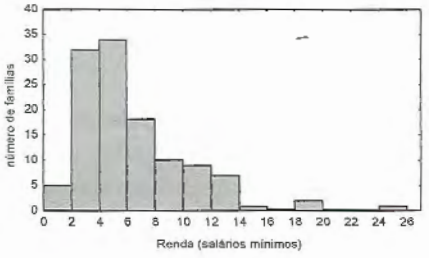


Figura 6.10 Medidas descritivas calculadas com o auxílio do Excel® e um histograma feito com apoio do *STATISTICA®*.

Em termos de posição central, temos a *média*, a *mediana* e a *moda*. Esta última medida apresenta o valor mais frequente do conjunto de dados. O fato de a média apresentar um valor maior que a mediana e a moda sugere uma distribuição assimétrica, com cauda mais longa para o lado direito, o que é confirmado pelo gráfico. Aliás, na lista de medidas, aparece o chamado *coeficiente de assimetria*, com valor igual a 1,73. Em distribuições simétricas esse coeficiente se aproxima de zero. Coeficiente de assimetria positivo (especialmente quando superior à unidade) indica cauda mais longa para o lado direito. Por outro lado, quando negativo (especialmente quando inferior a −1), indica cauda mais longa para o lado esquerdo.

A medida *erro padrão* será apresentada no Capítulo 9. A *curtose* é pouco usada e, por isso, não será discutida neste texto. O *intervalo* ou *amplitude* é outra medida de dispersão, definida como a distância entre os dois valores extremos; e a *contagem* é o número (n) de valores usados no cálculo das medidas descritivas.

[9] Sobre o uso do Excel, ver *Excel.doc* em www.inf.ufsc.br/~barbetta/livro1.htm. O histograma foi construído com o apoio do *STATISTICA®*. Ver www.statsoft.com.br.

6.4 Orientação para análise exploratória de dados

Na análise exploratória de grandes conjuntos de dados, é comum, inicialmente, construirmos uma distribuição de frequências para cada variável, verificando os valores ou categorias típicas, possíveis casos discrepantes, etc. É a descrição ou caracterização dos dados em estudo. Lembramos que a construção da distribuição e a representação gráfica dependem do tipo de variável em estudo, em termos do nível de mensuração (ver Figuras 6.11).

Numa fase seguinte, é comum buscarmos possíveis relações (associações ou correlações) entre as variáveis em estudo. Os procedimentos também dependem do tipo das variáveis (ver Figura 6.12).

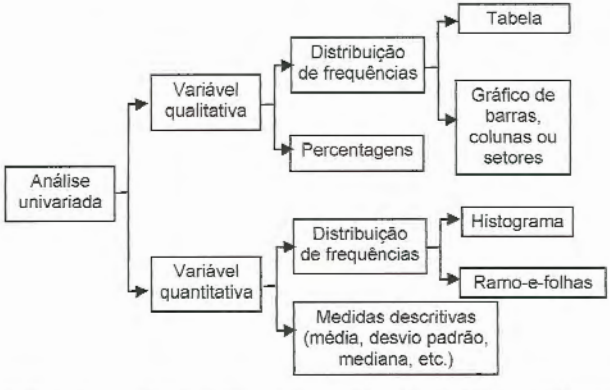


Figura 6.11 Esquema para análise de cada variável individualmente.

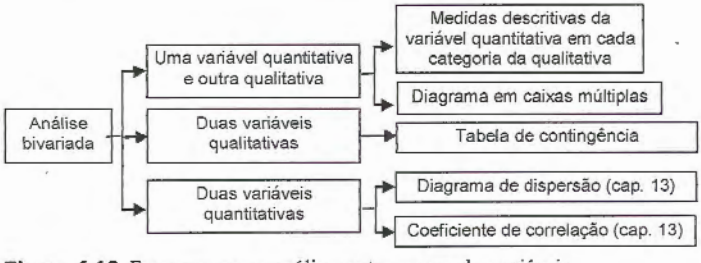


Figura 6.12 Esquema para análise entre pares de variáveis.

Exercícios

10) Calcule a mediana e os quartis dos seguintes dados:

a) {15, 9, 7, 20, 18, 19, 23, 32,14, 10, 11}

b) {15, 9, 7, 20, 18, 19, 23, 32,14, 10, 11, 16}

11) Obtenha a mediana e os quartis da distribuição de frequências do Exercício 5 (Seção 6.2).

12) Considere o anexo do Capítulo 2:
   a) Obtenha a mediana, os quartis e os extremos dos valores do índice de desempenho do aluno (item 5 do questionário) e interprete. Sugestão: apresente, inicialmente, os dados num *ramo-e-folhas*.
   b) Comparando o valor da mediana com o valor que você obteve para a média aritmética no Exercício 7 (igual a 2,311), o que você diria sobre a simetria da distribuição desses valores?

13) A tabela abaixo mostra a distribuição de frequências do número de filhos dos pais de alunos da UFSC, considerando uma amostra de 212 estudantes, entrevistados pelos alunos do Curso de Ciências Sociais, UFSC, 1990. Obtenha os extremos, a mediana e os quartis.

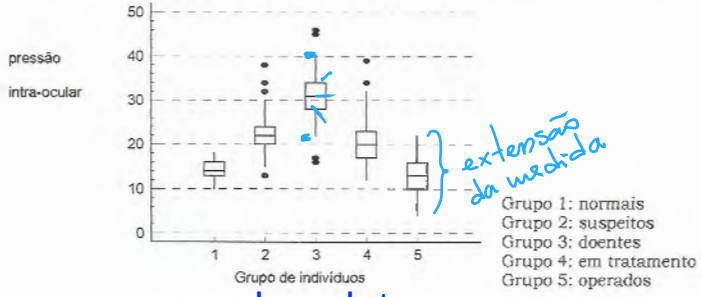
Nº de filhos	1	2	3	4	5	6	7	8	9	10	11	12
frequência	10	45	32	50	23	23	9	7	6	2	3	2

14) A tabela seguinte é composta de medidas descritivas, calculadas a partir de quatro conjuntos de valores, oriundos de uma amostra de 212 estudantes da UFSC. Os estudantes foram indagados acerca do número de filhos que planejam ter, do número de filhos de seus pais, do número de filhos de seus avós maternos e do número de filhos de seus avós paternos.

Medidas descritivas	número de filhos			
	planejados	dos pais	dos avós maternos	dos avós paternos
média	2,06	4,23	6,35	6,15
desvio padrão	1,26	2,29	3,21	3,12
extremo inferior	0	1	1	1
quartil inferior	1	2	4	4
mediana	2	4	6	6
quartil superior	2	5	8	8
extremo superior	12	12	18	16

Faça uma redação comparando os quatro conjuntos de valores, tomando por base as medidas descritivas apresentadas na tabela.

15) A figura seguinte apresenta cinco distribuições de frequências representadas por *diagramas em caixas*. São dados de pressão intraocular de uma amostra de 243 indivíduos, divididos em cinco grupos, segundo a condição clínica da doença *glaucoma*. Descreva as principais informações oriundas desta análise:

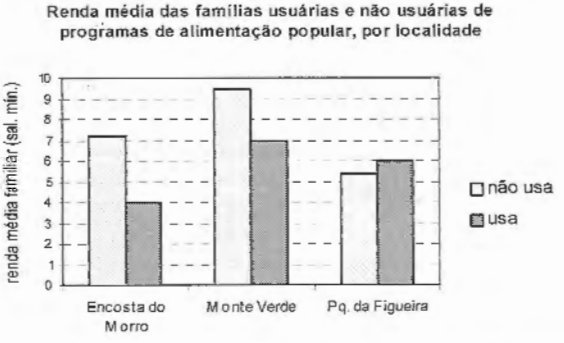


boxplot

Exercícios complementares

16) No Exemplo 6.2, calculamos a média aritmética da taxa de alfabetização de uma amostra de municípios brasileiros. Se esses municípios fossem os municípios de uma Unidade da Federação, o valor da média (76,75) poderia ser interpretado como a taxa de alfabetização dessa Unidade da Federação? Explique.

17) O gráfico seguinte foi construído com o auxílio da planilha *Excel*, a partir dos dados do anexo do Capítulo 4. Interprete.



18) Com o objetivo de comparar a distribuição da renda familiar em duas cidades, levantou-se a renda familiar de cada população e calcularam-se algumas medidas descritivas, apresentadas na tabela abaixo.

Medidas descritivas da renda familiar, em quantidade de salários mínimos, em duas cidades.

Cidade	média	desvio padrão	quartil inferior	mediana	quartil superior
A	4,8	3,2	3,4	4,9	6,5
B	4,9	6,2	3,0	3,8	9,0

Descreva um texto observando as principais informações verificadas nos dados da tabela.

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

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

a) Apresente esses dados em *ramo-e-folhas*.

b) Na empresa BBB, a distância (em km) até a residência dos seus 300 funcionários apresenta as seguintes medidas descritivas:

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

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

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

20) Apresentamos, abaixo, algumas medidas descritivas da distribuição de salários, em R$, de três empresas de um certo ramo.

Empresa	média	desvio padrão	extremo inferior	quartil inferior	mediana	quartil superior	extremo superior
A	300	100	100	200	302	400	510
B	400	180	100	250	398	550	720
C	420	350	100	230	300	650	10.000

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

21) Dada a tabela abaixo, compare os quatro departamentos da UFSC quanto aos escores de identidade social com o departamento. Quanto maior o escore, identidade social mais elevada.

Medidas descritivas do nível identidade social com o departamento.

Depto	Tamanho da amostra	Média	Mediana	Desvio padrão
Eng. Mecânica	40	46,9	47,0	2,1
Arquitetura	24	40,8	42,5	5,9
Psicologia	19	42,5	44,0	5,4
História	21	38,4	39,0	5,4

Fonte: Laboratório de Psicologia Social (Depto de Psicologia/UFSC).

Parte III

Modelos de probabilidade

Como usar modelos de probabilidade para entender melhor os fenômenos aleatórios

Capítulo 7

Modelos probabilísticos

Nos capítulos anteriores, procuramos entender uma variável estudando o comportamento de um conjunto de observações (amostra). Desta forma, estudamos a distribuição de frequências do uso (*sim* ou *não*) de programas de alimentação popular, com base numa amostra de famílias da região de interesse (Capítulo 4). Nessa abordagem, predomina o raciocínio indutivo: com base na organização e descrição de dados observados, procuramos fazer conjeturas sobre o universo (população) em estudo.

Neste capítulo, faremos o raciocínio de forma inversa, em que procuraremos entender como poderão ocorrer os resultados de uma variável, considerando certas suposições a respeito do problema em estudo (raciocínio dedutivo). Exemplo: supondo que 60% das famílias do bairro usam programas de alimentação popular, o que se pode deduzir sobre a percentagem de famílias que usam esses programas, numa amostra aleatória simples de dez famílias?

A resposta a esta indagação não é um simples número, pois, dependendo das dez famílias selecionadas na amostra, teremos resultados diferentes. Para responder adequadamente, precisamos apresentar quais são os possíveis resultados e como eles poderão ocorrer. Essa descrição é feita em termos dos chamados *modelos probabilísticos*.

A Figura 7.1 faz um paralelo entre modelos probabilísticos e um método de análise exploratória de dados, em termos do tipo de raciocínio.

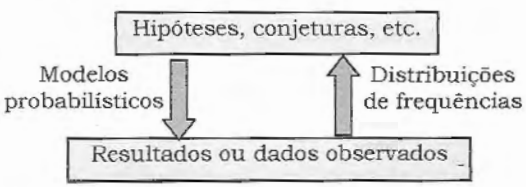


Figura 7.1 Distribuições de frequências e modelos probabilísticos.

7.1 DEFINIÇÕES BÁSICAS

Os *modelos probabilísticos* são construídos a partir de certas hipóteses ou conjeturas sobre o problema em questão e constituem-se de duas partes: (1) dos possíveis resultados e (2) de uma certa *lei* que nos diz quão provável é cada resultado (ou grupos de resultados).

Seja o experimento de lançar uma moeda e observar a face voltada para cima. Os possíveis resultados são *cara* e *coroa*. Se supusermos que a moeda é perfeitamente equilibrada, e se o lançamento for imparcial, podemos também dizer que a *probabilidade* de ocorrer *cara* é a mesma de ocorrer *coroa*.

ESPAÇO AMOSTRAL E EVENTOS

Seja um experimento aleatório, isto é, uma experiência ou situação em que deve ocorrer um, dentre vários resultados possíveis.

Espaço amostral é o conjunto de *todos* os resultados possíveis do experimento e será denotado por Ω.

EXEMPLO 7.1

a) Lançar uma moeda e observar a face voltada para cima. Temos, neste caso, dois resultados possíveis: *cara* e *coroa*. Então, o espaço amostral é o conjunto $\Omega = \{cara, coroa\}$.

b) Lançar um dado e observar o número de pontos marcado no lado voltado para cima. Temos: $\Omega = \{1, 2, 3, 4, 5, 6\}$.

c) Numa urna com bolas azuis e vermelhas, extrair uma bola e observar sua cor. Temos: $\Omega = \{azul, vermelha\}$.

d) Num certo bairro, indagar a uma família se ela costuma utilizar-se de algum programa de alimentação popular. Um possível espaço amostral para esta situação é $\Omega = \{sim, não\}$. Considerando, porém, a possibilidade do respondente não saber ou se negar a responder, podemos ser levados a tomar um espaço amostral mais amplo: $\Omega^* = \{sim, não, não\ resposta\}$.

e) Num certo bairro, selecionar uma amostra de dez famílias e verificar quantas utilizaram algum programa de alimentação popular nos últimos dois meses. Um espaço amostral adequado é $\Omega = \{0, 1, 2, ..., 10\}$.

f) Numa escola de ensino fundamental, selecionar uma criança e medir a sua altura. Como *altura* é uma variável *contínua*, o espaço amostral precisa ser construído como um conjunto de números reais possíveis, tal como $\Omega = \{x$, tal que $x \in \Re$ e $0 < x < 2{,}00$ m$\}$.

■

Ressaltamos que a especificação do espaço amostral pode não ser única, porque depende daquilo que estamos observando e de algumas considerações sobre o problema. Veja, por exemplo, o item (d).

Um espaço amostral é **discreto** quando podemos *listar* os possíveis resultados. É **contínuo** quando temos uma infinidade de resultados possíveis dentro de um intervalo de números reais.

No Exemplo 7.1, nos itens de (a) a (e) temos espaços amostrais discretos; já no item (f), temos um espaço amostral contínuo.

Evento é um conjunto de resultados do experimento.[1]

Por exemplo, no lançamento de um dado, podemos ter interesse nos seguintes eventos:

A = ocorrer um número par;
B = ocorrer um número menor que três;
C = ocorrer o ponto seis; e
D = ocorrer um ponto maior que seis.

Em termos de notação de conjunto, temos: $A = \{2, 4, 6\}$, $B = \{1, 2\}$, $C = \{6\}$ e $D = \{\ \}$. Repare que o último caso é um evento impossível e, por isso, é representado pelo conjunto vazio.

Vejamos, agora, a segunda parte de um modelo probabilístico: a alocação de probabilidades aos resultados possíveis.

Probabilidades

Probabilidade é um valor entre 0 (zero) e 1 (um). A soma das probabilidades de todos os resultados possíveis do experimento deve ser igual a 1 (um).

[1] Em linguagem matemática, podemos dizer que A é um evento se e somente se A é um subconjunto do espaço amostral Ω, pois Ω é o conjunto de <u>todos</u> os resultados possíveis.

EXEMPLO 7.1 (CONTINUAÇÃO) Vamos apresentar os modelos probabilísticos para alguns experimentos aleatórios, alocando, de forma intuitiva, a probabilidade de cada resultado do espaço amostral. O princípio que norteia a alocação dessas probabilidades será apresentado posteriormente.

a) No lançamento de uma moeda, se considerarmos a moeda perfeitamente equilibrada e lançamento imparcial, os resultados tornam-se equiprováveis. Assim, podemos alocar probabilidade 0,5 tanto para *cara* como para *coroa*, resultando no seguinte modelo probabilístico:

Resultado	Probabilidade
cara	0,5
coroa	0,5

b) No lançamento de um dado, se considerarmos o dado perfeitamente equilibrado e o lançamento imparcial, tem-se o seguinte modelo probabilístico:

Resultado	1	2	3	4	5	6
Probabilidade	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$

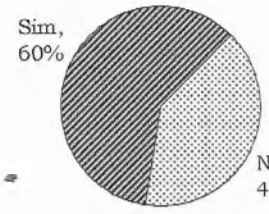
c) Na seleção de uma bola de uma urna, para construirmos um modelo para a cor da bola a ser extraída, precisamos conhecer a quantidade (ou a percentagem) de bolas de cada cor. Se tiverem sete bolas azuis e três vermelhas e, ainda, supusermos que a bola seja extraída *aleatoriamente*, temos o seguinte modelo:[2]

Resultado	Probabilidade
azul	0,7
vermelha	0,3

d) No problema de verificar se uma família de um bairro costuma utilizar programas de alimentação popular, vamos supor, por simplicidade, a inexistência de *não resposta*, ou seja, qualquer que seja a família selecionada, as possíveis respostas devem estar em $\Omega = \{sim, não\}$. Como no caso anterior, é necessário o conhecimento da distribuição desta característica na população. Vamos supor que em todo o bairro 60% das famílias utilizam e 40% não utilizam programas de alimentação popular. Se a família for selecionada aleatoriamente, podemos explicitar o modelo probabilístico, como mostra o esquema seguinte.

[2] Usaremos frequentemente o termo *seleção aleatória* para uma seleção que garanta que todos os elementos tenham a mesma probabilidade de serem selecionados. No caso de bolas numa urna, a seleção aleatória pode ser equivalente a uma seleção *ao acaso*, desde que todas as bolas tenham o mesmo tamanho e que estejam bem misturadas.

População de famílias dividida quanto ao uso de programas de alimentação popular (*sim* ou *não*).



sorteio de uma família

Modelo de probabilidades para o resultado (*sim* ou *não*) de uma família extraída ao acaso e indagada sobre a utilização de programas de alimentação popular.

Resultado	Probabilidade
sim	0,6
não	0,4

■

Para alocar probabilidades, podemos lançar mão do princípio da equiprobabilidade. Por exemplo, no problema da urna (Exemplo 7.1c), podemos fazer o seguinte raciocínio: como a seleção é aleatória, toda bola da urna tem a mesma probabilidade de ser selecionada. Como têm 7 bolas azuis dentre as 10 bolas da urna, a probabilidade de selecionar uma bola azul é $^7/_{10}$ (ou 0,7). Analogamente, a probabilidade de selecionar uma bola vermelha é $^3/_{10}$ (ou 0,3). O princípio da equiprobabilidade é usualmente enunciado em termos da probabilidade de algum evento, como apresentamos a seguir.

Princípio da equiprobabilidade: quando as características do experimento sugerem N resultados possíveis, todos com igual probabilidade de ocorrência, a probabilidade de um certo evento A, contendo N_A resultados, pode ser definida por:

$$P(A) = \frac{N_A}{N}$$

Usando este princípio, vamos alocar probabilidades aos seguintes eventos, baseados num lançamento imparcial de um dado perfeitamente equilibrado.

Evento	Probabilidade
A = *ocorrer um número par*	$P(A) = 3/6 = 1/2$ ou 0,5
B = *ocorrer um número menor que três*	$P(B) = 2/6 = 1/3$
C = *ocorrer o ponto seis*	$P(C) = 1/6$
D = *ocorrer um ponto maior que seis*	$P(D) = 0/6 = 0$

Uma forma mais geral de alocar probabilidades a eventos é somando as probabilidades dos resultados que compõem o evento. No exemplo do dado:

$$P(\textit{ocorrer um número par}) = 1/6 + 1/6 + 1/6 = 1/2$$

Este procedimento pode ser usado mesmo quando os resultados não são equiprováveis.[3]

EXEMPLO 7.2 Seja uma urna com 5 bolas brancas, 3 vermelhas e 2 pretas. Selecionar uma bola ao acaso. Qual a probabilidade da bola selecionada ser branca ou vermelha?

P(*branca* ou *vermelha*) = P(*branca*) + P(*vermelha*) = $^5/_{10} + ^3/_{10} = ^8/_{10}$ (ou 0,8).

Também chegaríamos a este resultado se lembrássemos que a soma de todos os resultados possíveis é igual a 1. Assim,

P(*branca*) + P(*vermelha*) + P(*preta*) = 1, ou:

P(*branca* ou *vermelha*) = 1 – P(*preta*) = $1 - ^2/_{10} = ^8/_{10}$.

■

Dois eventos são **independentes** quando a ocorrência de um deles não altera a probabilidade da ocorrência do outro.

Por exemplo, no lançamento imparcial de um dado e de uma moeda, os eventos *A* = *número par no dado* e *B* = *cara na moeda* podem ser admitidos como independentes, já que a ocorrência de *A* (ou de *B*) nada tem a ver com a ocorrência de *B* (ou de *A*).

Quando a ocorrência de um evento puder ser interpretada como resultante da ocorrência simultânea de dois outros eventos independentes, sua probabilidade pode ser obtida pelo *produto* das probabilidades individuais desses eventos.

EXEMPLO 7.3 Lançar duas vezes, de forma imparcial e independente, um dado perfeitamente equilibrado. Calcular a probabilidade de ocorrer número par em ambos os lançamentos.

P(número par em ambos os lançamentos) =

= P(nº par no 1º lançamento) × P(nº par no 2º lançamento) =

= $(^1/_2) \times (^1/_2) = ^1/_4$

■

[3] Estamos supondo que os resultados de um experimento são mutuamente exclusivos, ou seja, ao realizar o experimento vai ocorrer somente um resultado.

Ensaios de Bernoulli

Os *ensaios de Bernoulli* ocorrem em situações onde observamos apenas um elemento e verificamos se este tem (ou não) um certo atributo.

Exemplo 7.4 São exemplos de ensaios de Bernoulli:

a) Numa urna com bolas brancas e pretas, extrair, aleatoriamente, uma bola da urna e observar se é de cor branca.

b) Observar, ao acaso, um morador da cidade e verificar se ele é favorável a um certo projeto municipal. Admita que todos os moradores têm opinião formada.[4]

c) Lançar uma moeda e observar se ocorreu *cara*.

d) Lançar um dado e observar se ocorreu o ponto *seis*.[5]

e) Selecionar, aleatoriamente, um eleitor numa certa cidade e verificar se ele pretende votar em determinado candidato à prefeitura. Admita que todos os eleitores desta cidade já tenham definido seu voto.

f) Selecionar, aleatoriamente, uma peça que está saindo de uma linha de produção e verificar se ela é *defeituosa*.

■

Em todos esses casos o espaço amostral pode ser Ω = {*sim*, *não*}. Sob certas suposições a respeito do experimento e supondo conhecida a distribuição de *sim* e *não* na população, podemos especificar o modelo probabilístico.

Exemplo 7.4 (continuação)

b) Se admitirmos que 70% dos moradores são favoráveis ao projeto, temos o seguinte modelo probabilístico:

Resultado	sim (concorda)	não (discorda)
Probabilidade	0,7	0,3

c) Se admitirmos que o dado é perfeitamente equilibrado, e o lançamento imparcial, temos:

Resultado	sim (ponto 6)	não (outro ponto)
Probabilidade	1/6	5/6

■

[4] Na prática, é difícil supor que todos os moradores tenham opinião formada. Pode-se contornar este problema restringindo o estudo àqueles que tenham a opinião formada, descartando os indecisos.

[5] Neste exemplo, temos seis resultados possíveis, mas, considerando que o interesse é somente no ponto *seis*, podemos restringir o espaço amostral a Ω = {*seis*, *não seis*}.

Muitas vezes não conhecemos informações suficientes para especificar completamente o modelo probabilístico. No item (b), por exemplo, podemos não conhecer a percentagem de favoráveis na população. Nesse caso podemos apresentar apenas o *jeitão* do modelo, como segue:

Resultado	Probabilidade
sim	π
não	$1 - \pi$

onde π é um valor (desconhecido) entre 0 e 1. Por exemplo, se a probabilidade de *sim* é $\pi = 0{,}7$, então a probabilidade de *não* é $1 - \pi = 0{,}3$.

Chamamos de **parâmetro** a uma quantidade desconhecida do modelo, que se tornaria conhecida se tivéssemos informações adicionais sobre a população de onde está sendo tirada a amostra (ou sobre o fenômeno em que se está tirando algumas observações).

O número π, do modelo anterior, corresponde ao parâmetro *proporção de favoráveis ao projeto na população.*

VARIÁVEL ALEATÓRIA

Variável aleatória é uma característica numérica associada aos resultados de um experimento.[6]

Exemplo: X = número de caras em três lançamentos de uma moeda;

Y = percentagem de pessoas favoráveis a um projeto municipal, numa amostra de 500 moradores da cidade.

Podemos caracterizar um ensaio de Bernoulli por uma variável aleatória X, definida da seguinte forma:

$$X = \begin{cases} 0, & \text{se } \textit{não} \\ 1, & \text{se } \textit{sim} \end{cases}$$

e o modelo de probabilidade:

x	1	0
p(x)	π	$1 - \pi$

[6] Formalmente, *variável aleatória* é definida como uma *função*, que associa resultados do espaço amostral, Ω, ao conjunto de números reais.

onde: π é uma quantidade entre 0 e 1 (*parâmetro* do modelo);
x é um possível valor de X (no caso, 0 ou 1); e
$p(x)$ é a probabilidade de ocorrer o valor x. Assim, $p(0) = 1 - \pi$ é a probabilidade de *não* e $p(1) = \pi$ é a probabilidade de *sim*.

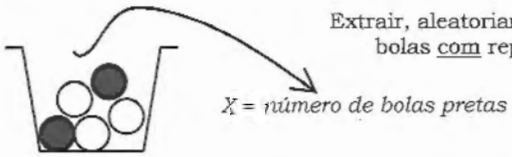
Um modelo probabilístico, quando apresentado em termos de uma variável aleatória, também é chamado de *distribuição de probabilidades*.

Dois ensaios de Bernoulli

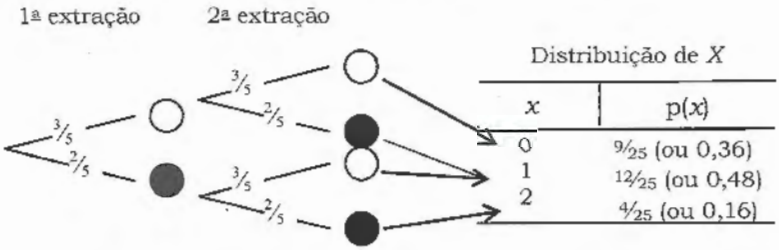
Quando temos dois ensaios de Bernoulli, geralmente o interesse está na variável aleatória:

X = número de ocorrências de sim nos dois ensaios.

Exemplo 7.5 Seja uma urna com três bolas brancas e duas pretas. Extrair, aleatoriamente, duas bolas, sendo uma após a outra, tal que repomos na urna a primeira bola antes de extrairmos a segunda – amostragem com reposição.



O esquema, a seguir, mostra a construção da distribuição de probabilidades de X = *número de bolas pretas extraídas na amostra.*



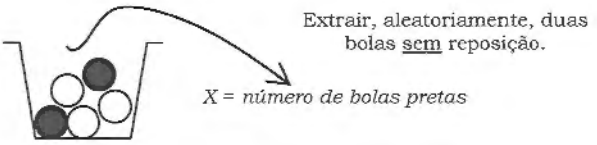
Probabilidade de $X = 0$: calcula-se a probabilidade de ocorrer *bola branca na 1ª extração* e *bola branca na 2ª extração*, ou seja, $(^3/_5)\cdot(^3/_5) = {}^9/_{25}$.

Probabilidade de $X = 2$: de forma análoga, $(^2/_5)\cdot(^2/_5) = {}^4/_{25}$.

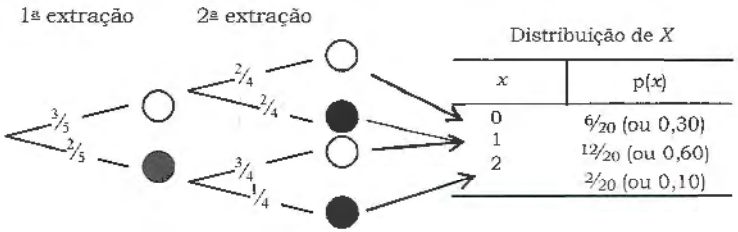
Probabilidade de $X = 1$:
bola branca na 1ª e *bola preta na 2ª* (com probabilidade $(^3/_5)\cdot(^2/_5) = {}^6/_{25}$) ou
bola preta na 1ª e *bola branca na 2ª* (com probabilidade $(^2/_5)\cdot(^3/_5) = {}^6/_{25}$).
Logo, a probabilidade de $X = 1$ é ${}^6/_{25} + {}^6/_{25} = {}^{12}/_{25}$.

■

EXEMPLO 7.6 Idem ao exemplo anterior, mas sem repor a primeira bola na segunda extração – amostragem *sem* reposição.



A configuração da urna na segunda extração depende do que aconteceu na primeira extração. Assim, o resultado da primeira extração condiciona as probabilidades da segunda extração.



Distribuição de X

x	p(x)
0	6/20 (ou 0,30)
1	12/20 (ou 0,60)
2	2/20 (ou 0,10)

■

Quando a amostragem é feita *com reposição*, como no Exemplo 7.5, há *independência* entre os ensaios, pois os resultados de um ensaio não alteram as probabilidades de outros. Isto não acontece quando a amostragem é feita *sem reposição*, como no Exemplo 7.6, onde os resultados de uma extração *dependem* do que ocorreu nas extrações anteriores.

Se compararmos as distribuições de probabilidades dos Exemplos 7.5 e 7.6, notamos que o efeito da *dependência* entre os ensaios provoca uma grande alteração na distribuição de probabilidades. Contudo, se o leitor refizer esses cálculos, considerando um grande número de bolas (digamos, 2.000 bolas brancas e 3.000 bolas pretas), as distribuições de probabilidades dos dois casos (com e sem reposição) serão praticamente as mesmas.

Em *grandes populações* podemos supor independência entre os ensaios mesmo que a amostragem seja feita *sem* reposição.[7]

[7] Como referência, vamos considerar a *população grande* quando o tamanho desta superar em vinte vezes o tamanho da amostra ($N > 20n$).

Exercícios

1) Numa urna com 10 bolas numeradas de 1 a 10, extrair, aleatoriamente, uma bola e observar o seu número.
   a) Construa um modelo probabilístico.
   b) Liste os resultados contidos nos eventos: *A = número par*, *B = número ímpar* e *C = número menor que 3*.
   c) Atribua probabilidades aos eventos do item (b).

2) Numa sala com 10 homens e 20 mulheres, sorteia-se um indivíduo, observando o sexo (masculino ou feminino). Construa um modelo probabilístico.

3) Numa eleição para prefeitura de uma cidade, 30% dos eleitores pretendem votar no Candidato A, 50% no Candidato B e 20% em branco ou nulo. Sorteia-se um eleitor na cidade e verifica-se o candidato de sua preferência.
   a) Apresente um modelo probabilístico.
   b) Qual é a probabilidade de o eleitor sorteado votar num dos dois candidatos?

4) Seja uma família sorteada de uma população de 120 famílias, as quais se distribuem conforme a seguinte tabela.

Distribuição conjunta de frequências do nível de instrução do chefe da casa e uso de programas de alimentação popular, num conjunto de 120 famílias.

Uso de programas	Nível de instrução do chefe da casa			Total
	nenhum	fundamental	médio	
sim	31	22	25	78
não	7	16	19	42
Total	38	38	44	120

Calcule a probabilidade de a família sorteada ser:
   a) usuária de programas de alimentação popular;
   b) tal que o chefe da casa tenha o nível médio;
   c) tal que o chefe da casa não tenha o nível médio
   d) usuária de programas de alimentação popular, e o chefe da casa ter o nível médio;
   e) usuária de programas de alimentação popular, e o chefe da casa não ter o nível médio;
   f) usuária de programas de alimentação popular, considerando que o sorteio tenha sido restrito às famílias cujo chefe da casa tenha o nível médio;
   g) tal que o chefe da casa tenha o nível médio, considerando que o sorteio tenha sido restrito às famílias usuárias de programas de alimentação popular.

5) Seja a população descrita no Exercício 4. Selecionam-se, aleatoriamente, duas famílias, sendo uma após a outra, repondo a primeira família selecionada antes de proceder a segunda seleção (amostragem *com* reposição). Qual é a probabilidade de que ambas as famílias sejam usuárias de programas de alimentação popular?

7.2 O MODELO BINOMIAL: CARACTERIZAÇÃO E USO DA TABELA

Nesta seção, vamos caracterizar um tipo de modelo probabilístico que se presta a diversas situações práticas, em especial às situações em que observamos a presença (ou ausência) de algum atributo. O interesse é no número ou na percentagem de elementos que têm o atributo, numa amostra de n elementos.

CARACTERIZAÇÃO DE UM EXPERIMENTO BINOMIAL

Um experimento é dito binomial, quando:

a) consiste de n ensaios;
b) cada ensaio tem apenas dois resultados de interesse: *sim* ou *não*; e
c) os ensaios são *independentes*, com probabilidade *constante* π de ocorrer *sim* $(0 < \pi < 1)$.

Vamos estudar a distribuição de probabilidades da variável aleatória

X = *número de ocorrências de sim nos n ensaios,*

conhecida como *distribuição binomial.* As quantidades n e π são os *parâmetros* da distribuição cujos valores dependem das características do problema que se está modelando.

No Exemplo 7.5, a variável aleatória X = *número de bolas pretas obtidas nas duas extrações* tem distribuição binomial de parâmetros: $n = 2$ (pois, estamos extraindo duas bolas) e $\pi = {}^{2}/_{5}$ (pois, a probabilidade de sair bola preta numa particular extração é ${}^{2}/_{5}$). No Exemplo 7.6 não temos um experimento binomial, pois não há *independência* entre os ensaios.

EXEMPLO 7.7 São exemplos de experimentos binomiais:

a) O número Y de caras, em três lançamentos imparciais de uma moeda perfeitamente equilibrada. Valores dos parâmetros: $n = 3$ e $\pi = 0{,}5$.
b) Dentre uma grande população de pessoas, em que 70% são favoráveis a um projeto municipal, o número X de favoráveis, numa amostra aleatória de dez pessoas. Parâmetros: $n = 10$ e $\pi = 0{,}7$.
c) O número F de eleitores, que se declaram a favor de um certo candidato, numa amostra de 3.000 eleitores, extraída aleatoriamente de uma população de 100.000 eleitores. Parâmetros: $n = 3.000$ e π = *proporção de eleitores favoráveis ao candidato na população.*

■

A tabela da distribuição binomial

Para conhecermos as probabilidades de uma variável com distribuição binomial, podemos fazer uso da Tabela 2 do apêndice (Tabela da distribuição binomial).[8]

Exemplo 7.8 Retornemos ao problema de extrair, aleatoriamente e com reposição, duas bolas de uma urna, que contém duas bolas pretas e três brancas. Seja X o número de bolas pretas extraídas.

Inicialmente, verificamos pelas características do problema que $n = 2$ e $\pi = 2/5 = 0{,}40$. Entrando com estes valores na tabela da distribuição binomial, como indica o esquema ao lado, encontramos a mesma distribuição de probabilidades que havíamos desenvolvido no Exemplo 7.5.

n	x	π				
		0,05	...	0,40	...	0,95
...	...	...		...		...
2	0	...		0,3600		...
	1	...		0,4800		...
	2	...		0,1600		...

■

Exemplo 7.9 Seja a população de pessoas de um município em que 70% são favoráveis a um certo projeto municipal. Qual é a probabilidade de que, numa amostra aleatória simples de 10 pessoas dessa população, a maioria seja favorável ao projeto?

Note que temos um experimento binomial, com $n = 10$ e $\pi = 0{,}70$. Usando a *tabela da distribuição binomial*, podemos especificar a distribuição de X = *número de favoráveis na amostra*. A probabilidade de ocorrer o evento *a maioria da amostra ser favorável*, corresponde, em termos da variável aleatória X, ao evento $X > 5$, como ilustramos ao lado. A probabilidade deste evento será a *soma dos resultados individuais*, ou seja:

$$P(X > 5) =$$
$$= p(6) + p(7) + p(8) + p(9) + p(10) =$$
$$= 0{,}2001 + 0{,}2668 + 0{,}2335 + 0{,}1211 + 0{,}0282 =$$
$$= 0{,}8497.$$

Parte da Tabela 2

n	x	π = 0,70
10	0	0,0000
	1	0,0001
	2	0,0014
	3	0,0090
	4	0,0368
	5	0,1029
	6	0,2001
	7	0,2668
$X > 5$	8	0,2335
	9	0,1211
	10	0,0282

■

[8] A Tabela 2 fornece as probabilidades para experimentos com até 15 ensaios. Uma fórmula geral para o cálculo dessas probabilidades será apresentada na próxima seção. Para experimentos compostos de muitos ensaios (n grande), podemos usar a distribuição normal, a qual será estudada no próximo capítulo.

Uma distribuição de probabilidades também pode ser apresentada sob forma gráfica, de maneira análoga às distribuições de frequências, substituindo o eixo das frequências por probabilidades. A Figura 7.2 mostra gráficos típicos para variáveis aleatórias discretas, como é o caso da binomial.

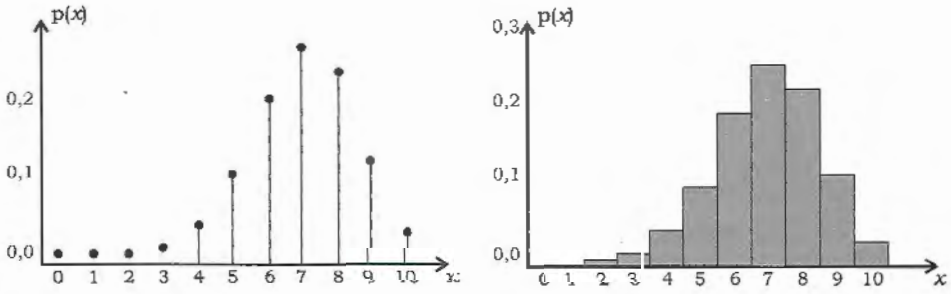


Figura 7.2 Representações gráficas da distribuição binomial com $n = 10$ e $\pi = 0{,}7$ (Exemplo 7.7b).

EXERCÍCIOS

6) Dos experimentos abaixo, verificar quais são binomiais, identificando, quando possível, os valores dos parâmetros n e π. Para aqueles que não são binomiais, apontar as razões.
   a) De uma sala com cinco mulheres e três homens, selecionar, aleatoriamente e com reposição, três pessoas. A variável aleatória de interesse é o número de mulheres selecionadas na amostra.
   b) Idem (a), mas considerando a amostragem *sem reposição*.
   c) De uma população de milhares de homens e mulheres, selecionar aleatoriamente e sem reposição, vinte pessoas. O interesse está no número de mulheres na amostra.
   d) Selecionar uma amostra aleatória simples de 500 pessoas no Estado de Santa Catarina. O interesse está no número de favoráveis à mudança da capital do município de Florianópolis para o município de Curitibanos.
   e) Selecionar, aleatoriamente, um morador de cada município de Santa Catarina. A variável aleatória de interesse é a mesma do item anterior.
   f) Observar uma amostra aleatória simples de 100 crianças recém-nascidas em Santa Catarina. O interesse é verificar quantas nasceram com menos de 2 kg.
   g) Observar uma amostra aleatória simples de 100 crianças recém-nascidas em Santa Catarina. A variável aleatória em questão é o peso, em kg, de cada criança da amostra.

7) Lançar, de forma imparcial, uma moeda perfeitamente equilibrada, cinco vezes. Calcule as seguintes probabilidades:
   a) ocorrer exatamente três caras;
   b) ocorrer 60% ou mais de caras, isto é, $P(X \geq 3)$, onde X é o número de caras.

8) Considere o experimento do exercício anterior, porém com dez lançamentos. Qual é a probabilidade de se obter 60% ou mais de caras? Intuitivamente você esperava que esta probabilidade fosse menor do que a do Exercício 7? Por quê?

9) Seja uma população em que 40% são favoráveis e 60% são contrárias a um projeto. Apresente a distribuição de probabilidades de X = *número de favoráveis numa amostra aleatória de* $n = 5$ *moradores.*

10) Construa um gráfico para a distribuição de probabilidades do exercício anterior.

11) Com respeito ao Exercício 9, calcule a probabilidade de a amostra acusar:
   a) dois ou mais favoráveis, ou seja, $P(X \geq 2)$;
   b) menos de dois favoráveis, ou seja, $P(X < 2)$;
   c) mais de 50% de favoráveis.

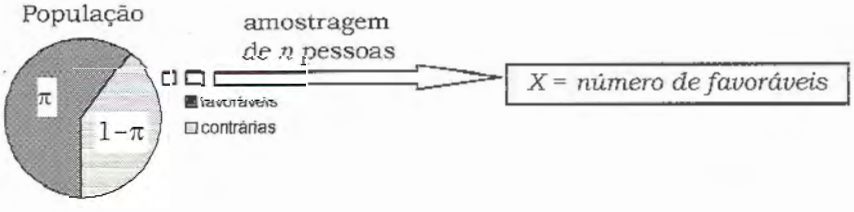
12) Considerando o Exercício 9, construa a distribuição de probabilidades da variável aleatória $P = X/5$ (proporção de indivíduos favoráveis, na amostra).

13) Sob a hipótese de que um certo programa de treinamento melhora o rendimento de 80% das pessoas a ele submetidas, qual é a probabilidade de, numa amostra de sete pessoas que sejam submetidas a esse programa de treinamento,
   a) exatamente cinco melhorarem de rendimento?
   b) menos de a metade melhorar de rendimento?

14) Um certo processo industrial pode, no máximo, produzir 10% de itens defeituosos. Uma amostra aleatória de 10 itens acusou 3 defeituosos. Calcule a probabilidade de ocorrerem, numa amostra de tamanho $n = 10$, três ou mais itens defeituosos, supondo que o processo esteja sob controle (digamos, com $\pi = 0{,}10$, onde π é a probabilidade de cada particular item sair defeituoso).

7.3 O modelo binomial: formulação matemática

Seja X o número de pessoas favoráveis a um certo projeto municipal, numa amostra aleatória simples de n pessoas, extraída de uma população em que a proporção de favoráveis é igual a π. Admitindo que o tamanho da população seja bastante superior ao tamanho da amostra, podemos supor que a variável aleatória X tenha distribuição binomial, com parâmetros n e π. Veja esquema a seguir:



Para cada uma das pessoas indagadas a respeito do projeto, vamos representar por S a resposta *sim* (favorável) e por N a resposta *não* (contrária). A Figura 7.3 apresenta as possíveis combinações de respostas S e N, numa amostra de $n = 4$ pessoas. Esta figura também mostra os valores da variável aleatória X e suas respectivas probabilidades.

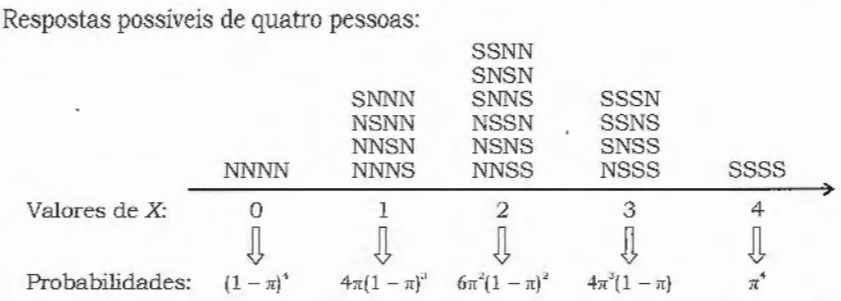


Figura 7.3 Possíveis sequências de respostas e construção de uma distribuição binomial de probabilidades com $n = 4$ e π genérico.

O evento $X = 0$ ocorre quando são sorteadas quatro pessoas contrárias ao projeto (NNNN), cuja probabilidade é $(1 - \pi)\cdot(1 - \pi)\cdot(1 - \pi)\cdot(1 - \pi) = (1 - \pi)^4$.

O evento $X = 1$ ocorre quando forem observadas três pessoas contrárias e uma favorável, em qualquer ordem (SNNN, NSNN, NNSN ou NNNS). Como cada um destes resultados tem probabilidade $\pi\cdot(1 - \pi)^3$, a probabilidade do evento $X = 1$ é $4\cdot\pi\cdot(1 - \pi)^3$. As outras probabilidades podem ser obtidas de forma análoga.

Coeficientes binomiais

No cálculo da probabilidade do evento $X = 1$, contamos quatro maneiras diferentes de aparecer uma resposta S nos n ensaios (SNNN, NSNN, NNSN e NNNS). Em geral, para calcular a probabilidade do evento $X = x$ da distribuição binomial, onde x é um valor possível da variável aleatória X, precisamos calcular o número de maneiras em que podemos combinar as x respostas S dentre as n respostas. Esse número, conhecido como coeficiente binomial, pode ser obtido na *Tabela dos coeficientes binomiais* (Tabela 3 do apêndice), ou calculado pela seguinte expressão:

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

onde $n! = n(n-1)(n-2)...1$ (lê-se *n fatorial*) e, por convenção, $0! = 1$. Por exemplo, para $n = 4$ temos os seguintes coeficientes binomiais:

$x = 0$: $\binom{4}{0} = \frac{4!}{4!0!} = \frac{4!}{4!} = 1$

$x = 1$: $\binom{4}{1} = \frac{4!}{3!1!} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1 \cdot 1} = 4$

$x = 2$: $\binom{4}{2} = \frac{4!}{2!2!} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 2 \cdot 1} = 6$

$x = 3$: $\binom{4}{3} = \frac{4!}{1!3!} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{1 \cdot 3 \cdot 2 \cdot 1} = 4$

$x = 4$: $\binom{4}{4} = \frac{4!}{0!4!} = \frac{4!}{4!} = 1$

Expressão geral da distribuição binomial

Para generalizar o raciocínio que fizemos ao obter as probabilidades na Figura 7.3, considere X uma variável aleatória com distribuição binomial de parâmetros n e π (sendo $0 < \pi < 1$). A probabilidade de X assumir um certo valor x, pertencente ao conjunto $\{0, 1, 2, \ldots, n\}$, é dada pela expressão:

$$p(x) = \binom{n}{x} \cdot \pi^x \cdot (1-\pi)^{n-x}$$

Exemplo 7.10 Seja a população de pessoas de um município em que 70% são favoráveis a um certo projeto municipal. Qual é a probabilidade de, numa amostra aleatória simples de quatro pessoas desta população, encontrarmos exatamente três pessoas favoráveis ao projeto?

Solução: X tem distribuição binomial com parâmetros $n = 4$ e $\pi = 0{,}7$. Então, a probabilidade pedida é dada por:

$$p(3) = \binom{4}{3} \cdot (0{,}7)^3 \cdot (0{,}3)^1 = 4 \cdot (0{,}7)^3 \cdot (0{,}3) = 0{,}4116$$

Se o leitor procurar na tabela da distribuição binomial (Tabela 2 do apêndice), deve encontrar o mesmo resultado.

Exercícios

15) Refazer o Exercício 9, sem usar a tabela da distribuição binomial.

16) (BUSSAB; MORETTIN, 2002, p.122) Uma companhia de seguros vendeu apólices a cinco pessoas, todas da mesma idade e com boa saúde. De acordo com as tábuas atuariais, a probabilidade de que uma pessoa daquela idade esteja viva daqui a 30 anos é de $^2/_3$. Calcular a probabilidade de que, daqui a 30 anos:
   a) exatamente duas pessoas estejam vivas;
   b) todas as pessoas estejam vivas;

c) pelo menos 3 pessoas estejam vivas.

Indique as suposições necessárias para a aplicação do modelo binomial.

17) Dentre sessenta alunos do Curso de Ciências da Computação da UFSC, observamos que quatro estavam plenamente satisfeitos com o curso que estavam realizando (anexo do Capítulo 2). Se fizermos cinco sorteios com reposição dessa população, encontre a probabilidade de:

a) nenhuma resposta "plenamente satisfeito";
b) a maioria "plenamente satisfeito";
c) pelo menos um "plenamente satisfeito".

Exercícios complementares

18) De uma sala com quatro homens e duas mulheres, selecionar, ao acaso e sem reposição, duas pessoas. Qual é a probabilidade de se obter exatamente uma mulher?

19) Uma sala contém vinte mulheres e oitenta homens. Se forem feitos seis sorteios, um após o outro e com reposição, qual é a probabilidade de que se observe:

a) cinco ou mais homens?
b) exatamente duas mulheres?
c) pelo menos uma mulher?

20) Numa população onde 32% dos indivíduos têm alguma descendência indígena, retira-se uma amostra aleatória de seis pessoas. Qual é a probabilidade de se encontrar

a) exatamente duas pessoas com descendência indígena?
b) mais de uma pessoa com descendência indígena?

21) Suponha que 10% dos clientes que compram a crédito em uma loja deixem de pagar regularmente as suas contas (prestações). Se num particular dia, a loja vende a crédito para dez pessoas, qual a probabilidade de que:

a) exatamente uma deixe de pagar?
b) mais de 20% delas deixem de pagar?

Suponha que as dez pessoas que fizeram crediário nesse dia correspondam a uma amostra aleatória de clientes potenciais dessa loja.

22) Admitamos igualdade de probabilidade para o nascimento de menino e menina. De todas as famílias com seis filhos:

a) que proporção tem três meninos e três meninas?
b) que proporção tem quatro ou mais meninas?

23) Um exame de múltipla escolha consiste em dez questões, cada uma com quatro possibilidades de escolha. A aprovação exige, no mínimo, 50% de acertos. Qual é a probabilidade de aprovação se o candidato comparece ao exame sem saber absolutamente nada, apelando apenas para o "palpite"?

Capítulo 8

Distribuições contínuas e modelo normal

Neste capítulo estudaremos o modelo de probabilidades mais conhecido da Estatística: a chamada *distribuição normal de probabilidade*. Diversas aplicações deste modelo estarão presentes ao longo dos demais capítulos. Para podermos estudá-la, vamos inicialmente estender o conceito de equiprobabilidade para variáveis aleatórias contínuas.

Dizemos que uma variável aleatória é **contínua** quando não conseguimos enumerar seus possíveis resultados, por esses formarem um conjunto infinito, num dado intervalo de números reais.

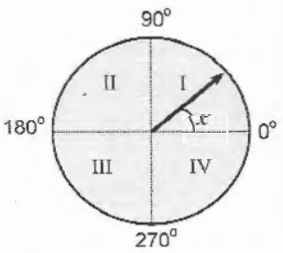
Por exemplo, a altura de um indivíduo, tomado ao acaso, é uma variável aleatória contínua, pois não é possível enumerar todos os valores possíveis de altura de indivíduos, mas podemos dizer, por exemplo, que o resultado será um número real do intervalo de zero a dois metros e meio, o qual contém infinitos números.

Distribuições contínuas

Em variáveis aleatórias contínuas, não existe interesse em atribuir probabilidade a cada particular valor, mas sim, para eventos formados por intervalos de valores. Ao observar a altura de um indivíduo, não importa a probabilidade de ele medir 1,682333... metros; mas o interesse pode estar na probabilidade de ele ter altura no intervalo *de 1,60 a 1,80 m*; ou *acima de 1,90 m*; e assim por diante.

A especificação da distribuição de probabilidades de uma variável aleatória contínua é realizada por um modelo matemático que permite calcular probabilidades em qualquer intervalo de números reais. O Exemplo 8.1 ilustra a construção de um modelo para uma variável aleatória contínua.

EXEMPLO 8.1 Considere um círculo, com medidas de ângulos, em graus, a partir de uma determinada origem, como mostra a figura ao lado. Neste círculo, tem um ponteiro que é colocado a girar no sentido anti-horário.



Seja X a variável aleatória que indica o ponto em que o ponteiro para de girar. Como existem infinitos pontos no intervalo de 0 a 360°, esta variável aleatória é contínua. Vejamos, inicialmente, a probabilidade de o ponteiro parar no quadrante I, isto é, a probabilidade de X assumir um valor entre 0 e 90°.

Supondo que não exista região de preferência para o ponteiro parar, podemos deduzir, pelo *princípio da equiprobabilidade*, que as probabilidades de parada são iguais para os quatro quadrantes. Assim, a probabilidade de o ponteiro parar no primeiro quadrante deve ser igual a $^1/_4$.

Podemos representar o evento *ponteiro parar no quadrante I* por $0 \leq X < 90$; e esta probabilidade por $P(0 \leq X < 90)$. Em termos de variáveis aleatórias contínuas, os sinais "$<$" e "$\leq$" são equivalentes, pois, considerando a equiprobabilidade de todos os pontos e a existência de infinitos pontos, podemos definir a probabilidade de ocorrência de um particular ponto como nula.

■

A distribuição de probabilidades de uma variável aleatória contínua pode ser representada por uma função não negativa, com a área entre o eixo-X e a curva igual a 1 (um). Os eventos podem ser representados por intervalos no eixo-X, enquanto as probabilidades, pelas correspondentes áreas sob a curva (ver Figura 8.1).

A função descrita na Figura 8.1a é uma constante no intervalo de 0 a 360°, porque o experimento sugere que todos os intervalos de mesmo tamanho devem ser igualmente prováveis. Para que a área total seja igual à unidade, a constante deve ser $^1/_{360}$.[1] Construída a distribuição, qualquer

[1] A área de um retângulo é dada por *base* × *altura*. Como a base é 360 e a área é 1, então a altura tem que ser $^1/_{360}$.

probabilidade associada à variável X pode ser obtida pelo cálculo de certa área. Neste contexto, a Figura 8.1b ilustra a probabilidade de o ponteiro parar no quadrante I, que é igual a: $90 \cdot \frac{1}{360} = \frac{1}{4}$.

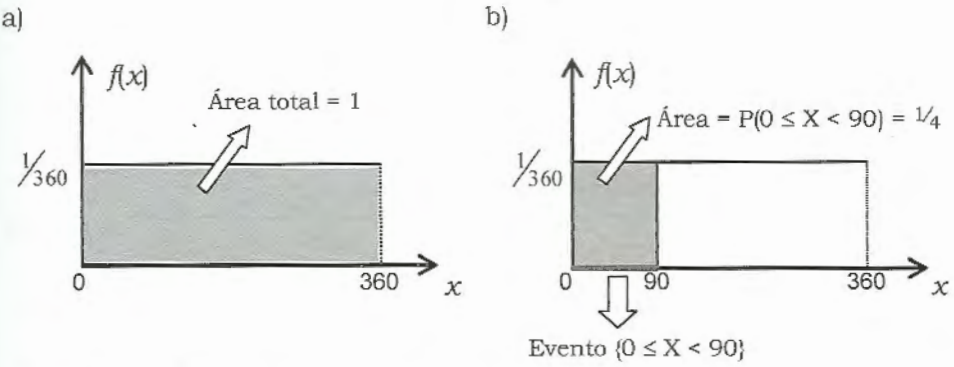


Figura 8.1 Ilustração de: (a) uma distribuição de probabilidades para a variável aleatória do Exemplo 8.1; e (b) a probabilidade do evento $\{0 \leq X < 90\}$.

Exemplo 8.2 Selecionar, aleatoriamente, de uma certa universidade, um estudante do sexo masculino. Seja X o valor de sua altura, em centímetros.

Temos, novamente, uma variável aleatória contínua, mas, desta vez, não é razoável atribuir a mesma probabilidade para diferentes faixas de altura. Por exemplo, é intuitivo que a probabilidade do estudante ter altura entre 165 e 175 cm seja bem maior do que entre 190 e 200 cm, mesmo que ambos os intervalos tenham a mesma amplitude.

A Figura 8.2a sugere um modelo mais adequado para a presente situação. Por este modelo, conhecido como *distribuição normal de probabilidades*, existe um *valor típico*, ou *valor médio*, que no caso de alturas de homens adultos, deve estar em torno de 170 cm. Intervalos em torno deste valor médio têm altas probabilidades de ocorrência, mas as probabilidades diminuem na medida em que nos afastamos deste valor médio, indiferentemente se do lado esquerdo (para valores menores) ou do lado direito (para valores maiores). A Figura 8.2b identifica a probabilidade do evento *o estudante sorteado ter mais de 180 cm.*

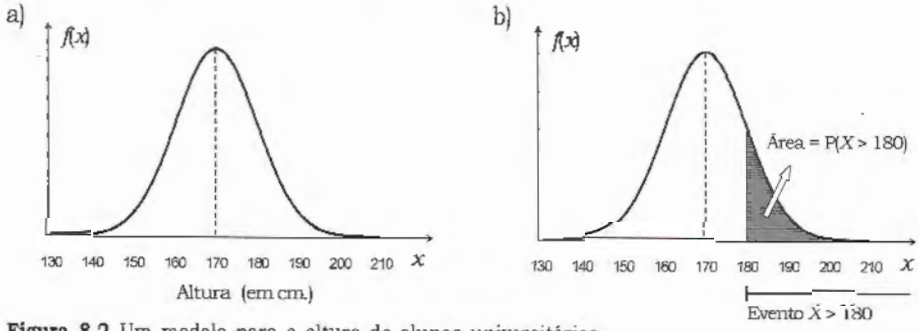


Figura 8.2 Um modelo para a altura de alunos universitários.

8.1 Distribuições normais

A distribuição normal é caracterizada por uma função, cujo gráfico descreve uma curva em forma de sino. Esta distribuição depende de dois parâmetros, a saber:

- μ (*média* ou *valor esperado*): especifica a posição central da distribuição de probabilidades;
- σ (*desvio padrão*): especifica a variabilidade da distribuição de probabilidades.[2]

A Figura 8.3 apresenta a forma gráfica de um modelo normal genérico, com parâmetros μ e σ. A curva é perfeitamente simétrica em torno da média μ e, independentemente dos valores de μ e σ, a área total entre a curva e o eixo-X é igual a 1 (um), permitindo identificar probabilidades de eventos como áreas sob a curva, como já ilustramos na Figura 8.2b.

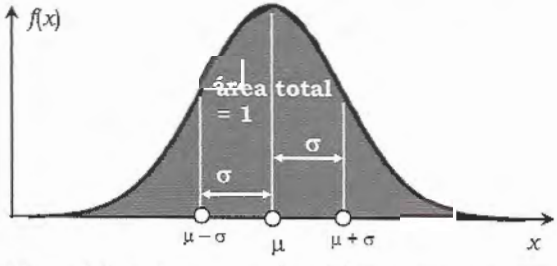


Figura 8.3 Gráfico de uma distribuição normal com parâmetros μ e σ.

[2] Os parâmetros μ e σ do modelo normal têm analogia com as estatísticas $\bar{X}$ e S (Capítulo 6), usadas para medir, respectivamente, a posição central e a dispersão de uma distribuição de frequências.

A Figura 8.4 mostra diferentes modelos normais, em termos dos parâmetros μ e σ. Estes modelos podem representar, por exemplo, a distribuição de alturas de crianças, em diferentes populações.

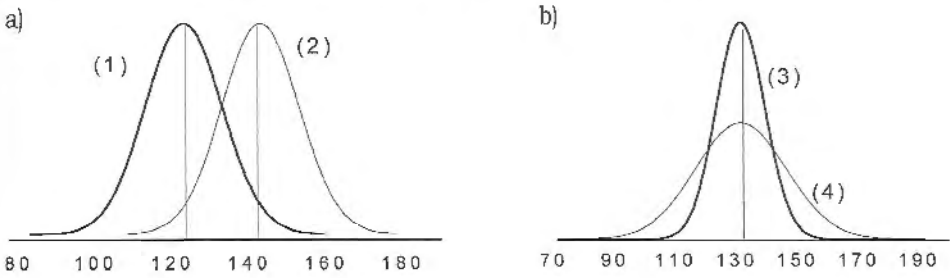


Figura 8.4 Distribuições normais em função dos parâmetros μ e σ.

As duas distribuições da Figura 8.4a podem representar, por exemplo, (1) *alturas de estudantes da primeira série do ensino fundamental* e (2) da *quarta série*. Podemos admitir que ambas as distribuições apresentam, aproximadamente, a mesma dispersão ($\sigma_1 \approx \sigma_2$), porém, na quarta série os estudantes devem ter, em média, alturas maiores do que os estudantes da primeira série ($\mu_2 > \mu_1$). Por outro lado, as distribuições da Figura 8.4b podem representar (3) *alturas de estudantes da terceira série* e (4) *alturas de estudantes da primeira à quinta série*. É razoável supor, neste caso, que a média das alturas dos dois grupos de estudantes deve ser aproximadamente igual ($\mu_3 \approx \mu_4$), mas a dispersão deve ser maior no grupo formado da primeira à quinta série ($\sigma_4 > \sigma_3$).

Valores padronizados e a distribuição normal padrão

Com o objetivo de facilitar a obtenção de determinadas áreas sob uma curva normal, podemos fazer uma transformação na variável, levando-a para a distribuição normal com média 0 (zero) e desvio padrão 1 (um).

> A distribuição normal com média 0 (zero) e desvio padrão 1 (um) é conhecida como *distribuição normal padrão*.

Para transformar um valor x, de uma distribuição normal com média μ e desvio padrão σ, em um valor z da distribuição normal padrão, basta fazer a seguinte operação:

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

O valor z conhecido como *valor padronizado* é uma medida relativa. Mede o quanto x se afasta da média (μ), em unidade de desvio padrão (σ).

EXEMPLO 8.3 Suponha que numa certa universidade, a altura dos estudantes do sexo masculino tenha distribuição normal com média $\mu = 170$ cm e desvio padrão $\sigma = 10$ cm. A Figura 8.5 mostra a relação entre a escala dos valores das alturas de universitários masculinos (x) e seus correspondentes valores padronizados (z). Por exemplo, para um estudante de altura $x = 180$ cm, temos o valor padronizado:

$$z = \frac{180 - 170}{10} = 1$$

Podemos dizer que este estudante de altura 180 cm encontra-se a 1 (um) desvio padrão acima da altura média dos estudantes do sexo masculino da universidade.

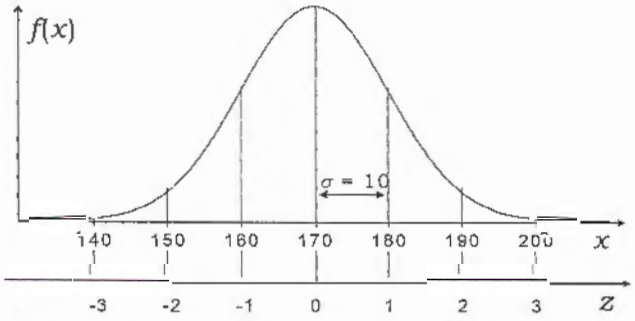


Figura 8.5 Transformação de valores de alturas de universitários (x) em valores padronizados (z).

Seja X a altura, em centímetro, de um estudante do sexo masculino, selecionado ao acaso. Considere que temos interesse no evento $X > 180$. A Figura 8.6 mostra a equivalência da probabilidade deste evento, $P(X > 180)$, com área na distribuição normal padrão. Para facilitar a notação, identificaremos por Z uma variável aleatória com distribuição normal padrão.

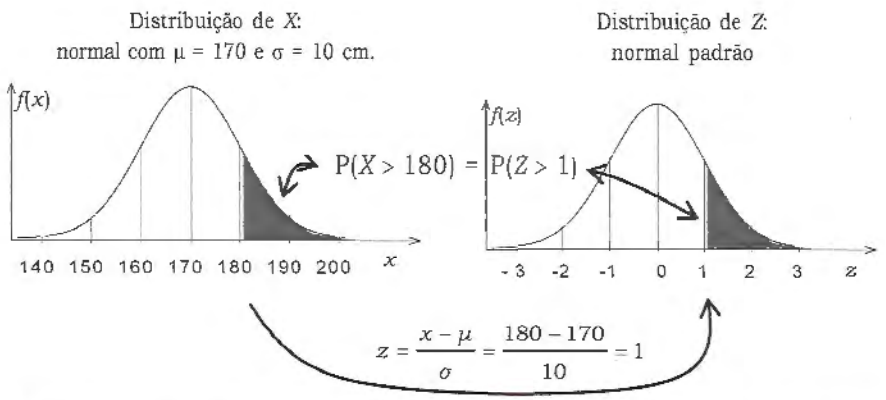


Figura 8.6 Transformação de um evento da distribuição normal de parâmetros μ = 170 cm e σ = 10 cm em um evento da distribuição normal padrão.

Exercícios

1) Supondo que as alturas dos estudantes de uma universidade tenham distribuição normal com média 170 cm e desvio padrão 10 cm, encontre os valores padronizados de:
   a) x = 190 cm; b) x = 185 cm; c) x = 170 cm; d) x = 165 cm.
2) Considerando o exercício anterior e lembrando que a distribuição normal é perfeitamente simétrica em torno da média μ, qual é a probabilidade de um estudante sorteado dessa universidade apresentar altura acima de 170 cm?
3) Suponha que as notas X de um vestibular tenham distribuição normal com média de 60 pontos e desvio padrão de 15 pontos.
   a) Se você prestou esse vestibular e obteve nota x = 80 pontos, qual é a sua posição relativa em relação à média dos vestibulandos, em unidade de desvio padrão?
   b) Se foram considerados aprovados os candidatos que obtiveram nota mínima correspondente a 1 (um) desvio padrão acima da média, qual é a nota mínima de aprovação na escala original?

8.2 Tabela da distribuição normal padrão

Como vimos na seção precedente, as probabilidades de uma distribuição normal podem ser representadas por áreas sob a curva da distribuição normal padrão. A Tabela 4 do apêndice relaciona valores positivos de z com áreas sob a cauda superior da curva. Os valores de z são apresentados com duas decimais. A primeira decimal fica na coluna

da esquerda e a segunda decimal na linha do topo da tabela. A Figura 8.7 mostra como podemos usar essa tabela.

	Segunda decimal de z				
z	0	1	2	...	9
0,0		↓			
0,1					
0,2	→	0,4168			
...					

Área na cauda superior

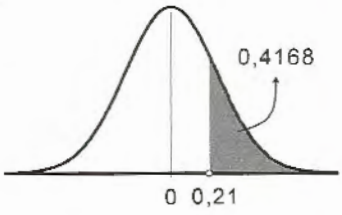


Figura 8.7 Ilustração do uso da tabela da distribuição normal padrão (Tabela 4 do apêndice) para encontrar a área na cauda superior relativa ao valor de $z = 0,21$.

EXEMPLO 8.3 (CONTINUAÇÃO) Suponhamos que a altura X de um estudante do sexo masculino, tomado ao acaso de uma universidade, tem distribuição normal com média 170 cm e desvio padrão 10 cm. Vimos que a probabilidade de ele acusar altura superior a 180 cm corresponde à área acima de $z = 1$ da curva normal padrão, isto é, $P(X > 180) = P(Z > 1)$. Usando a Tabela 4 do apêndice, podemos encontrar esta área (probabilidade), como ilustra o esquema seguinte:

	Segunda decimal de z		
z	0	...	9
...			
1,0	0,1587		
...			

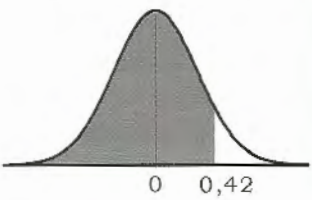
Portanto,
$P(X > 180) = 0,1587$

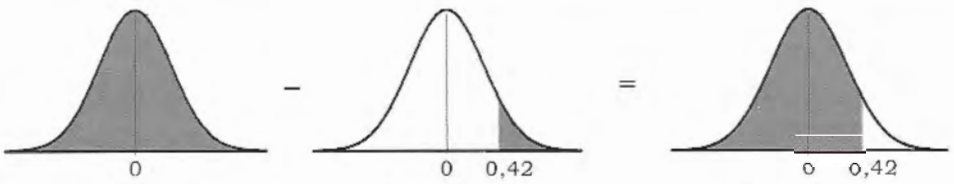
■

A Tabela 4 considera valores de z entre 0 (zero) e 5 (cinco). Além de $z = 5$ a área pode ser consaiderada nula. Aliás, a partir de 3 (três) a área já é praticamente nula. Áreas para valores negativos de z podem ser obtidas por simetria, considerando os correspondentes valores positivos.

EXEMPLO 8.4 Seja Z uma variável aleatória com distribuição normal padrão. Vamos usar a Tabela 4 para encontrar as seguintes probabilidades:

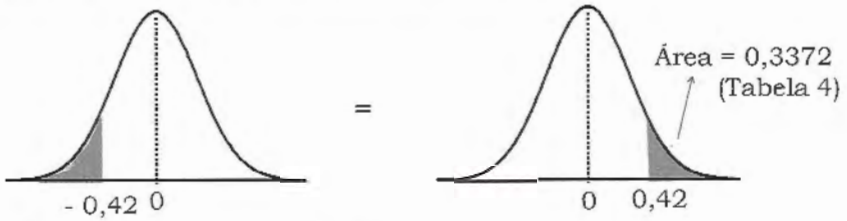
a) $P(Z < 0,42)$. Esta probabilidade corresponde à área da distribuição normal padrão indicada ao lado. Podemos obter esta área, fazendo a seguinte operação:





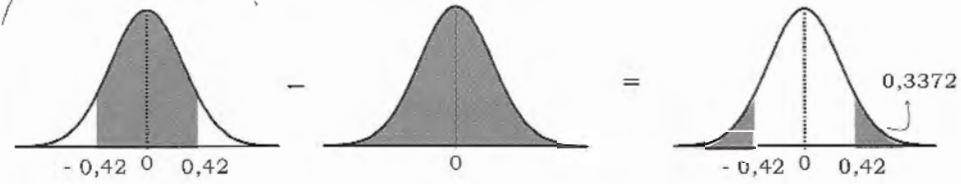
Portanto, $P(Z < 0,42) = 0,6628$.

b) $P(Z < -0,42)$. O esquema seguinte mostra como podemos usar a simetria da curva para obter a área pedida na Tabela 4.



Portanto, $P(Z < -0,42) = 0,3372$.

c) $P(-0,42 < Z < 0,42)$.

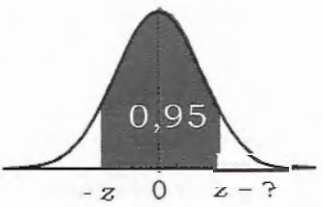


Então, $P(-0,42 < Z < 0,42) = 1 - 2\times(0,3372) = 0,3256$.

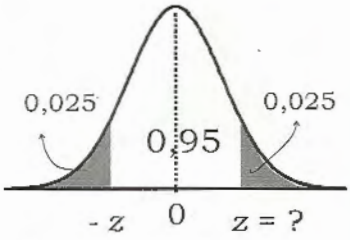
■

Como vimos nos exemplos precedentes, podemos obter a probabilidade de qualquer evento relativo a uma variável normal padrão, por manipulações adequadas com áreas sob a curva. O Exemplo 8.5 mostra como obter um valor de z a partir da fixação de uma certa área de interesse.

Exemplo 8.5 Qual é o valor de z, tal que $P(-z < Z < z) = 0,95$? Ou seja, precisamos obter z, tal que no intervalo de $-z$ até z resulte numa área sob a curva de 0,95, como ilustra a figura ao lado.



Considerando a simetria da curva normal e o fato de a área total sob a curva ser igual a 1 (um), podemos transformar esta pergunta em: *qual é o valor de z que deixa uma área de* 0,025 *além dele*? A figura ao lado ilustra a equivalência entre as duas perguntas.



Entrando com o valor de área 0,025 na Tabela 4 do apêndice, encontramos o valor de z igual a 1,96. Este processo está ilustrado ao lado.

z	0,00	0,01	...	0,06	...	0,09
...						
1,9				0,025		
...						

■

EXEMPLO 8.6 Suponha que o desempenho dos alunos das três últimas fases do Curso de Ciências da Computação da UFSC tenha distribuição normal de média 2,5 e desvio padrão de 0,6.[3] Selecionando aleatoriamente um aluno desta população, qual a probabilidade de ele acusar desempenho entre 2 e 3,5?

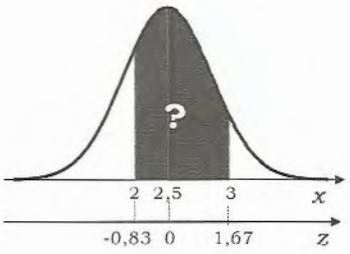
Solução: Primeiramente precisamos transformar os valores de desempenho, x, em valores padronizados:

$$z = \frac{x - \mu}{\sigma} = \frac{x - 2{,}5}{0{,}6}$$

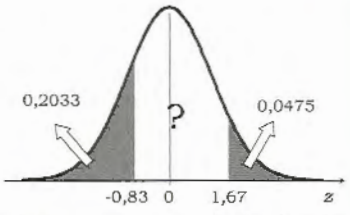
Para $x = 2$: $z = \dfrac{2 - 2{,}5}{0{,}6} = -0{,}83$

Para $x = 3{,}5$: $z = \dfrac{3{,}5 - 2{,}5}{0{,}6} = 1{,}67$

(veja a figura ao lado).



Usando a Tabela 4 do apêndice, encontramos para $z = -0{,}83$ e $z = 1{,}67$ as áreas nas extremidades da curva: 0,2033 e 0,0475, respectivamente (lembrando que para valores negativos de z, como -0,83, procuramos na Tabela 4 o seu valor simétrico positivo, no caso, $z = 0{,}83$). É fácil observar, pela figura ao lado, que a probabilidade desejada corresponde ao complemento da soma destas áreas, ou seja:



$$P(2 < X < 3{,}5) = 1 - (0{,}2033 + 0{,}0475) = 0{,}7492.$$

[3] Foram usados como estimativas de μ e σ, os valores das estatísticas $\overline{X}$ e S, calculadas a partir dos dados observados nesta população (anexo do Capítulo 2).

Exercícios

4) Seja Z uma variável aleatória com distribuição normal padrão. Calcule:
   a) $P(Z > 1{,}65)$; b) $P(Z < 1{,}65)$; c) $P(-1 < Z < 1)$;
   d) $P(-2 < Z < 2)$; e) $P(-3 < Z < 3)$; f) $P(Z > 6)$;
   g) o valor de z, tal que $P(-z < Z < z) = 0{,}90$;
   h) o valor de z, tal que $P(-z < Z < z) = 0{,}99$.

5) Sendo X a variável aleatória que representa a altura de um estudante tomado ao acaso de uma universidade, supostamente com distribuição normal de média 170 cm e desvio padrão 10 cm, calcule:
   a) $P(X > 190)$; b) $P(150 < X < 190)$; c) $P(X \quad 160)$;
   d) a percentagem esperada de estudantes com altura entre 150 e 190 cm.

6) Admitindo que a distribuição do quociente de inteligência (Q.I.) de crianças de uma certa escola seja normal com média 100 pontos e desvio padrão 10 pontos, calcule:
   a) a probabilidade de uma criança, tomada ao acaso desta escola, acusar Q.I. superior a 120 pontos;
   b) a percentagem esperada de crianças com Q.I. na faixa de 90 a 110 pontos.

7) Suponha que numa certa região, o peso dos homens adultos tenha distribuição normal com média 70 kg e desvio padrão 16 kg. E o peso das mulheres adultas tenha distribuição normal com média 60 kg e desvio padrão 12 kg. Ao selecionar uma pessoa ao acaso, o que é mais provável: *uma mulher com mais de 75 kg* ou *um homem com mais de 90 kg*? Responda calculando essas probabilidades.

8.3 Dados observados e modelo normal

A Figura 8.8 mostra um histograma de frequências das médias diárias de pressão intraocular, numa amostra de 43 indivíduos sadios. Observamos que o traçado do gráfico se aproxima de uma curva em forma de sino, donde podemos inferir que um modelo normal pode representar razoavelmente bem a distribuição desta variável, em indivíduos sadios.

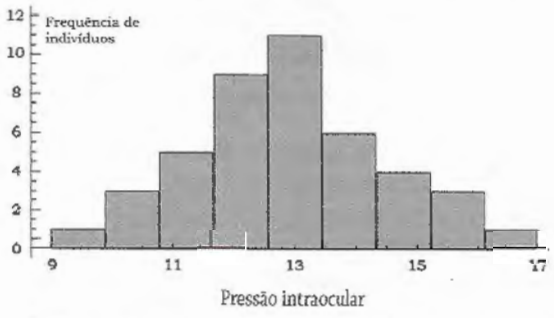


Figura 8.8 Histograma de frequências das médias diárias de pressão intraocular, numa amostra de 43 indivíduos sadios.

Uma variável que possa ser identificada como uma *soma*, ou *média*, de vários itens, geralmente se distribui de forma parecida com uma distribuição normal. É o caso do exemplo anterior, em que cada valor corresponde à média aritmética de sete medidas de pressão intraocular, observadas ao longo do dia. As medidas físicas ou comportamentais, tais como altura, peso, quociente de inteligência e índices de aptidões, também costumam se distribuir de forma parecida com um modelo normal, porque elas podem ser vistas como *somas* de uma infinidade de componentes inerentes ao indivíduo e ao seu meio.

Quando temos uma variável que acreditamos ter distribuição aproximadamente normal, podemos usar algumas propriedades desta distribuição na análise dos dados dessa variável. Uma propriedade da distribuição normal, muito usada na análise exploratória de dados, é a seguinte:

- ao afastar um desvio padrão, em ambos os lados da média (intervalo de $\mu - \sigma$ até $\mu + \sigma$), a área sob a curva atinge, aproximadamente, 0,683;
- ao afastar dois desvios padrões (intervalo de $\mu - 2\sigma$ até $\mu + 2\sigma$), a área cresce para 0,955;
- o afastamento de três desvios padrões (intervalo de $\mu - 3\sigma$ até $\mu + 3\sigma$) gera uma área de 0,997 (veja a Figura 8.9).

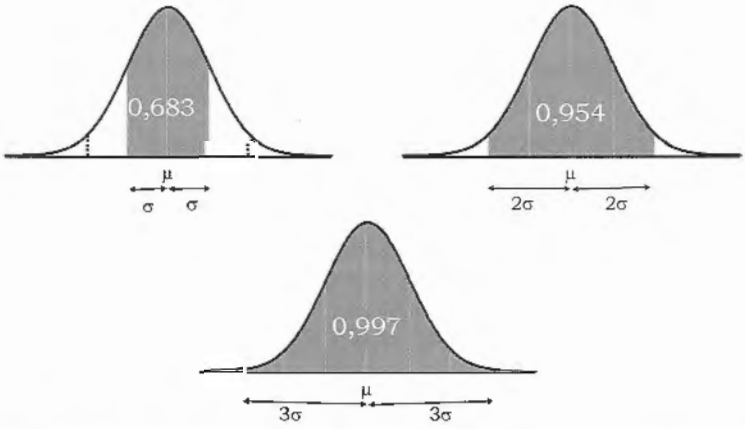


Figura 8.9 Áreas sob a curva normal em função de afastamentos de desvios padrões σ, em torno da média μ.

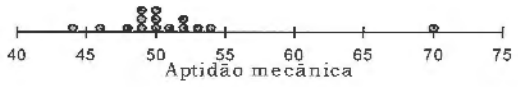
Dado um conjunto de valores, podemos calcular a média $\overline{X}$ e o desvio padrão S, como vimos no Capítulo 6. Se os dados em análise se distribuem de forma parecida com um modelo normal, devemos esperar:

- em torno de 95% dos dados em $\overline{X} \pm 2S$ (isto é, no intervalo de $\overline{X} - 2S$ até $\overline{X} + 2S$); e
- mais de 99% dos dados em $\overline{X} \pm 3S$ (isto é, no intervalo de $\overline{X} - 3S$ até $\overline{X} + 3S$).

Assim, algum valor que esteja fora do intervalo $\overline{X} \pm 3S$ pode ser considerado um valor *discrepante* dos demais. Valores fora do intervalo $\overline{X} \pm 2S$ podem ser vistos como *suspeitos*.

Exemplo 8.7 Sejam os seguintes valores de aptidão mecânica, numa turma de crianças.

44 52 50 49 52 46 53 48
50 70 54 49 51 50 49



Pelo diagrama de pontos, observamos que, com exceção do valor 70, os demais apresentam-se de maneira compatível com um modelo normal. Calculando a média aritmética e o desvio padrão desses dados, temos:

$\overline{X} = 51{,}1$ pontos e $S = 5{,}8$ pontos.[4]

Daí:

$\overline{X} \pm 2S = 51{,}1 \pm 2(5{,}8) = 51{,}1 \pm 11{,}6$ ➔ intervalo de 39,5 a 62,7 pontos;

$\overline{X} \pm 3S = 51{,}1 \pm 3(5{,}8) = 51{,}1 \pm 17{,}4$ ➔ intervalo de 33,7 a 68,5 pontos.

Verificamos que, com exceção do 70, todos os demais valores estão no intervalo $\overline{X} \pm 2S$. Aliás, o 70 também não pertence ao intervalo $\overline{X} \pm 3S$, caracterizando um *ponto discrepante*. A criança que obteve 70 no teste de aptidão mecânica é, neste contexto, *anormal* perante as demais crianças pesquisadas.

8.4 Aproximação normal à binomial

Em muitas situações práticas, a distribuição normal pode ser usada como uma aproximação razoável de outras distribuições. É o que acontece, por exemplo, em experimentos binomiais com n grande. Apesar de a distribuição verdadeira ser a binomial, os cálculos das probabilidades podem ser feitos com a distribuição normal. Seja o problema de amostragem e as variáveis aleatórias binomiais X e Y definidas na Figura 8.10.

[4] Os cálculos de $\overline{X}$ e S foram vistos no Capítulo 6.

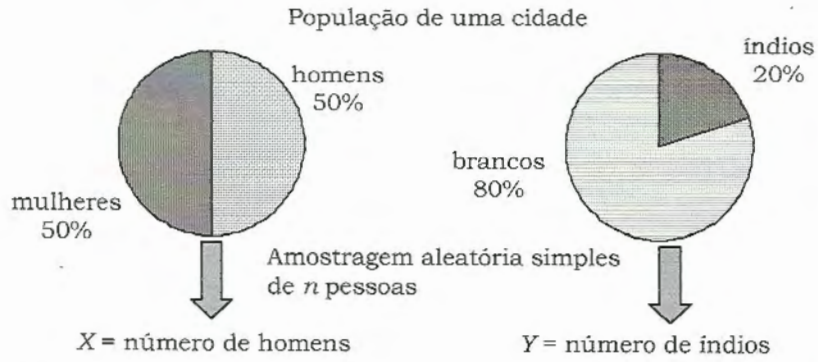


Figura 8.10 Ilustração de duas variáveis aleatórias binomiais.

Ambas as variáveis aleatórias têm distribuição binomial com n igual ao tamanho da amostra. Quanto ao parâmetro π, temos X com $\pi = 0{,}5$ e Y com $\pi = 0{,}2$. A Figura 8.11 apresenta as distribuições de probabilidades de X e Y para $n = 2$, 10 e 50.

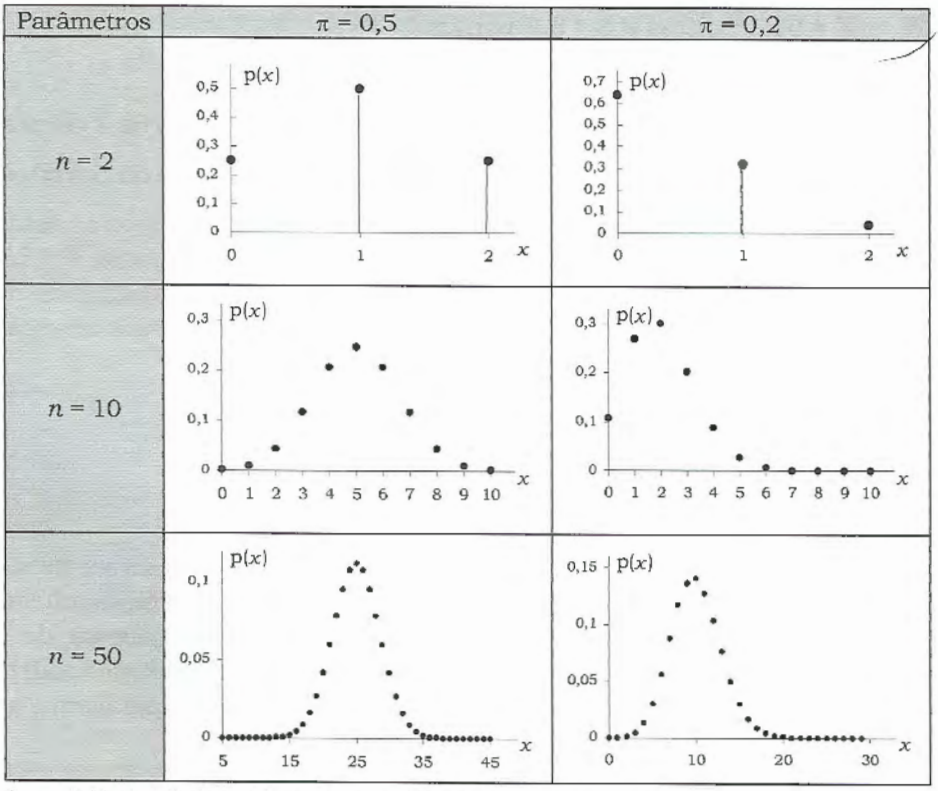


Figura 8.11 Distribuições binomiais para diferentes valores de n e π.

Verificamos pela Figura 8.11 que, para $n = 50$, a forma da distribuição binomial aproxima-se da curva de uma distribuição normal. Quando $\pi = 0{,}5$, a aproximação já parece razoável para $n = 10$.

De maneira geral, as condições para se fazer uma aproximação da distribuição binomial para a normal são:

1) n grande e

2) π não muito próximo de 0 (zero) ou de 1 (um).

Uma regra prática considera a aproximação razoável se as duas seguintes inequações forem satisfeitas:

a) $n \cdot \pi \geq 5$

b) $n \cdot (1 - \pi) \geq 5$

Ao aproximar uma *distribuição binomial* para uma *normal*, podemos obter os parâmetros μ e σ da normal, em função dos parâmetros n e π da binomial, segundo as expressões seguintes:

$$\mu = n \cdot \pi$$

$$\sigma = \sqrt{n \cdot \pi \cdot (1 - \pi)}$$

Exemplo 8.8 Observar o número, Y, de respostas favoráveis, numa amostra aleatória de $n = 50$ pessoas, as quais foram indagadas a respeito da opinião (favorável ou contrária) sobre um projeto municipal. Suponha que na população existam 40% de favoráveis.

Pelas características do experimento, a variável aleatória Y tem distribuição binomial com parâmetros $n = 50$ e $\pi = 0{,}4$. Como n é grande e π não é um valor muito próximo de *zero* ou de *um*, podemos usar a aproximação normal.[5] Esta distribuição normal deve ter média μ e desvio padrão σ dados, respectivamente, por:

$$\mu = n \cdot \pi = 50 \cdot (0{,}4) = 20$$

$$\sigma = \sqrt{n \cdot \pi \cdot (1 - \pi)} = \sqrt{50 \cdot (0{,}4) \cdot (1 - 0{,}4)} = 3{,}464$$

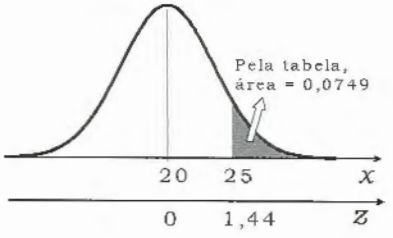
Calculemos, como exemplo, a probabilidade de se ter na amostra 25 ou mais de favoráveis, isto é, $P(X \geq 25)$. Esta probabilidade pode ser aproximada por uma área sob a curva da distribuição normal de média

[5] Poderíamos usar a regra prática: (a) $n \cdot \pi = 50(0{,}4) = 20$ e (b) $n \cdot (1 - \pi) = 50(1 - 0{,}4) = 30$. Como ambos os resultados são não inferiores a cinco, podemos usar a aproximação normal.

$\mu = 20$ e desvio padrão $\sigma = 3{,}464$. O valor $x = 25$ corresponde ao seguinte valor padronizado:

$$z = \frac{x - \mu}{\sigma} = \frac{25 - 20}{3{,}464} = 1{,}44$$

Usando a Tabela 4 (apêndice), encontramos a probabilidade 0,0749. Esquematicamente:



CORREÇÃO DE CONTINUIDADE

Ao calcular probabilidades de eventos oriundos de experimentos binomiais como áreas sob uma curva normal, estamos fazendo uma aproximação de uma variável aleatória discreta, que só assume valores inteiros, para uma variável contínua, cujos eventos constituem intervalos de números reais. Por isso, devemos fazer alguns ajustes, como mostra o exemplo seguinte.

EXEMPLO 8.9 Seja Y o número de caras obtidas em 10 lançamentos imparciais de uma moeda perfeitamente equilibrada.

Pelas características do experimento, podemos deduzir que Y tem distribuição binomial com $n = 10$ e $\pi = 0{,}5$, a qual pode ser aproximada pela distribuição normal de média e desvio padrão dados por:

$$\mu = n \cdot \pi = 10 \cdot (0{,}5) = 5$$

$$\sigma = \sqrt{n \cdot \pi \cdot (1 - \pi)} = \sqrt{10 \cdot (0{,}5) \cdot (1 - 0{,}5)} = \sqrt{2{,}5} = 1{,}58$$

Seja o seguinte evento de interesse: $\{Y = 4\}$, isto é, *ocorrer quatro caras*. Ao expressar este evento em termos de uma variável aleatória contínua X, com distribuição normal, devemos considerar um intervalo em torno do valor 4, porque, para variáveis contínuas, só faz sentido avaliar probabilidades em *intervalos*. O intervalo adequado, neste caso, é construído pela subtração e soma de meia unidade ao valor quatro, ou seja, $\{3{,}5 < X < 4{,}5\}$, como mostra a Figura 8.12.

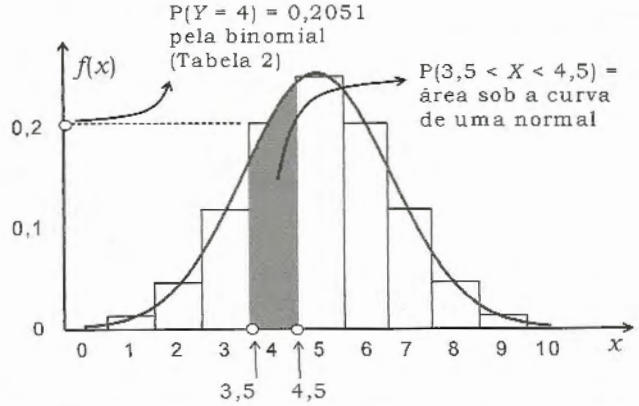


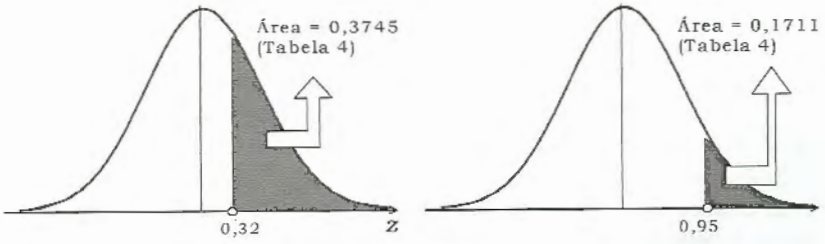
Figura 8.12 Aproximação da probabilidade do evento {$Y = 4$} (da distribuição binomial) para a probabilidade do evento {$3,5 < X < 4,5$} (da distribuição normal).

Usando a distribuição normal, a probabilidade do evento {$3,5 < X < 4,5$} deve ser colocada em termos de valores padronizados:

$$z = \frac{x - \mu}{\sigma} = \frac{x - 5}{1,58}$$

Para $x = 3,5$, temos $z = -0,95$ e para $x = 4,5$, temos $z = -0,32$, encontrando a probabilidade 0,2034, conforme mostra o esquema a seguir:

que pode ser obtida pela diferença das duas áreas representadas abaixo:



Então, P(3,5 < X < 4,5) = 0,3745 – 0,1711 = 0,2034.

É claro que neste exemplo é bem mais fácil usar a distribuição binomial. A probabilidade pedida é encontrada diretamente na Tabela 2 do apêndice, sendo igual a 0,2051. Mas quando n é grande, a aproximação normal é mais fácil.

Exercícios

8) Sejam dez lançamentos imparciais de uma moeda perfeitamente equilibrada. Calcule a probabilidade de ocorrer *mais de 6 caras*, usando:
   a) a distribuição binomial e
   b) a aproximação normal.
   Obs: ao usar a aproximação normal você deve considerar o evento $\{X > 6{,}5\}$ (correção de continuidade).

9) Com respeito ao exercício anterior, calcule a probabilidade de ocorrer o evento *cinco ou mais caras* (use a distribuição normal).

10) Resolva novamente o Exemplo 8.8, aplicando a correção de continuidade.

11) Numa amostra aleatória de 3.000 eleitores, qual é a probabilidade de a maioria se declarar favorável a um certo candidato, se na população existem 52% de favoráveis a este candidato?

Exercícios complementares

12) Um teste padronizado é aplicado a um grande número de estudantes. Os seus resultados são normalmente distribuídos com média de 500 pontos e desvio padrão de 100 pontos. Se João conseguir 650 pontos, qual é a percentagem esperada de estudantes com mais pontos do que João?

13) Suponha que as notas de um teste de aptidão tenham distribuição normal com média 60 e desvio padrão 20. Qual é a proporção de notas que
   a) excedem 85?
   b) estão abaixo de 50?

14) Considere que na cidade Paraíso, composta de um milhão de habitantes, existam 40% de homens e 60% de mulheres. Numa amostra extraída por sorteio (amostra aleatória), calcule a probabilidade de se obter mais mulheres do que homens, considerando:
   a) que a amostra tenha sido de cinco pessoas;
   b) que a amostra tenha sido de cinquenta pessoas.

15) a) Um exame de múltipla escolha consiste em dez questões, cada uma com quatro possibilidades de escolha. A aprovação exige, no mínimo, 50% de acertos. Qual é a chance de aprovação se o candidato comparece ao exame sem saber absolutamente nada, apelando apenas para o "palpite"?
   b) E se o exame tivesse cem questões?

16) Calculou-se em 70 minutos o tempo médio para o vestibular de uma universidade, com desvio padrão de 12 minutos. Quanto deve ser a duração da prova, de modo a permitir tempo suficiente para que 90% dos vestibulandos terminem a prova? Admita distribuição normal para o tempo de duração da prova.

Parte IV

Inferência estatística

Como generalizar resultados de uma amostra para a população de onde ela foi extraída

Como testar hipóteses com base em amostras

Capítulo 9

Estimação de parâmetros

Neste capítulo, estudaremos o problema de avaliar certas características dos elementos da população (*parâmetros*), com base em operações com os dados de uma amostra (*estatísticas*). É o que acontece nas pesquisas eleitorais, em que queremos conhecer as percentagens de cada candidato na população de eleitores (parâmetros), mas observamos apenas uma parte da população (uma amostra), na qual podemos calcular as percentagens de intenção de voto relativas a cada candidato (estatísticas).

Na estimação de parâmetros fazemos um raciocínio tipicamente indutivo, porque generalizamos resultados *da parte* (amostra) para *o todo* (população). É um caso especial de inferência estatística (ver Figura 9.1).

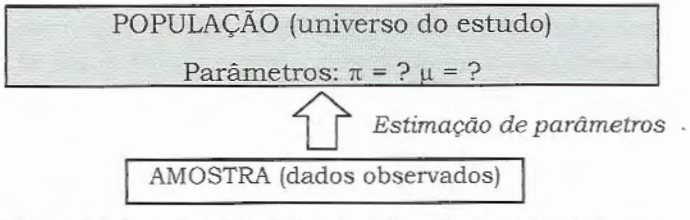


Figura 9.1 O raciocínio indutivo da estimação de parâmetros: uma forma de inferência estatística.

Reforçando algumas definições:

População é o conjunto de elementos para os quais desejamos que as conclusões da pesquisa sejam válidas, com a restrição de que esses elementos possam ser observados ou mensurados sob as mesmas condições.

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

Amostra aleatória simples: uma parte da população, sendo que os elementos são extraídos por sorteio.

Estatística: alguma medida associada com os dados de uma amostra a ser extraída da população. Quando usada com o objetivo de avaliar (*estimar*) o valor de algum parâmetro, também é chamada de **estimador**.

Erro amostral é a diferença entre uma estatística e o parâmetro que se quer estimar.

Estimativa: valor da estatística (estimador), calculado com base na amostra efetivamente observada.

EXEMPLO 9.1 A prefeitura pretende avaliar a aceitação de um projeto de mudança no transporte coletivo. Depois de apresentá-lo aos usuários, os responsáveis por sua execução pretendem conhecer, mesmo que de forma aproximada, o parâmetro:

π = proporção de favoráveis ao projeto (na população de usuários do transporte coletivo do município).

Para estimar este parâmetro, a prefeitura planeja uma amostragem aleatória simples de $n = 400$ usuários. Dessa amostra, calcular a estatística:

P = proporção de moradores favoráveis ao projeto (na amostra).

Observada efetivamente a amostra, devemos ter $P \neq \pi$, devido ao *erro amostral*. Então, pensaremos em avaliar a *margem de erro* que podemos estar cometendo por examinar apenas uma amostra e não toda a população.

■

EXEMPLO 9.2 Para estudar o efeito da merenda escolar, introduzida nas escolas de um município, planeja-se acompanhar uma amostra de $n = 100$ crianças, que estão entrando na rede municipal de ensino. Dentre diversas características de interesse, pretende-se avaliar o parâmetro:

μ = ganho médio de peso durante o primeiro ano letivo (na população de crianças da rede municipal de ensino)

Da amostra de crianças em estudo, pode-se calcular a estatística:

$\overline{X}$ = ganho médio de peso, durante o primeiro ano letivo, das 100 crianças em observação.

A estatística $\overline{X}$ pode ser usada como um estimador do parâmetro μ, mas, como no exemplo anterior, devemos ter $\overline{X} \neq \mu$ devido ao erro amostral. Nas próximas seções, vamos estudar um processo que permite avaliar a *margem de erro* que podemos estar cometendo por examinar apenas uma amostra e não toda a população.

■

Quando estivermos estudando a incidência de algum atributo numa certa população, geralmente o interesse está na *proporção*, ou *percentagem*, de elementos com o atributo, como no Exemplo 9.1. Por outro lado, quando estamos pesquisando alguma característica quantitativa, como no Exemplo 9.2, é comum o interesse em estimar uma *média*.

Apresentamos, a seguir, alguns parâmetros e as respectivas estatísticas que geralmente são usadas para estimá-los.[1]

PARÂMETROS (características da **população**)	ESTATÍSTICAS (características da **amostra**)
π = proporção de algum atributo, dentre os elementos da população.	P = proporção de elementos com o atributo, dentre os que serão observados na amostra.
μ = média de alguma variável quantitativa, nos elementos da população.	$\overline{X}$ = média da variável, a ser calculada com os elementos da amostra.
σ = desvio padrão de uma variável, dentre os elementos da população.	S = desvio padrão da variável, a ser calculado com os elementos da amostra.

Em geral, os parâmetros são *números desconhecidos* (somente serão conhecidos se for feito um *censo* – pesquisa de toda a população). Já as estatísticas são *variáveis aleatórias*, pois seus valores dependem dos elementos a serem sorteados na amostragem. Ao observar efetivamente uma amostra, a estatística se identifica com um valor (resultado do cálculo), chamado de *estimativa*. Por exemplo, se na amostra de $n = 400$ moradores do Exemplo 9.1, encontrarmos 240 favoráveis, então temos a seguinte estimativa para o parâmetro π:[2]

$$P = \frac{240}{400} = 0{,}60 \ \text{(ou, 60\%)}$$

---

[1] Lembramos que as expressões para o cálculo de algumas estatísticas, tais como a média $\overline{X}$ e o desvio padrão S, foram vistas no Capítulo 6.

[2] Na literatura de Estatística, geralmente são usadas letras minúsculas para as estimativas. Em nosso exemplo, $p = 0{,}60$. Neste livro, usaremos a mesma notação para estimador (uma variável aleatória) e estimativa (um número).

Contudo, não devemos esperar que este valor coincida com o parâmetro π, devido ao que chamamos de *erro amostral*. Um dos principais objetivos na teoria da estimação é estimar um *limite superior provável* para o erro amostral. Esse valor será a base para avaliarmos a precisão de nossa estimativa.

> Dizemos que uma estimativa é tão mais **precisa** quanto menor for o *limite superior provável* de seu erro amostral.

Toda a formulação que apresentaremos parte da suposição de que os dados em análise constituem uma amostra aleatória simples da população de interesse.

EXERCÍCIOS

1) O esquema seguinte representa uma população de noventa domicílios, situados em quadras residenciais. Os valores dentro dos quadradinhos (domicílios) indicam o número de cômodos. Esses valores, na verdade, somente serão conhecidos após a realização da pesquisa.

4	5	2	9
4		7	
1	2	6	4

1	4	4	6
4		5	
2	3	2	3

7	2	2	4
6		8	
2	4	5	6

8	5	2	3
8		5	
2	4	5	9

4	1	6	3
4		2	
5	6	4	3

2	3	5	4
4		3	
4	5	4	2

9	8	18	
22		8	9
7	7	9	9

8	7	9	6
14		9	9
8	7	12	

14		8	9
8	8	15	
8	9	8	8

Calcular os seguintes parâmetros:

a) π = *proporção de domicílios com mais de cinco cômodos*;

b) μ = *número médio de cômodos por domicílio.*

2) Selecione uma amostra aleatória simples de vinte domicílios da população do Exercício 1.3 Com base na amostra selecionada, calcule o valor das seguintes estatísticas:

a) P = *proporção de domicílios com mais de cinco cômodos, na amostra*;

b) $\overline{X}$ = *número médio de cômodos por domicílio, na amostra.*

---

[3] Se você não se lembrar de como extrair uma amostra aleatória simples, leia novamente a Seção 3.1 (Capítulo 3). Lembre que o primeiro passo é numerar os domicílios.

9.1 Distribuição amostral

Considere a seguinte pergunta, relativa ao Exemplo 9.1:

– o valor de P (proporção de favoráveis numa amostra de $n = 400$ usuários do transporte coletivo) vai ser um valor próximo da verdadeira proporção π, a qual se refere a *todos* os usuários do município?

Como na prática o valor de π é desconhecido, tentaremos responder a esta pergunta de forma indireta, através do conhecimento de como se distribuem os possíveis valores de P. Diferentes valores de P podem ser obtidos por diferentes amostras de n elementos, extraídas da população de interesse, sob as mesmas condições. Para cada amostra observada, temos um valor para P. A distribuição do conjunto de todos os possíveis valores de P, correspondentes às possíveis amostras de tamanho n, forma a chamada *distribuição amostral* de P.

A **distribuição amostral** de uma estatística é a distribuição dos possíveis valores dessa estatística, se examinássemos *todas* as possíveis amostras de tamanho n, extraídas aleatoriamente de uma população.

Para simplificar, vamos supor que a população em estudo seja bastante grande, de tal forma que, para cada elemento observado, a probabilidade de ele ser favorável seja sempre igual a π, independentemente dos elementos já observados. A Figura 9.2 mostra o modelo de probabilidades, referente a cada observação.

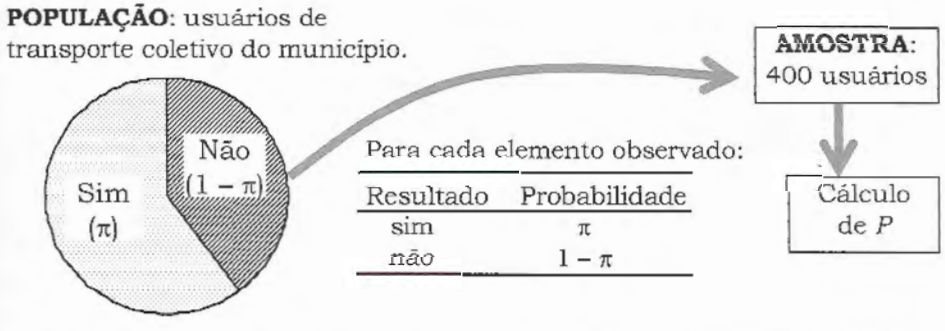


Figura 9.2 Modelo de probabilidades associado ao processo de amostragem do Exemplo 9.1.

UMA SIMULAÇÃO

Para ilustrarmos a distribuição amostral de P, conforme a situação da Figura 9.2, podemos simular várias amostras de tamanho $n = 400$. Como exemplo, suporemos, artificialmente, que o parâmetro é $\pi = 0{,}7$ (população com 70% de favoráveis). A simulação pode ser realizada com o apoio de uma tabela de números aleatórios (Tabela 1 do apêndice). Cada número de um algarismo da tabela simula a observação de um elemento da população, da seguinte forma:

- quando o algarismo extraído da tabela de números aleatórios for um valor do conjunto {0, 1, 2, 3, 4, 5, 6}, que acontece com probabilidade $^{7}/_{10} = 0{,}7$, simula a observação de um indivíduo *favorável* ao projeto;
- quando o algarismo for um valor do conjunto {7, 8, 9}, que ocorre com probabilidade $^{3}/_{10} = 0{,}3$, simula a observação de um indivíduo *contrário* ao projeto.

Ao observarmos 400 algarismos da tabela de números aleatórios, podemos calcular:

> P = proporção de números no conjunto {0, 1, 2, 3, 4, 5, 6}, simulando a proporção de indivíduos favoráveis ao projeto.

Para avaliarmos a distribuição amostral de P e termos informações sobre o erro amostral, precisamos repetir esse processo várias vezes, sob as mesmas condições. Os valores da Figura 9.3 referem-se a valores de P, oriundos da simulação de 100 amostras de tamanho $n = 400$.

Valores simulados de P							
0,70	0,67	0,69	0,70	0,74	0,71	0,68	0,69
0,69	0,71	0,71	0,68	0,71	0,71	0,70	0,70
0,72	0,71	0,68	0,69	0,69	0,66	0,69	0,69
0,67	0,72	0,73	0,71	0,70	0,67	0,70	0,71
0,70	0,72	0,71	0,68	0,70	0,72	0,70	0,72
0,73	0,67	0,71	0,76*	0,73	0,70	0,68	0,66
0,73	0,69	0,69	0,68	0,69	0,71	0,69	0,74
0,74	0,70	0,70	0,67	0,70	0,71	0,69	0,72
0,70	0,69	0,73	0,70	0,74	0,72	0,70	0,68
0,69	0,70	0,74	0,75	0,71	0,69	0,71	0,70
0,72	0,65*	0,69	0,69	0,70	0,71	0,66	0,71
0,66	0,65	0,68	0,69	0,69	0,68	0,71	0,71
0,72	0,68	0,68	0,73				

* Valor máximo e valor mínimo.

Histograma de frequências

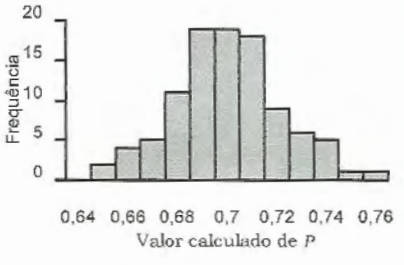


Figura 9.3 Cem observações da distribuição amostral de P, considerando amostras de tamanho $n = 400$ e $\pi = 0{,}70$.

Pela Figura 9.3, verificamos que em nenhuma amostra, dentre as 100 simuladas, resultou em um valor de P fora do intervalo de 0,65 a 0,76. Nesta situação fictícia, adotamos o valor de $\pi = 0{,}70$. Na simulação, verificamos que o valor mais distante foi 0,76, apontando um erro amostral igual a $0{,}76 - 0{,}70 = 0{,}06$. Podemos dizer que temos uma altíssima confiança de que uma estimativa P, obtida através de uma amostra aleatória simples de tamanho $n = 400$, sob as mesmas condições da simulação realizada, não carregará um erro amostral superior a 0,06 (ou seja, 6%).

O fato de nenhuma das amostras simuladas ter carregado um erro amostral superior a 0,06 não garante que, numa amostra efetivamente extraída da população em estudo, o erro amostral não possa ser superior a 0,06, pois sempre existe o efeito do *azar* ao sortearmos os elementos que irão compor a amostra. Neste contexto, as afirmações são sempre feitas com um certo *nível de confiança*.

Para entendermos melhor o significado do termo *nível de confiança*, podemos fazer o seguinte raciocínio em termos da nossa simulação: observamos que 96 valores de P, dentre os 100 simulados, resultaram em erros amostrais inferiores a 0,05 (veja a Figura 9.3). Assim, podemos afirmar que uma estimativa construída sob um modelo análogo ao da simulação deverá ter um erro amostral inferior a 0,05, com nível de confiança em torno de $^{96}/_{100} = 96\%$.

> Na prática examinamos apenas *uma* amostra, resultando em um único valor para a estatística – uma **estimativa**. Porém, o conhecimento da *distribuição amostral* da estatística permite avaliarmos um limite superior para o erro amostral (*margem de erro*), com certo *nível de confiança*.

Usando a distribuição normal

Na maioria dos problemas de estimação de parâmetros, não é necessário realizar simulações para avaliar a precisão de uma estimativa. Por exemplo, na estimação de uma *proporção*, com base em uma amostra aleatória simples, o experimento é tipicamente *binomial*, com parâmetros n (tamanho da amostra) e π (proporção do atributo em questão). No capítulo anterior, vimos que se n for grande, a distribuição binomial se aproxima de uma *distribuição normal*. No caso da estatística *proporção*, a média e o desvio padrão são determinados em função de n e π, da seguinte forma:[4]

[4] No capítulo anterior, trabalhamos mais com a variável aleatória X = *número de favoráveis na amostra*. Aqui estamos trabalhando com a *proporção* $P = X/n$, razão pela qual as expressões da média e do desvio padrão são diferentes. O subíndice P nas notações usuais de média e desvio padrão, μ e σ, é para lembrar que esses parâmetros referem-se à distribuição amostral de P.

$$\mu_p = \pi$$

$$\sigma_p = \sqrt{\frac{\pi \cdot (1-\pi)}{n}}$$

A Figura 9.4(a) mostra a forma aproximada da distribuição amostral de P. Note que esta distribuição está centrada no próprio valor do parâmetro de interesse, π. Pela teoria da distribuição normal, é sabido que existe 95% de probabilidade de que um valor seja observado a menos de 1,96 desvios padrões da média (Exemplo 8.5, Capítulo 8). Assim, com probabilidade de 95%, o erro amostral não deve exceder 1,96 desvios padrões, como mostra a Figura 9.4(b).

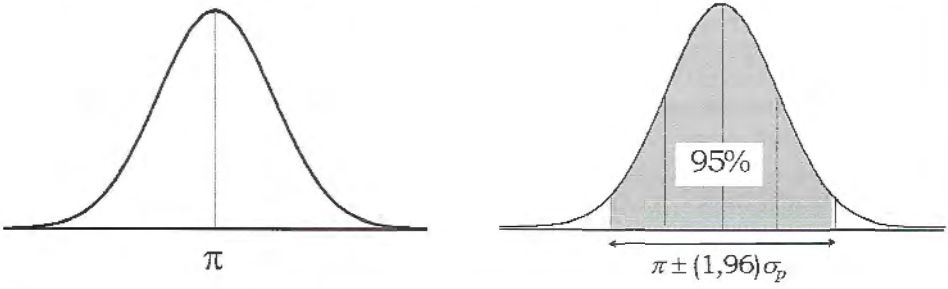


Figura 9.4 (a) Forma aproximada da distribuição amostral de P; (b) Faixa em que deve estar o valor de P calculado com base na amostra (95% de probabilidade).

O desvio padrão da distribuição amostral de uma estatística é comumente chamado de **erro padrão** da estatística.

9.2 Estimação de uma proporção

No que segue, estaremos considerando que já examinamos uma amostra aleatória simples da população de interesse.

Limitaremos o estudo para o caso em que o tamanho da amostra é razoavelmente grande e o atributo em observação não seja muito raro ou quase certo, de tal forma que seja válida a aproximação da distribuição binomial para a normal.[5] Nesta e na próxima seção, também suporemos

[5] Desde que π não seja próximo de 0 ou de 1, podemos usar a distribuição normal para $n \geq 30$. Uma discussão mais detalhada sobre esta aproximação foi feita na Seção 8.4.

que a população de onde foi extraída a amostra seja muito grande, não necessitando considerar o seu tamanho nos cálculos.

Com as suposições anteriores, o *erro padrão* de P pode ser estimado com os dados da própria amostra, usando a expressão:

$$S_p = \sqrt{\frac{P \cdot (1-P)}{n}}$$

onde P é a proporção do atributo, na amostra; e n é o tamanho da amostra.

Nível de confiança de 95%

Fixado o nível de confiança em 95%, como é usual na prática, o limite máximo para o erro amostral fica em torno de $(1{,}96)S_p$, pois, como ilustra a Figura 9.4(b), temos, aproximadamente, 95% de probabilidade de o valor de P cair a menos de 1,96 desvios padrões de π.

Exemplo 9.1 (continuação) Suponha que na amostra de $n = 400$ pessoas, encontramos 60% de favoráveis. Temos, então, $P = 0{,}60$ (ou 60%), com erro padrão:

$$S_p = \sqrt{\frac{P \cdot (1-P)}{n}} = \sqrt{\frac{(0{,}60) \cdot (0{,}40)}{400}} = 0{,}0245$$

Usando nível de confiança de 95%, temos um *limite superior para o erro amostral* de:

$$E = (1{,}96) \cdot S_p = (1{,}96) \cdot (0{,}0245) = 0{,}048 \text{ (ou } 4{,}8\%)$$

Representaremos por:

$$60{,}0\% \pm 4{,}8\%$$

o intervalo de limite inferior 60,0% – 4,8% = 55,2% e de limite superior 60,0% + 4,8% = 64,8%.

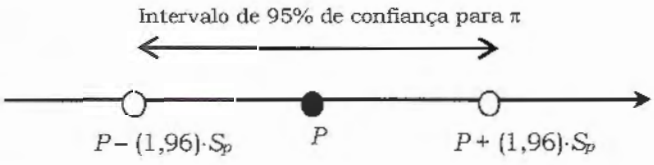
Podemos dizer, com nível de confiança de 95%, que o intervalo 60,0% ± 4,8% contém o parâmetro π (*proporção de favoráveis em toda a população*).

■

De modo geral, o intervalo centrado em P e com semiamplitude $E = (1{,}96) \cdot S_p$, representado por:

$$P \pm E \quad \text{ou} \quad (P - E, \quad P + E)$$

é dito um **intervalo de confiança** para o parâmetro π, com nível de confiança de 95%. O esquema seguinte ilustra este intervalo sobre a reta de números reais:



OUTROS NÍVEIS DE CONFIANÇA

A Figura 9.5 mostra uma tabela, construída com base na Tabela 4 do apêndice (tabela da distribuição normal padrão), que associa os níveis usuais de confiança com os respectivos valores de z.

Área	0,800	0,900	0,950	0,980	0,990	0,995	0,998
z	1,282	1,645	1,960	2,326	2,576	2,807	3,090

Figura 9.5 Valores de z para alguns níveis de confiança

Fixado o nível de confiança, podemos obter o correspondente valor de z, como ilustra a Figura 9.5. Depois, calculamos uma estimativa para o *limite superior do erro amostral* por:

$$E = z \cdot S_P$$

e o *intervalo de confiança* para π:

$$P \pm E$$

EXEMPLO 9.1 (CONTINUAÇÃO) Adote o nível de confiança de 99%. Então, pelo esquema da Figura 9.5, temos:

$$Área = 0,99 \rightarrow z = 2,576$$

resultando no seguinte limite provável para o erro amostral:

$$E = z \cdot S_P = (2,576).(0,0245) = 0,063 \text{ (ou } 6,3\%)$$

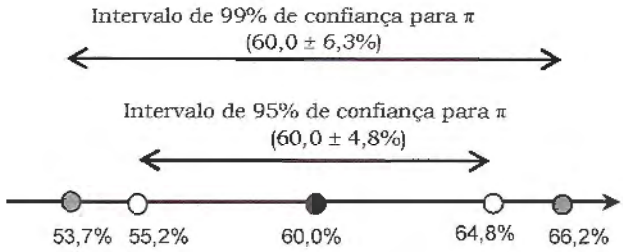
Então, com nível de confiança de 99%, o intervalo:

$$60,0\% \pm 6,3\%$$

deve conter o verdadeiro parâmetro π.

■

O esquema seguinte ilustra os intervalos de confiança para π com níveis de confiança de 95% e de 99%, referentes à amostra descrita no Exemplo 9.1.



Observe que, ao exigir maior nível de confiança, o intervalo de confiança aumenta em magnitude. Tente entender o porquê disto!

Para um *dado nível de confiança*, dizemos que uma estimativa é tão mais **precisa** quanto menor for a amplitude de seu intervalo de confiança.

Observe, pela expressão do intervalo de confiança, que a maneira natural de aumentarmos a precisão de uma estimativa é através do aumento do tamanho n da amostra.

Exercícios

3) (Para fazer em sala de aula.) Com respeito à população do Exemplo 9.1, mas agora considerando $\pi = 0,60$, simule 50 amostras de tamanho $n = 10$ (cada aluno deve simular uma ou duas amostras). Para cada amostra simulada, calcule P. Apresente os valores encontrados de P num histograma. Com base nessa simulação, discuta sobre o erro amostral associado a uma amostra de

tamanho $n = 10$, para estimar o parâmetro π, relativo à proporção de algum atributo da população.

4) Seja o problema de construir um intervalo de confiança para a proporção π de alunos favoráveis à presença da Polícia Militar no *Campus* de uma grande universidade, com base numa amostra aleatória simples de n alunos. Faça os itens abaixo e, com base nos resultados, discuta sobre a precisão das estimativas ao variar n e π.
   a) nível de confiança de 90%, $n = 400$, com 60% de favoráveis na amostra.
   b) nível de confiança de 90%, porém considerando que a amostra tenha sido de $n = 1.000$ alunos, sendo que 600 disseram ser favorável.
   c) nível de confiança de 95%, $n = 400$, com 80 favoráveis.
   d) nível de confiança de 95%, $n = 400$, com 320 favoráveis.
   e) nível de confiança de 95%, $n = 400$, com 200 favoráveis.

5) Numa pesquisa mercadológica, deseja-se estimar a proporção π de consumidores que passariam a usar certo produto após experimentá-lo pela primeira vez. Para atingir esse objetivo, selecionou-se uma amostra aleatória simples de $n = 200$ consumidores potenciais, fornecendo-lhes amostras grátis do produto. Depois de um mês, voltou-se a contatar os consumidores da amostra, oferecendo-lhes o produto por um certo preço. Trinta por cento da amostra decidiu adquirir o produto. Construa uma estimativa intervalar para π, com nível de confiança de 95%.

6) O vestibular COPERVE-1991 teve como tema de redação a possível mudança da capital de Florianópolis para Curitibanos.
   a) Foram observadas 400 redações, extraídas por sorteio, dentre todas as redações. Nessa amostra, 120 mostraram-se favoráveis à mudança da capital. O que se pode dizer a respeito da proporção de vestibulandos favoráveis à mudança, na amostra? E na população de vestibulandos?
   b) Foram observadas 400 redações, correspondentes aos alunos que prestaram o vestibular num dos locais de realização das provas (por exemplo, na região de Curitibanos). Nessa amostra, 250 eram favoráveis à mudança da capital. O que se pode dizer a respeito da proporção de favoráveis à mudança, na população de vestibulandos?

7) Num trabalho de auditoria nas contabilidades das empresas, para estimar a proporção de empresas que deixaram de pagar algum tributo no ano anterior, foi selecionada uma amostra aleatória simples de 40 empresas. Os resultados foram os seguintes (1 = deixou de pagar, 0 = pagou corretamente):

   0 0 1 0 1 0 0 0 1 1 0 1 0 0 1 0 0 0 0 0

   1 1 0 0 1 0 0 0 0 1 0 1 0 1 0 0 1 1 0 0

   Construa um intervalo de 90% de confiança para a população de empresas que deixaram de pagar corretamente os tributos no ano anterior.

   NOTA: Observe que quando os dados estão codificados com 0 e 1, o cálculo de P coincide com o cálculo da média aritmética $\bar{X}$, ou seja, a proporção é uma média em dados do tipo 0 e 1.

8) No anexo do Capítulo 4, temos o resultado de uma amostra aleatória simples de 120 famílias do bairro Saco Grande II, Florianópolis – SC, 1988. Uma das

características pesquisadas foi o uso (*sim* ou *não*) de programas de alimentação popular (PAP). Com base nessa amostra, construa um intervalo de 95% de confiança para o parâmetro π (proporção de famílias que usam programas de alimentação popular, em todo o bairro).

9) A amostra descrita no Exercício 8 está, na verdade, dividida em três localidades. Construa intervalos de 95% de confiança para a proporção de famílias que usam programas de alimentação popular, para cada localidade. Interprete esses intervalos de confiança.

NOTA: Observe que, ao trabalhar com subgrupos de uma amostra, as precisões das estimativas tendem a ser piores (intervalos de confiança mais longos), quando comparadas com a análise de toda a amostra.

9.3 Estimação de uma média

Quando a variável em estudo é quantitativa, normalmente se tem interesse no parâmetro μ (média). Tendo uma amostra aleatória simples da população de interesse, podemos ter uma estimativa de μ através do cálculo da média dos valores da amostra:

$$\overline{X} = \frac{\sum X}{n}$$

Como o valor de $\overline{X}$ vai depender da amostra selecionada, podemos falar em *erro padrão* e em *distribuição amostral* de $\overline{X}$. O *erro padrão* de $\overline{X}$ pode ser estimado com os dados da amostra por:

$$S_{\overline{X}} = \frac{S}{\sqrt{n}}$$

onde S é o desvio padrão dos dados, conforme apresentado no Capítulo 6. Por exemplo, se uma amostra de 9 alunos em que se observaram as seguintes notas:

8 10 9 6 7 9 8 7 8

temos a soma dos valores: $\sum X = 72$;

a média da amostra: $\overline{X} = \frac{72}{9} = 8{,}0$;

a soma dos valores quadráticos:

$$\sum X^2 = 8^2 + 10^2 + 9^2 + 6^2 + 7^2 + 9^2 + 8^2 + 7^2 + 8^2 = 588$$

a variância da amostra:

$$S^2 = \frac{\sum X^2 - n \cdot \overline{X}^2}{n-1} = \frac{588 - 9 \cdot (8)^2}{8} = 1{,}5$$

o desvio padrão da amostra: $S = \sqrt{1{,}5} = 1{,}225$;

e o erro padrão da média: $S_{\overline{X}} = \frac{S}{\sqrt{n}} = \frac{1{,}225}{\sqrt{9}} = 0{,}408$

■

Formalmente, o erro padrão de $\overline{X}$ é:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

onde σ é o desvio padrão de todos os elementos da população. Como, em geral, o parâmetro σ é desconhecido, usamos em seu lugar S, resultando na estimativa $S_{\overline{X}}$, apresentada anteriormente.

No Exercício 7, vimos que, se o conjunto de valores é formado por *zeros* e *uns*, sendo 1 quando o indivíduo tem uma certa característica, e 0 quando não tem, então a média aritmética desses valores é igual à proporção P de indivíduos com a característica. Da mesma forma, o erro padrão da média, $S_{\overline{X}}$, iguala-se ao erro padrão da proporção, S_p.[6] Ou seja, o estudo da *proporção* (seção anterior) é caso particular do estudo da *média*. Vimos, também, que a distribuição amostral de P é aproximadamente normal para amostras grandes. O mesmo acontece com a distribuição amostral de $\overline{X}$.

> Para amostras aleatórias grandes ($n \geq 30$), a *distribuição amostral* de $\overline{X}$ é aproximadamente normal.

$\frac{X - M}{\sigma} = z$

$X - M = z\sigma$

$X = M \pm z\sigma$

AMOSTRAS GRANDES

Quando temos uma amostra grande ($n \geq 30$), podemos estimar o *limite superior para o erro amostral* por:

$$E = z \cdot S_{\overline{X}}$$

onde z é obtido conforme indicado na Figura 9.5, em função do *nível de confiança* previamente fixado.

---

[6] O cálculo dos dois erros padrões deve acusar pequena diferença, porque usamos o denominador $n - 1$ no desvio padrão da amostra.

Exemplo 9.2 (continuação) O objetivo é estimar o parâmetro:

μ = ganho médio de peso durante o primeiro ano letivo, na população de crianças da rede municipal de ensino, devido a uma merenda especial.

Numa amostra aleatória simples de $n = 100$ crianças do primeiro ano letivo, em que se estava servindo a merenda especial, foram obtidos os seguintes resultados:

Ganho médio de peso: $\overline{X} = 6{,}0$ kg;
Desvio padrão: S = 2,0 kg.

Procedendo a estimativa do erro padrão de $\overline{X}$:

$$S_{\overline{X}} = \frac{S}{\sqrt{n}} = \frac{2{,}0}{\sqrt{100}} = 0{,}2 \text{ kg}$$

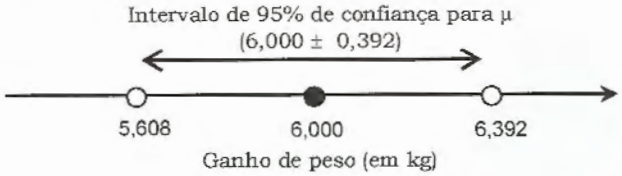
O limite superior para o erro amostral (nível de confiança de 95%):

$$E = (1{,}96)\cdot(0{,}2) = 0{,}392 \text{ kg}$$

donde resulta o seguinte intervalo de 95% de confiança para μ:

$$6{,}000 \pm 0{,}392 \text{ kg}.$$

Ou seja, a partir do acompanhamento da amostra das cem crianças, chegamos à conclusão de que o intervalo de 5,608 a 6,392 kg contém, com 95% de confiança, o ganho médio de peso, μ, de todas as crianças do primeiro ano da rede municipal de ensino, que venham a ser submetidas à merenda especial.[7] Esquematicamente:



Amostras pequenas

Para os casos em que a variável em estudo tiver distribuição razoavelmente simétrica, parecida com uma *normal*, é possível construir estimativas intervalares para a média populacional, μ, mesmo que a amostra seja pequena ($n < 30$). Nesse caso, é necessário usar a chamada *distribuição t* de *Student* (Tabela 5 do Apêndice).

[7] Note que o intervalo de confiança de uma média é apresentado na mesma unidade de medida da variável em estudo.

A *distribuição t* de *Student*, como mostra a Figura 9.6, tem forma parecida com a *normal padrão*, sendo um pouco mais dispersa. Esta dispersão varia com o tamanho da amostra, sendo bastante dispersa para amostras pequenas, mas aproximando-se da normal padrão para amostras grandes. Sua dispersão é função de um parâmetro denominado *graus de liberdade*, *gl*. No problema de estimação de uma média, tem-se: $gl = n - 1$.

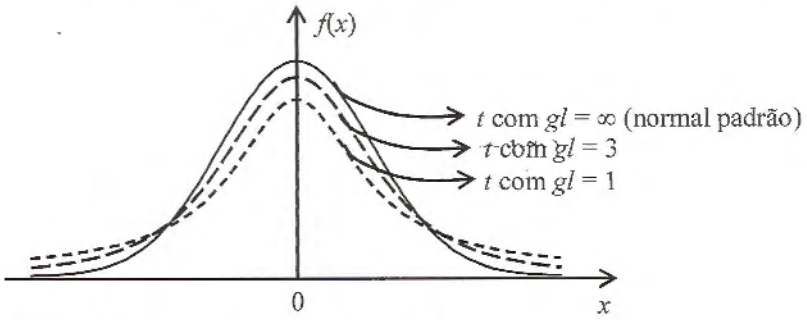


Figura 9.6 Gráficos de distribuições *t* de *Student* e *normal padrão*.

Para obtermos o valor *t* da distribuição *t* de *Student*, basta calcular os graus de liberdade: $gl = n - 1$; fixar o nível de confiança desejado; e usar a Tabela 5 do apêndice. Por exemplo, para $gl = 9$ e nível de confiança de 95%, devemos usar a Tabela 5, como mostra a Figura 9.7.

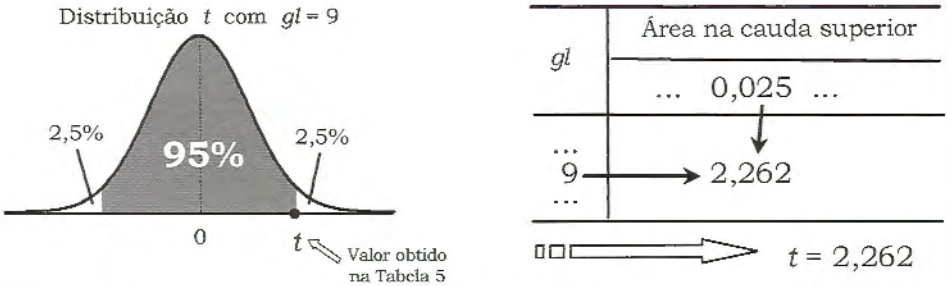


Figura 9.7 Uso da tabela da distribuição *t* de *Student*. Ilustração com $gl = 9$ e nível de confiança de 95%.

Se $n < 30$ e a variável em estudo tiver distribuição aproximadamente *normal*, podemos estimar o *limite superior para o erro amostral* por:

$$E = t \cdot S_{\overline{X}}$$

onde *t* é obtido na Tabela 5 com $gl = n - 1$. Para amostras grandes, $t \approx z$, permitindo o uso de qualquer uma das duas distribuições.

Exemplo 9.3 Para verificar a eficácia de um programa de prevenção de acidentes de trabalho, foi realizado um estudo experimental, implementando esse programa em dez empresas da construção civil, escolhidas ao acaso, numa certa região. Os dados abaixo se referem aos *percentuais de redução de acidentes de trabalho* nas dez empresas observadas.

Amostra					Estatísticas	
20	15	23	11	29	Média:	$\overline{X} = 18$
5	20	22	18	17	Desvio padrão:	S = 6,65

O objetivo é estimar o parâmetro:

μ = média da redução percentual de acidentes de trabalho, em todas as empresas da construção civil da região, que venham a ser submetidas ao programa preventivo.

Estimativa do erro padrão de $\overline{X}$:

$$S_{\overline{X}} = \frac{6,65}{\sqrt{10}} = 2,10$$

Usando nível de confiança de 95%, graus de liberdade $gl = 9$ (pois, $n = 10$ e $gl = n - 1$), obtemos na Tabela 5 (apêndice) o valor $t = 2,262$. Assim:

$$E = t \cdot S_{\overline{X}} = (2,262) \times (2,10) = 4,75 \approx 4,8$$

Então, temos o seguinte intervalo de 95% de confiança para o parâmetro μ:

$$18,0 \pm 4,8 \text{ pontos percentuais}^{8}$$

Exercícios

10) Quer se avaliar o tempo médio, μ, que um cliente leva para ser atendido num posto de serviço, no horário de maior movimento. Uma amostra aleatória simples de 14 clientes apontou para os seguintes tempos de espera (em minutos):

15 17 19 10 13 14 20 18 16 15 16 22 13 16

Calcule:

a) a média da amostra;

b) o desvio padrão da amostra;

[8] O intervalo de confiança foi colocado em termos da unidade *pontos percentuais*, porque era esta a unidade dos dados originais (redução *percentual* de acidentes de trabalho).

c) o erro padrão da média amostral;
d) um intervalo de 95% de confiança para μ.

11) A tabela seguinte mostra as médias e os desvios padrões da renda familiar, calculados com base em uma amostra de 120 famílias, estratificada em três localidades. Essa tabela foi construída com os dados do anexo do Capítulo 4.

Localidade	Tamanho da amostra	Renda familiar (sal. mín.)	
		média	desvio padrão
Monte Verde	40	8,1	4,3
Pq. da Figueira	42	5,8	2,6
Encosta do Morro	37	5,0	4,5

Construa um intervalo de confiança, ao nível de confiança de 95%, para a renda familiar média de cada localidade. Interprete as estimativas.

12) Suspeita-se que um certo fiscal tende a favorecer os devedores, atribuindo multas mais leves. Fazendo-se uma auditoria numa amostra aleatória de oito empresas, verificaram-se os seguintes valores que deixaram de ser cobrados, em reais:

200 340 180 0 420 100 460 340

a) Apresente um intervalo de 95% de confiança para o parâmetro μ.
b) Qual é o significado, no presente problema, do parâmetro μ?
c) Interprete a estimativa do item (a).

13) Considerando a amostra do Exercício 2, construa um intervalo de 99% de confiança para o número médio de cômodos por domicílio, no bairro em estudo. Verifique se o parâmetro μ, calculado no Exercício 1, pertence a este intervalo.

14) Considere as informações do anexo do Capítulo 2. Selecione uma amostra aleatória simples de 10 alunos e observe os dados relativos à variável *desempenho no curso*. Com esta amostra, faça os seguintes itens:

a) Apresente um intervalo de 90% de confiança para o parâmetro μ.
b) Qual é o significado do parâmetro μ, neste caso?
c) Interprete a estimativa do item (a).
d) Usando toda a população, calcule o parâmetro μ e verifique se o intervalo que você construiu no item (a) contém este parâmetro. Consulte seus colegas de sala. Verifique quantos obtiveram intervalos de confiança contendo o parâmetro μ.

9.4 Correções para tamanho de população conhecido

O leitor pode estar estranhando que, na avaliação da precisão das estimativas, o tamanho N da população não tenha sido considerado. Na verdade, o conhecimento deste valor só é relevante em populações pequenas. Neste caso, basta fazer as seguintes mudanças nas estimativas dos erros padrões de P e $\overline{X}$:

$$S_p = \sqrt{\frac{P \cdot (1-P)}{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

$$S_{\bar{x}} = \frac{S}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

O restante dos cálculos dos intervalos de confiança mantém-se inalterado. Cabe também observar que se N for muito grande (digamos, mais que vinte vezes o tamanho da amostra), então o segundo fator das fórmulas acima será aproximadamente igual a um, podendo ser desprezado, resultando nas fórmulas anteriormente apresentadas.

Exemplo 9.4

a) Vamos refazer o Exemplo 9.3, considerando que existam $N = 30$ empresas na região. Neste caso:

$$S_{\bar{x}} = \frac{S}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} = (2,10) \cdot \sqrt{\frac{30-10}{30-1}} = (2,10) \cdot (0,83) = 1,74$$

$$E = t \cdot S_{\bar{x}} = (2,262) \cdot (1,74) \approx 3,9$$

Resultando no seguinte intervalo de 95% de confiança para a média μ:

18,0 ± 3,9 pontos percentuais.

b) E se a população fosse constituída de $N = 400$ empresas?

Neste caso:

$$S_{\bar{x}} = \frac{S}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} = (2,10) \sqrt{\frac{400-10}{400-1}} = (2,10) \cdot (0,99) = 2,08$$

$$E = t \cdot S_{\bar{x}} = (2,262) \cdot (2,08) = 4,7$$

E o intervalo de 95% de confiança para a média m:

18,0 ± 4,7 pontos percentuais.

■

Comparando os resultados dos Exemplos 9.3 e 9.4, verificamos que a inclusão do tamanho da população, N, no cálculo do erro padrão, somente acarretou alteração relevante no caso (a). Observe que no caso (b) o tamanho da população é mais que vinte vezes o tamanho da amostra ($N > 20n$). Nesse caso, poderíamos ter usado a fórmula mais simples do erro padrão.

EXERCÍCIOS

15) Numa amostra aleatória simples de 120 domicílios, realizada num certo bairro da cidade, observou-se que apenas 33,3% possuíam instalações sanitárias adequadas. Considerando que existam 460 domicílios no bairro, encontre um intervalo de 95% de confiança para a proporção de domicílios com instalações sanitárias adequadas.

16) Refazer os Exercícios 13 e 14, considerando o tamanho da população.

9.5 TAMANHO MÍNIMO DE UMA AMOSTRA ALEATÓRIA SIMPLES

Na fase de planejamento de pesquisa que envolva um levantamento por amostragem, uma das principais preocupações é o número de elementos que precisarão ser pesquisados (tamanho da amostra, n).

No Capítulo 3, descrevemos algumas técnicas para a seleção de uma amostra e apresentamos uma primeira fórmula para a determinação de seu tamanho. Com a teoria discutida neste capítulo, temos condições de complementar a questão da determinação do tamanho da amostra, supondo o plano de uma amostragem aleatória simples.

As fórmulas para o cálculo do tamanho da amostra são extraídas das expressões dos intervalos de confiança, fixando *a priori* o nível de confiança e o erro amostral tolerado. Suporemos, também, que haja condições para a observação de uma amostra razoavelmente grande, que permita o uso da distribuição normal, na representação das distribuições amostrais de $\overline{X}$ e de P.

Tendo o valor z da distribuição normal, em função do *nível de confiança desejado*, como também E_0 (*erro amostral tolerado*), podemos obter o tamanho da amostra por uma das duas seguintes fórmulas, dependendo se o objetivo final é estimar uma proporção ou uma média:

a) para estimar uma proporção π:

$$n_0 = \frac{z^2 \cdot \pi \cdot (1-\pi)}{E_0^{\,2}}$$

b) para estimar uma média μ:

$$n_0 = \frac{z^2 \cdot \sigma^2}{E_0^{\,2}}$$

Se a população for muito grande (digamos $N > 20n_0$), então n_0 já é o tamanho da amostra:

$$n = n_0$$

Se o tamanho da população for conhecido e não for muito grande, o tamanho da amostra é dado por (expressão aproximada):

$$n = \frac{N \cdot n_0}{N + n_0}$$

Pelas fórmulas apresentadas, podemos observar que, depois de fixado o *nível de confiança* e o *erro tolerável*, o tamanho da amostra depende basicamente da variabilidade da variável em estudo, representada pela sua variância (quadrado do desvio padrão), σ^2. No caso da estimação de uma proporção, a variância é expressa em função do parâmetro π por: $\sigma^2 = \pi \cdot (1 - \pi)$.

Como o parâmetro σ^2 aparece no numerador das expressões do cálculo de n, concluímos que, quanto mais heterogênea for a população em estudo, maior deverá ser o tamanho da amostra.

Uma dificuldade existente na fase do planejamento amostral de uma pesquisa é que o parâmetro σ^2 é, em geral, desconhecido. Apresentaremos duas sugestões para contornar este problema: (1) observação empírica e (2) argumentos teóricos.

Observação empírica

Podemos usar no lugar de σ^2 uma estimativa, S_0^2, obtida de algum estudo anterior ou de uma amostra piloto, isto é, uma pequena amostra realizada na fase de planejamento da pesquisa, com propósitos de avaliar o instrumento (questionário), treinar pesquisadores ou obter alguma estimativa inicial da população.

Exemplo 9.5 Considere, novamente, o problema de estimar o ganho médio de peso das crianças da rede municipal de ensino, durante o primeiro ano letivo (Exemplo 9.2). Suponha que um estudo similar tenha sido realizado num outro município, onde observaram uma amostra de 80 crianças, que resultou num desvio padrão igual a 1,95 kg. Fixando o nível de confiança em 95%, e tolerando um erro amostral de até 200 gramas (isto é, E_0 = 0,2 kg), qual deve ser o tamanho da amostra?

Solução: Nível de confiança de 95% acarreta z = 1,96 (ver Figura 9.5). Usaremos, no lugar de σ^2, o valor da variância da amostra do outro município: $S_0^2 = (1{,}95)^2 = 3{,}8$. Assim, o tamanho mínimo de uma amostra aleatória simples é:

$$n_0 = \frac{z^2 \cdot \sigma^2}{E_0^2} \approx \frac{z^2 \cdot S^2}{E_0^2} = \frac{(1{,}96)^2 \cdot (3{,}8)}{(0{,}2)^2} = 365$$

Como N é desconhecido, este já é o tamanho da amostra ($n = n_0 = 365$ crianças).

■

É comum, no cálculo do tamanho da amostra, aproximar o valor $z = 1{,}96$ para $z = 2$, pois, além de facilitar as contas, compensa, em termos, o erro introduzido pela substituição de σ^2 por S_0^2. No Exemplo 9.5, usando $z = 2$, obtemos como resultado: $n = 380$ crianças. No caso de se usar uma amostra piloto pequena, digamos, de tamanho $m < 30$, é melhor substituir z por t com $gl = m - 1$.

ARGUMENTOS TEÓRICOS: O CASO DE ESTIMAÇÃO DE PROPORÇÕES

Muitas vezes, pela forma de mensuração da variável, é possível obter alguma avaliação sobre σ^2, ou, pelo menos, algum limite superior para este parâmetro. Uma situação particularmente interessante é na estimação de uma proporção π. Neste caso, a variância pode ser expressa em termos do parâmetro π, da seguinte forma:

$$\sigma^2 = \pi \cdot (1 - \pi) = \pi - \pi^2$$

Ou seja, σ^2 é uma função de segundo grau de π, cujo gráfico (parábola) é mostrado na Figura 9.8. Observe que o valor máximo de σ^2 ocorre quando $\pi = ½$. Nesse caso, $\sigma^2 = ½ \cdot ½ = ¼$.

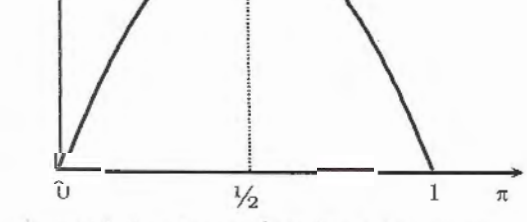


Figura 9.8 O parâmetro σ^2 em função da proporção π.

Nos problemas de estimação de uma proporção, em que não temos qualquer avaliação inicial sobre π, ou quando acreditamos que a proporção π esteja próxima de $^1/_2$, podemos usar, no lugar de σ^2, o seu valor máximo, $^1/_4$. Assim,

$$n_0 = \frac{z^2 \cdot \frac{1}{4}}{E_0^2} = \frac{z^2}{4 \cdot E_0^2}$$

Em pesquisas de levantamento por amostragem, normalmente queremos estimar várias proporções (vários parâmetros π), com dada margem de erro, E_0. A expressão precedente garante a precisão estabelecida, E_0, para as várias estimativas. Nesta expressão, se usarmos o nível usual de confiança de 95%, temos $z \approx 2$. Então, a fórmula do tamanho da amostra para várias proporções é:

$$n_0 = \frac{1}{E_0^2}$$

Exemplo 9.6 Com o objetivo de avaliar a preferência do eleitor na véspera de uma eleição para a prefeitura de um município, planeja-se um levantamento por amostragem aleatória simples. Considere que seja admissível um erro amostral de até 2%, com 95% de confiança, para as estimativas dos percentuais dos vários candidatos. Quantos eleitores devem ser pesquisados?

Solução: $n_0 \approx \frac{1}{E_0^2} = \frac{1}{(0{,}02)^2} = 2.500$

Como N é desconhecido, este já é o tamanho da amostra ($n = n_0 = 2.500$ eleitores).

■

Exemplo 9.7 Numa pesquisa epidemiológica deseja-se estimar, com 90% de confiança, o parâmetro:

π = proporção de pessoas infectadas

com erro amostral máximo de 1%. Qual deve ser o tamanho de uma amostra aleatória simples, supondo que, na população em estudo, não existam mais que 20% de indivíduos infectados?

Solução: Dada a informação que $\pi \leq 0{,}20$, então o valor máximo de σ^2 é:

$$\pi \cdot (1-\pi) = (0{,}20) \cdot (1-0{,}20) = 0{,}16$$

Assim,

$$n = n_0 = \frac{z^2 \cdot \pi(1-\pi)}{E_0^2} \approx \frac{(1{,}645)^2 \cdot (0{,}16)}{(0{,}01)^2} = 4.330$$

Exercícios

17) Com o objetivo de estimar o tempo médio de um caixa eletrônico para atender um cliente, planeja-se fazer um levantamento por amostragem. Qual deve ser o tamanho de uma amostra aleatória simples de clientes, para garantir uma estimativa com erro não superior a 2 segundos, ao nível de confiança de 95%? Suponha que se verificou, através de estudos anteriores, que o desvio padrão não passa de 8 segundos.

18) Deseja-se estudar as percentagens de ocorrências de diversos atributos, numa comunidade de 600 famílias. Qual deve ser o tamanho de uma amostra aleatória simples, considerando erro máximo de 4% e nível de confiança de 95%?

Exercícios complementares

19) Nas situações descritas abaixo, descreva qual é a população, a amostra, o parâmetro de interesse e uma estatística que pode ser usada para estimar o parâmetro.
   a) Para avaliar a proporção de alunos do Curso de Administração favoráveis à eliminação da disciplina de Estatística do currículo, selecionou-se aleatoriamente 80 alunos do Curso.
   b) Para avaliar a eficácia de um curso que orienta como fazer boa alimentação e exercícios físicos, selecionou-se uma amostra aleatória de 20 pessoas obesas de uma certa cidade.
   c) Para avaliar uma campanha contra o fumo, conduzida pela prefeitura de uma cidade, acompanhou-se uma amostra aleatória de 100 fumantes.

20) Um instituto de pesquisa observou uma amostra aleatória de 800 habitantes de uma grande cidade. Verificou que 320 indivíduos desta amostra apoiam a administração da prefeitura, enquanto que os outros 480 a criticam.
   a) O que se pode dizer sobre a percentagem de indivíduos que apoiam a administração da prefeitura, dentre os indivíduos da amostra?
   b) O que se pode dizer sobre a percentagem de indivíduos que apoiam a administração da prefeitura, dentre os habitantes da cidade?

Obs.: Em caso de estimativa, usar nível de confiança de 95%.

21) Com o objetivo de avaliar a aceitação de um novo produto no mercado, planeja-se fazer um levantamento amostral para estimar a proporção de futuros consumidores desse produto.
   a) Qual deve ser o tamanho de uma amostra aleatória simples, que garanta uma estimativa com erro máximo de 5% e nível de confiança de 99%?
   b) Efetuou-se a amostragem conforme o tamanho calculado no item (a). Foi verificado que 200 pessoas desta amostra passariam a usar regularmente o produto. Construa um intervalo de 99% de confiança para o parâmetro de interesse. Interprete o resultado.

22) Numa pesquisa realizada sobre uma amostra de 647 adolescentes em Santa Catarina, 88 responderam que se sentiam frustrados sexualmente. Admitindo que a amostragem tenha sido aleatória, construa um intervalo de 95% de

confiança para o percentual de adolescentes catarinenses que se dizem frustrados sexualmente.

23) Numa amostra aleatória de 12 estudantes do Curso de Administração, que contém cerca de 500 alunos, levantou-se o grau de satisfação do aluno com o Curso, numa escala de 1 a 5. Os resultados foram os seguintes:

2 2 3 3 3 3 4 4 4 4 5 5

a) Construa um intervalo de 95% de confiança para o nível médio de satisfação dos alunos com o Curso.
b) Admitindo que a amostra do item anterior era apenas um estudo piloto, qual deve ser o tamanho de uma amostra aleatória simples para que o erro amostral não seja superior a 0,2 unidade, com 95% de confiança?

24) Para verificar a eficácia de uma dieta de emagrecimento, realizou-se um experimento com 10 indivíduos, que se submeteram à dieta por um período de um ano. A variação de peso de cada indivíduo, medido em kg, é apresentada abaixo.

−5 −10 5 −20 −8 10 0 −2 −8 −1

a) Calcule a média, mediana e desvio padrão da amostra.
b) Construa um intervalo de 95% de confiança para o parâmetro μ, sendo μ a redução de peso esperada em um ano de dieta.
c) Considerando o resultado do item anterior, você pode afirmar, com nível de confiança de 95%, que a dieta em questão realmente tende a emagrecer os indivíduos?

25) Uma empresa tem 2.400 empregados. Deseja-se extrair uma amostra de empregados para verificar o grau de satisfação em relação à qualidade da comida no refeitório. Em uma amostra piloto, numa escala de 0 a 10, obteve-se para o grau de satisfação nota média igual a 6,5 e desvio padrão igual a 2,8.
a) Determine o tamanho mínimo da amostra, supondo um planejamento por amostragem aleatória simples, com erro máximo de 0,5 unidade e nível de confiança de 99%.
b) Considere que a amostra planejada no item anterior tenha sido realizada. A média dos dados da amostra foi 5,3 e o desvio padrão foi 2,6 pontos. Faça um intervalo de 99% de confiança para o parâmetro μ.
c) Considerando o resultado do item anterior, você pode afirmar, com nível de confiança de 99%, que se a pesquisa fosse aplicada nos 2.400 funcionários, a nota média seria superior a cinco? Justifique.
d) Realizada a amostra planejada no item (a), suponha, agora, que 120 atribuíram notas iguais ou superiores a cinco. Apresente um intervalo de 90% de confiança para a percentagem de indivíduos da população que atribuiriam notas iguais ou superiores a cinco.

26) Uma pesquisa realizada por pesquisadores da Universidade Federal de Minas Gerais, que se baseou em amostras de sangue de 250 pessoas brancas das regiões Norte, Nordeste, Sudeste e Sul, concluiu que por parte das ancestrais mulheres, 39% da herança genética dos brancos é europeia, 28% é negra e 33% é indígena.[9] Supondo que a amostragem tenha sido aleatória, qual a margem de erro de cada uma dessas estimativas, considerando nível de confiança de 95%?

[9] Divulgado no Jornal Hoje – Rede Globo, em 18 abr. 2000.

Capítulo 10

Testes estatísticos de hipóteses

Muitas vezes o pesquisador tem alguma ideia ou conjetura sobre o comportamento de uma variável, ou de uma possível associação entre variáveis. Neste caso, o planejamento da pesquisa deve ser de tal forma que permita, com os dados amostrais, testar a veracidade de suas ideias sobre a população em estudo. Adotamos que a população seja o mundo real e as ideias sejam as hipóteses de pesquisa que poderão ser testadas por técnicas estatísticas denominadas *testes de hipóteses* ou *testes de significância.*

Exemplo 10.1

a) Na problemática de verificar se existe relação entre tabagismo e sexo, em certa região, pode-se lançar a seguinte hipótese: *Na região em estudo, a propensão de fumar nos homens é diferente da que ocorre nas mulheres.*

b) Para se verificar o efeito de uma propaganda nas vendas de certo produto, tem-se interesse em verificar a veracidade da hipótese: *A propaganda produz um efeito positivo nas vendas.*

c) Na condução de uma política educacional, pode-se ter interesse em comparar dois métodos de ensino. Hipótese: *Os métodos de ensino tendem a produzir resultados diferentes de aprendizagem.*

■

Para verificar estatisticamente a veracidade de uma hipótese, precisamos de um conjunto de dados, observados adequadamente na população em estudo.

Antes de executar a coleta dos dados, torna-se fundamental fixar claramente a população a ser estudada, bem como a maneira pela qual se vai observar as variáveis descritas nas hipóteses. Por exemplo, numa hipótese de associação entre sexo e tabagismo, devemos definir a região de abrangência da pesquisa ou, mais precisamente, a *população* a ser estudada. Também devemos estabelecer uma forma de *medir* a variável *tabagismo*. Uma maneira razoavelmente simples de mensurar *tabagismo* é, a partir de critérios previamente estabelecidos, classificar os indivíduos em *fumantes* e *não fumantes*, gerando dados categorizados.

A Tabela 10.1 apresenta os resultados da classificação de 300 indivíduos, selecionados aleatoriamente de uma determinada população, segundo o sexo (*masculino* ou *feminino*) e tabagismo (*fumante* ou *não fumante*).

Tabela 10.1 Distribuição de 300 pessoas, classificadas segundo o sexo e tabagismo

Tabagismo	Sexo		Total
	masculino	feminino	
fumante	92 (46%)	38 (38%)	130 (43%)
não fumante	108 (54%)	62 (62%)	170 (57%)
Total	200 (100%)	100 (100%)	300 (100%)

Na amostra, a percentagem de homens fumantes (46%) é diferente da percentagem de mulheres fumantes (38%); os dados parecem comprovar a hipótese de que existe diferença entre homens e mulheres, quanto à variável tabagismo. Contudo, não devemos nos esquecer de que estamos examinando uma amostra e, consequentemente, as diferenças observadas podem ter ocorrido por fatores casuais, de tal forma que, se tomássemos outras amostras da mesma população, sob as mesmas condições, as conclusões poderiam ser diferentes.

A aplicação de um teste estatístico (ou teste de significância) serve para verificar se os dados fornecem evidência suficiente para que se possa aceitar como verdadeira a hipótese de pesquisa, precavendo-se, com certa segurança, de que as diferenças observadas nos dados não são meramente casuais.

10.1 As Hipóteses de um Teste Estatístico

Dado um problema de pesquisa, o pesquisador precisa saber escrever a chamada **hipótese de trabalho** ou **hipótese nula**, H_0. Essa hipótese é

descrita em termos de parâmetros populacionais e é, basicamente, uma negação daquilo que o pesquisador deseja provar. Sob essa hipótese, as diferenças observadas nos dados são consideradas casuais.

Exemplo 10.1 (continuação) Podemos ter as seguintes hipóteses nulas para os problemas descritos anteriormente.

a) H_0: A proporção de homens fumantes *é igual* à proporção de mulheres fumantes, na população em estudo.

b) H_0: Em média, as vendas *não aumentam* com a introdução da propaganda.

c) H_0: Em média, os dois métodos de ensino produzem *os mesmos* resultados.

■

Quando os dados mostrarem evidência suficiente de que a hipótese nula, H_0, é falsa, o teste a rejeita, aceitando em seu lugar a chamada **hipótese alternativa**, H_1. A hipótese alternativa é, em geral, aquilo que o pesquisador quer provar, ou seja, a própria hipótese de pesquisa, considerando a forma do planejamento da pesquisa.

Exemplo 10.1 (continuação) As hipóteses alternativas.

a) H_1: A proporção de homens fumantes *é diferente* da proporção de mulheres fumantes, na população em estudo.

b) H_1: Em média, as vendas *aumentam* com a introdução da propaganda.

c) H_1: Em média, os dois métodos de ensino produzem resultados *diferentes*.

■

É comum H_0 ser apresentada em termos de *igualdade* de parâmetros populacionais, enquanto H_1 em forma de *desigualdade* (maior, menor ou diferente).

No Exemplo 10.1, item (a), H_0 é descrita em termos de igualdade de duas proporções (H_0: $\pi_h = \pi_m$, onde π_h é a proporção de homens fumantes e π_m é a proporção de mulheres fumantes, na população em estudo). Por outro lado, a hipótese alternativa pode ser escrita como H_1: $\pi_h \neq \pi_m$. Já no item (b), as hipóteses podem ser escritas em termos de médias da seguinte maneira: H_0: $\mu_c = \mu_s$ e H_1: $\mu_c > \mu_s$, onde μ_c é o valor médio das vendas com propaganda e μ_s é o valor médio das vendas sem propaganda. E em (c)?

EXEMPLO 10.2 Suponha, por exemplo, que se suspeite que uma certa moeda, usada num jogo de azar, é *viciada*, isto é, há uma tendência de ocorrerem mais caras do que coroas, ou mais coroas do que caras. Entendendo-se como *moeda honesta* aquela que tem a mesma probabilidade de dar cara e coroa, podemos formular as hipóteses da seguinte maneira:

H_0: a moeda é honesta e H_1: a moeda é viciada

Se chamarmos π à probabilidade de ocorrer cara num lançamento dessa moeda, podemos escrever:

H_0: $\pi = 0,5$ e H_1: $\pi \neq 0,5$

■

10.2 CONCEITOS BÁSICOS

Usaremos o Exemplo 10.2 para apresentar alguns conceitos.

Suponhamos, inicialmente, H_0 como verdadeira. H_0 somente vai ser rejeitada em favor de H_1, se houver evidência suficiente que a contradiga. A existência dessa possível evidência será verificada num conjunto de observações relativas ao problema em estudo (amostra). No presente exemplo, a amostra consiste de n lançamentos imparciais da moeda.

Em cada lançamento da moeda, observamos um resultado: *cara* ou *coroa*. Ao observar n lançamentos, podemos computar o valor da estatística:

Y = número total de caras nos n lançamentos

A estatística Y poderá ser usada na definição de um critério de decisão:

Aceitar H_0 ou
Rejeitar H_0 em favor de H_1.

Neste contexto, a estatística Y é chamada de **estatística do teste**. Sejam $n = 10$ lançamentos e as duas seguintes amostras:

AMOSTRA A – Suponha que nos 10 lançamentos observamos $Y = 10$ caras. Podemos rejeitar H_0 em favor de H_1?
AMOSTRA B – E se tivéssemos observado $Y = 7$ caras?

Na amostra A, é intuitivo que existe mais evidência para rejeitar H_0. Contudo, em nenhuma das duas situações podemos rejeitar H_0 com a certeza de que H_0 é falsa, pois estamos trabalhando com um fenômeno aleatório, em que é plenamente possível nos 10 lançamentos de uma moeda

sabidamente honesta (H_0 verdadeira), ocorrerem 7, 8, 9 ou até mesmo 10 caras! Por outro lado, se a ocorrência de um certo resultado for muito pouco provável para uma moeda honesta, é natural decidirmos por H_1.

Para realizar o teste estatístico, é necessário conhecer a probabilidade de ocorrerem $Y = 10$ caras (amostra A), ou $Y = 7$ caras (amostra B), em 10 lançamentos de uma moeda honesta. Mais geralmente, precisamos da distribuição de probabilidades da estatística do teste Y, supondo H_0 verdadeira. Esta distribuição de probabilidades será a referência básica para analisarmos o resultado da amostra e decidirmos entre H_0 e H_1.

Distribuição de referência

No exemplo em questão, Y tem distribuição binomial com parâmetros $n = 10$ e $\pi = 0{,}5$ (supondo H_0 verdadeira). Esta será a *distribuição de referência* para o valor calculado da estatística do teste, Y. A Figura 10.1 apresenta a distribuição de referência do presente teste, sob forma gráfica. As probabilidades, p(y), foram obtidas na tabela da distribuição binomial (Tabela 2 do apêndice). Para facilitar a exposição, essas probabilidades foram arredondadas para três decimais.

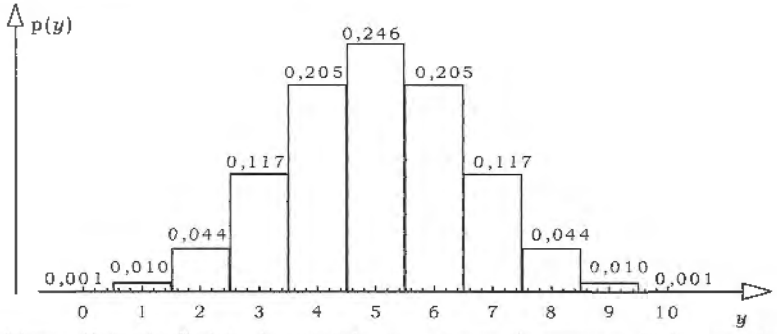


Figura 10.1 Distribuição da estatística Y = *número de caras em 10 lançamentos da moeda*, sob H_0 (binomial com $n = 10$ e $\pi = 0{,}5$).

Com a distribuição de probabilidades da estatística do teste, podemos avaliar melhor a adequação de H_0 com o resultado de Y, calculado com base na amostra. A Figura 10.1 mostra que se H_0 for verdadeira, os resultados mais prováveis estão em torno de 5 caras. Chamaremos este valor central da distribuição de probabilidades de **valor esperado** ou **valor médio**, e o denotaremos por μ.

VALOR p

Supondo H_0 verdadeira, a **probabilidade de significância**, ou **valor *p***, é a probabilidade de a estatística do teste acusar um resultado tão ou mais distante do esperado por H_0, como o resultado da amostra observada.

EXEMPLO 10.3 Retomemos à amostra A, em que observamos $Y = 10$ caras em $n = 10$ lançamentos da moeda em estudo. Considerando o número esperado de caras sob H_0 ($\mu = 5$) como referência, verificamos que tão ou mais distante do que o valor observado na amostra ($Y = 10$), encontra-se o valor 0 e o próprio valor 10, como ilustra a Figura 10.2.

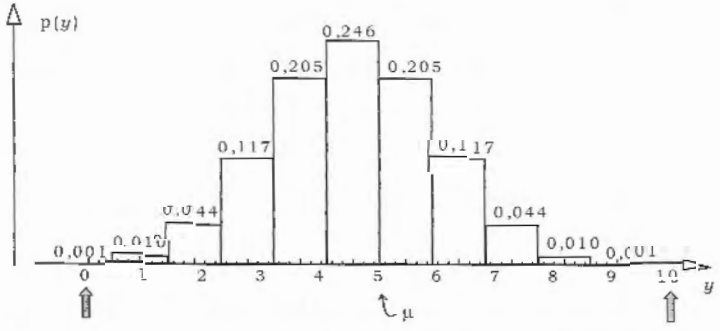


Figura 10.2 Distribuição de Y, sob H_0. As setas indicam os valores que distam do esperado, $\mu = 5$, tão ou mais do que o valor $Y = 10$, observado na amostra A.

Consequentemente, a probabilidade de significância será:

$$p = p(0) + p(10) = 0{,}001 + 0{,}001 = 0{,}002 \text{ (ou } 0{,}2\%)$$

Ou seja, para uma moeda honesta (H_0 verdadeira), tem-se a pequena probabilidade $p = 0{,}002$ de ocorrer um resultado tão ou mais distante do valor esperado, como o que, de fato, ocorreu neste caso ($Y = 10$ caras). Como $p = 0{,}002$ é uma probabilidade muito pequena, é natural rejeitar a hipótese de que a moeda é honesta (H_0), decidindo-se pela hipótese de que a moeda é viciada (H_1).

Os dados mostram evidência suficiente para dizer que a moeda é viciada!

■

EXEMPLO 10.4 Vejamos, agora, a amostra B, em que observamos $Y = 7$ caras em $n = 10$ lançamentos. Nesta amostra, tão ou mais distante do que o valor $Y = 7$ são encontrados os valores: 7, 8, 9, 10, 0, 1, 2 e 3, como ilustra a Figura 10.3.

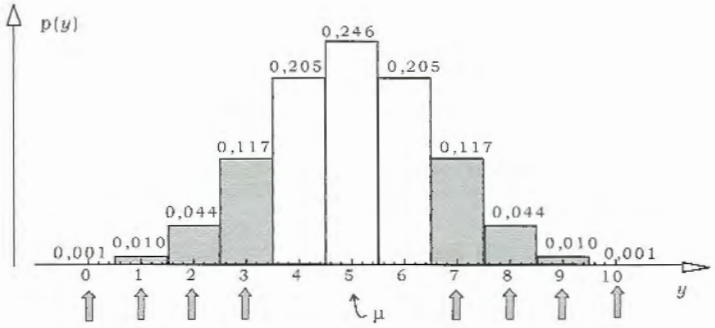


Figura 10.3 Distribuição de Y, sob H_0. As setas indicam os valores que distam do esperado, $\mu = 5$, tão ou mais do que o valor $Y = 7$, observado na amostra B.

Temos, então, a seguinte probabilidade de significância:

$p = p(0) + p(1) + p(2) + p(3) + p(7) + p(8) + p(9) + p(10) =$
$= 0{,}001 + 0{,}010 + 0{,}044 + 0{,}117 + 0{,}117 + 0{,}044 + 0{,}010 + 0{,}001 =$
$= 0{,}344$ (ou, 34,4%).

Esta segunda situação mostra que, para uma moeda honesta (H_0 verdadeira), tem-se a probabilidade $p = 0{,}344$ de ocorrer um resultado tão ou mais distante do valor esperado, como o que, de fato, ocorreu neste caso ($Y = 7$ caras). Como $p = 0{,}344$ não é uma probabilidade desprezível, é mais prudente não rejeitar H_0.

Não há evidência suficiente para afirmar que a moeda é viciada!

■

O ***valor p*** aponta o quão *estranho* foi o resultado da amostra, se supusermos H_0 a hipótese verdadeira.

Quanto menor for o valor p, maior a evidência para rejeitar H_0. O valor p também pode ser interpretado como o risco de se tomar a decisão errada após a observação da amostra, caso se rejeite H_0. Por exemplo, se afirmássemos que a moeda é viciada com a evidência de $Y = 7$ caras, em $n = 10$ lançamentos, estaríamos incorrendo num risco de 34,4% de estarmos fazendo uma afirmação errada.

Nível de significância

Ainda na fase do planejamento de uma pesquisa, quando desejamos confirmar ou refutar alguma hipótese, é comum estabelecer o valor da probabilidade tolerável de incorrer no erro de rejeitar H_0, quando H_0 é

verdadeira. Este valor é conhecido como **nível de significância do teste** e é designado pela letra grega α. Em pesquisa social, é comum adotar nível de significância de 5%, isto é, $\alpha = 0,05$.

Estabelecido o nível de significância α, tem-se a seguinte regra geral de decisão de um teste estatístico:

$p > \alpha \rightarrow$ aceita H_0
$p \leq \alpha \rightarrow$ rejeita H_0, em favor de H_1

Exemplo 10.3 (continuação) Seja o nível de significância de 5% ($\alpha = 0,05$). Na amostra A, quando observamos dez caras em dez lançamentos, o teste estatístico *rejeita* H_0, *em favor de* H_1 (pois a probabilidade de significância, calculada com base na amostra, foi $p = 0,002$ e, portanto, *menor* do que o valor adotado para α).

■

Exemplo 10.4 (continuação) Seja $\alpha = 0,05$. Na amostra B, quando observamos sete caras em dez lançamentos, o teste estatístico *não* rejeita H_0, porque a probabilidade de significância, calculada com base na amostra, foi $p = 0,344$; que *não é menor* do que o valor adotado para α.

■

Quando o teste *rejeita* H_0 em favor de H_1 ($p \leq \alpha$), a probabilidade de se estar tomando a decisão errada é, no máximo, igual ao nível de significância α adotado. Desta forma, temos certa garantia da veracidade de H_1.

Uma interpretação um pouco diferente é dada quando o teste *aceita* a hipótese nula H_0 ($p > \alpha$). Neste caso, podemos dizer: *os dados estão em conformidade com a hipótese nula!* Isto não implica, contudo, que H_0 seja realmente a hipótese verdadeira, mas que os dados não mostraram evidência suficiente para rejeitá-la e, por isso, continuamos acreditando em sua veracidade. Conforme Ronald A. Fisher, conhecido como o pai da estatística experimental (FISHER, 1956, p.16):

> A hipótese nula pode ou não ser impugnada pelos resultados de um experimento. Ela nunca pode ser provada, mas pode ser desaprovada no curso da experimentação.

Estabelecido um nível de significância α antes da observação dos dados, temos as seguintes possibilidades:

Realidade (desconhecida)	Decisão do teste	
	Aceita H_0	Rejeita H_0
H_0 verdadeira	Decisão correta	Erro tipo I (Probab. = α)
H_0 falsa	Erro tipo II (Probab. = β)	Decisão correta

Observamos no esquema que, se o teste *rejeitar* H_0, temos controle do risco de erro (probabilidade igual a α). Por outro lado, se o teste *aceitar* H_0, não temos controle do risco de erro. No esquema, representamos a probabilidade de ocorrer o erro tipo II como β, mas, ao contrário de α, a probabilidade β não é fixada *a priori*. Em razão disso, estamos usando uma linguagem mais enfática quando o teste rejeita H_0 (p. ex., *os dados provaram estatisticamente que a moeda é viciada*) e uma linguagem mais suave quando o teste aceita H_0 (p. ex., *os dados não mostraram evidência suficiente de que a moeda é viciada, portanto admite-se que ela é honesta*).

Exercícios

1) Seja π a probabilidade de cara de uma certa moeda. Sejam H_0: $\pi = 0,5$ e H_1: $\pi \neq 0,5$. Lança-se 12 vezes esta moeda, observando-se o número de caras. Usando a tabela da distribuição binomial (Tabela 2 do Apêndice), obtenha a probabilidade de significância para cada um dos seguintes resultados:
   a) 1 cara; b) 4 caras e c) 11 caras.
2) Adotando o nível de significância de 5%, qual é a conclusão do teste em cada item do Exercício 1.
3) É possível, para uma mesma amostra, aceitar H_0 ao nível de significância de 1%, mas rejeitá-la ao nível de 5%? E o inverso? Exemplifique.

10.3 Testes unilaterais e bilaterais

No teste discutido no tópico anterior, a rejeição de H_0: $\pi = 0,5$, em favor de H_1: $\pi \neq 0,5$, se dá tanto quando ocorre um valor muito pequeno, quanto muito grande de caras. Essa é uma situação típica de *teste bilateral*.

Existem situações em que pretendemos rejeitar H_0 somente num dos sentidos. Por exemplo, se suspeitamos que a moeda tende a dar mais caras do que coroas, então, sendo π a probabilidade de ocorrer cara, o teste pode ser formulado da seguinte maneira:

H_0: $\pi = 0{,}5$ (a moeda é honesta) e
H_1: $\pi > 0{,}5$ (a moeda tende a dar mais caras do que coroas).

Com essas hipóteses, só faz sentido rejeitar H_0, em favor de H_1, se na amostra ocorrer um número significativamente maior de caras do que de coroas, resultando no que chamamos de *teste unilateral*.

> Um teste pode ser unilateral ou bilateral, dependendo do problema em estudo. Nos testes *unilaterais*, a probabilidade de significância é computada em apenas *um dos lados* da distribuição de referência.

EXEMPLO 10.5 Considere que, para testar H_0: $\pi = 0{,}5$ contra H_1: $\pi > 0{,}5$, tenhamos lançado a moeda $n = 10$ vezes e observado $Y = 7$ caras. A probabilidade de significância será:

$$p = p(7) + p(8) + p(9) + p(10) = 0{,}117 + 0{,}044 + 0{,}010 + 0{,}001 = 0{,}172$$

que corresponde à metade da probabilidade de significância do teste bilateral, discutido no Exemplo 10.4. Com o nível de significância de 5%, o teste *não* rejeita H_0 (pois, $p > \alpha$). Veja a Figura 10.4.

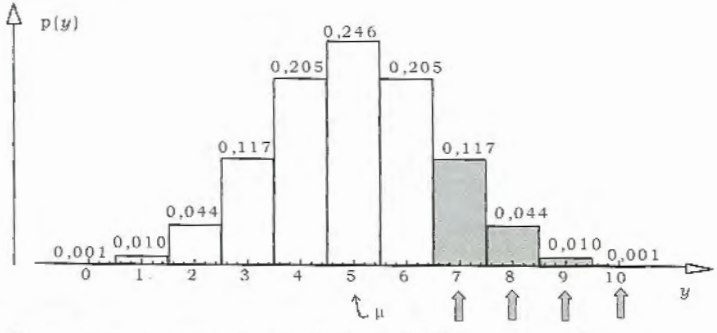


Figura 10.4 Ilustração do cálculo da probabilidade de significância de um teste unilateral (Exemplo 10.5).

■

EXEMPLO 10.6 (TEIXEIRA; MEINERT; BARBETTA, 1987, p.137) Com o objetivo de testar se a diferença de odor em sorvetes de morango é percebida por degustadores, efetuou-se o seguinte experimento: para cada um dos 8 (oito) degustadores selecionados para o experimento, foram dadas, em ordem aleatória e sem identificação, duas amostras de sorvete, sendo uma com odor mais forte e outra normal. As amostras de sorvete foram elaboradas de forma tão similar quanto possível, com exceção da intensidade de odor, que é a característica em estudo.

Chamando de π a probabilidade de o degustador acusar corretamente a amostra de sorvete com odor mais intenso, temos interesse em testar as seguintes hipóteses.

H_0: $\pi = 0,5$ (o degustador *chuta* a resposta, isto é, o odor mais intenso não é detectado) e

H_1: $\pi > 0,5$ (existe uma tendência de o degustador perceber o sorvete que tem o odor mais intenso).

Seja Y o número de degustadores que indicam corretamente o sorvete com odor mais intenso. Pelas características do experimento, podemos deduzir que se H_0 for correta, a estatística Y tem distribuição binomial com $n = 8$ e $\pi = 0,5$.

Os resultados do experimento mostraram que dos oito degustadores, seis indicaram corretamente o sorvete de odor mais intenso ($Y = 6$). Usando a distribuição binomial (Tabela 2 do apêndice), podemos computar a probabilidade de significância:

$$p = p(6) + p(7) + p(8) = 0,109 + 0,031 + 0,004 = 0,144$$

Assim, se estamos trabalhando com o nível de significância de 5% ($\alpha = 0,05$), a hipótese nula *não* pode ser rejeitada. Portanto, concluímos que os dados resultantes do experimento são insuficientes para se afirmar que a diferença de odor em sorvetes de morango seja percebida pelos degustadores.

Exercícios

4) Para cada um dos itens do Exemplo 10.1, descrever qual abordagem (unilateral ou bilateral) é mais apropriada.

5) Seja π a probabilidade de cara de uma certa moeda. Sejam H_0: $\pi = 0,5$ e H_1: $\pi < 0,5$. Lança-se 12 vezes esta moeda, observando-se o número de caras. Usando a tabela da distribuição binomial (Tabela 2 do apêndice), obtenha a probabilidade de significância para cada um dos seguintes resultados: a) 1 cara; b) 4 caras; e c) 6 caras.

Usando nível de significância de 5%, em quais casos acima o teste rejeita H_0?

10.4 Uso de distribuições aproximadas

Os exemplos de testes de hipóteses discutidos até aqui usavam amostras de tamanho pequeno, o que permitia o uso da tabela da distribuição binomial no cálculo das probabilidades de significância. Em

experimentos binomiais, quando o tamanho da amostra, n, for grande, a probabilidade de significância pode ser obtida, de forma aproximada, pela distribuição normal de parâmetros:[1]

$$\mu = n \cdot \pi \qquad \text{e} \qquad \sigma = \sqrt{n \cdot \pi \cdot (1 - \pi)}$$

EXEMPLO 10.7 Considere que, para testar H_0: $\pi = 0{,}5$ contra H_1: $\pi > 0{,}5$, onde π é a probabilidade de *cara* de uma certa moeda, tenham sido realizados $n = 40$ lançamentos, dos quais 28 deram caras. Este resultado leva à rejeição de H_0, em favor de H_1, ao nível de significância de 5%?

Solução: Como n é grande, vamos calcular a probabilidade de significância pela distribuição normal. Levando-se em conta que o teste é unilateral (H_1: $\pi > 0{,}5$), a probabilidade de significância vai se identificar com uma área na cauda superior da curva normal. Considerando o resultado observado, $Y = 28$ caras, e aplicando a correção de continuidade (Seção 8.4, Capítulo 8), a probabilidade de significância corresponde à área acima do ponto 27,5, como ilustra a Figura 10.5.

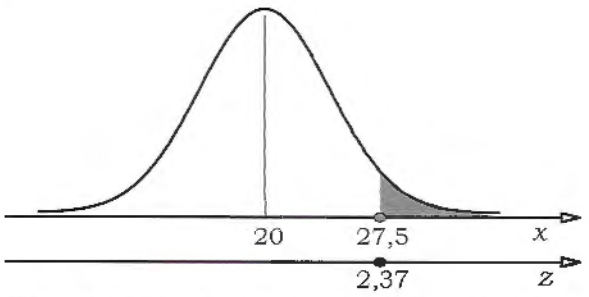


Figura 10.5 Obtenção de uma probabilidade de significância através do modelo normal.

Para realizar o cálculo da área indicada na Figura 10.5, precisamos calcular os parâmetros do modelo normal:

$$\mu = (40) \cdot (0{,}5) = 20 \qquad \text{e} \qquad \sigma = \sqrt{(40) \cdot (0{,}5) \cdot (0{,}5)} = 3{,}16$$

O valor 27,5 da escala original (escala x) corresponde ao seguinte valor padronizado (escala z):

$$z = \frac{x - \mu}{\sigma} = \frac{27{,}5 - 20}{3{,}16} = 2{,}37$$

[1] Vimos no Capítulo 8 que vale a aproximação normal se: (a) $n \cdot \pi \geq 5$ e (b) $n \cdot (1 - \pi) \geq 5$, onde π é o valor declarado em H_0.

Usando a tabela da distribuição normal padrão (Tabela 4 do apêndice), encontramos para $z = 2,37$ a área na cauda superior da curva igual a 0,0089. Temos, então, $p = 0,0089$. Sendo o teste unilateral, este já é o valor p. Como $p = 0,0089$ é menor do que o nível de significância adotado ($\alpha = 0,05$), o teste *rejeita* H_0, concluindo que a moeda tende a dar mais caras do que coroas.

Exercícios

6) Refaça os cálculos do Exercício 1, usando a distribuição normal. Compare os resultados.

7) Seja π a probabilidade de *coroa* de uma certa moeda. Com o objetivo de testar H_0: $\pi = 0,5$ contra H_1: $\pi > 0,5$; fizeram-se 50 lançamentos desta moeda, obtendo-se 31 coroas.
   a) O teste rejeita H_0 ao nível de significância de 5% ($\alpha = 0,05$)?
   b) E se estivéssemos trabalhando com o nível de significância de 1% ($\alpha = 0,01$)?

8) (LEVIN, 1985, p. 274) Para testar se consumidores habituais de determinada margarina eram capazes de identificá-la num teste comparativo com outra margarina, foi realizado o seguinte experimento: 20 consumidores habituais da margarina A provaram, cada um, em ordem aleatória, 2 pedaços de pão – um com A e outro com B (margarina desconhecida); cada degustador, após provar os 2 pedaços de pão com margarina, procurou identificar A, dizendo o número 1 ou 2, conforme a ordem – sempre casual – em que tenha recebido os pedaços de pão. Não houve comunicação entre os degustadores. Ao cabo do experimento, verificou-se que 15 respostas estavam corretas. Pode-se afirmar, com nível de significância de 5%, que há uma tendência de os degustadores conseguirem, de fato, reconhecerem A?

9) Quarenta pessoas se matricularam num curso de escrita criativa. Na primeira aula foi aplicado um teste para verificar a capacidade de escrever de cada aluno. Ao final do curso foi aplicado novo teste. Um especialista verificou quem melhorou e quem piorou sua capacidade de escrever, encontrando 30 que melhoraram e 10 que pioraram. Estes dados mostram evidência suficiente para se afirmar que o curso tende a melhorar a capacidade de escrita?

10.5 Aplicação de testes estatísticos na pesquisa

Formulada uma pergunta ou uma hipótese de pesquisa, o pesquisador precisa planejar a coleta de dados e um teste estatístico adequado à situação. Nos capítulos seguintes, serão apresentados alguns testes bastante aplicados em pesquisas nas áreas das ciências humanas

e sociais. Eles se diferenciam, basicamente, pelo tipo de problema que se pretende resolver e pelo tipo de variável em estudo. Existem testes voltados para variáveis quantitativas, em que normalmente as hipóteses são apresentadas em termos de *médias* e testes voltados para variáveis qualitativas, em que as hipóteses são apresentadas em termos de *proporções* ou *probabilidades* de eventos. Os exemplos deste capítulo estão no segundo caso.

Em geral, na aplicação de um teste estatístico, devemos saber:

a) formular H_0 e H_1 em termos de parâmetros populacionais;

b) como obter a estatística do teste (no exemplo da moeda, Y = *número de caras*);

c) qual é a distribuição de referência para calcular o valor p (no exemplo da moeda é a distribuição *binomial* – ou a *normal* quando n é grande);

d) quais as suposições básicas para o uso do teste escolhido (no exemplo da moeda, supusemos que os lançamentos foram imparciais e realizados sob as mesmas condições – *amostragem aleatória simples*).

A decisão do teste estatístico é feita pela comparação do valor p com o nível de significância α preestabelecido, mas a implicação do resultado estatístico depende da aplicação em questão. Por exemplo, num estudo experimental, normalmente a decisão do teste estatístico implica numa relação de causa e efeito, mas num estudo de levantamento, o resultado do teste usualmente leva apenas a uma conclusão de diferença entre grupos.

Hoje em dia, o cálculo da estatística do teste e a obtenção do valor p tornaram as tarefas relativamente fáceis com o auxílio do computador. Ou seja, o pesquisador não mais precisa ter habilidades em cálculos algébricos para realizar testes estatísticos. Por outro lado, a análise do problema de pesquisa, o planejamento da coleta dos dados, a escolha do teste estatístico, a verificação das suposições e a correta interpretação do resultado estatístico exigem conhecimento, raciocínio lógico e maturidade.

EXERCÍCIOS COMPLEMENTARES

10) Para cada um dos itens a seguir, apresente as hipóteses nula e alternativa, indicando qual abordagem (unilateral ou bilateral) é a mais adequada.

a) Um método de treinamento tende a aumentar a produtividade dos funcionários.

b) A velocidade de um veículo num percurso é, em média, menor do que o valor anunciado.
c) Dois métodos de treinamento tendem a produzir resultados diferentes na produtividade.

11) Para verificar as hipóteses de seu trabalho, um pesquisador fez vários testes estatísticos (um para cada hipótese de pesquisa), adotando para cada teste o nível de significância de 5%. Responda aos seguintes itens:
a) Num dado teste, o valor *p* foi igual a 0,0001. Com base no resultado da amostra, qual deve ser a conclusão (decide-se pela hipótese nula ou pela hipótese alternativa)? Com base no resultado da amostra, qual é o risco de o pesquisador estar tomando a decisão errada?
b) Em outro teste, o valor *p* foi igual a 0,25. Qual a conclusão? Qual é o risco de o pesquisador estar tomando a decisão errada?
c) Em outros dois testes, o valor *p* foi 0,0001 e 0,01, respectivamente. Supondo que se tenha adotado nível de significância de 5%, em qual dos dois testes o pesquisador deve estar mais convicto da rejeição de H_0? Por quê?

12) Com o objetivo de testar se uma certa moeda está *viciada*, decide-se lançá-la várias vezes de forma imparcial e sempre sob as mesmas condições.
a) Se em 8 lançamentos ocorreram 2 caras (e 6 coroas), qual é a conclusão do teste ao nível de significância de 5%?
b) Se em 80 lançamentos ocorreram 20 caras (e 60 coroas), qual é a conclusão do teste ao nível de significância de 5%?

13) Para testar se uma criança tem algum conhecimento sobre determinado assunto, foram elaboradas 12 questões do tipo *certo-errado*. A criança acertou 11. Qual é a conclusão ao nível de significância de 5%?

14) Para testar se uma criança tem algum conhecimento sobre determinado assunto, foram elaboradas 12 questões, cada uma com 4 possibilidades de escolha. A criança acertou 5.
a) Formule as hipóteses em termos do parâmetro π = probabilidade de acerto de cada questão.
b) Qual é o número esperado de acertos sob H_0.
c) Calcule o valor *p*.
d) Qual é a conclusão do teste ao nível de significância de 5%?

15) Para testar se um sistema computacional "inteligente" adquiriu algum conhecimento sobre determinado assunto, foram elaboradas 60 questões do tipo *certo-errado*. O sistema acertou 40. Qual é a conclusão do teste ao nível de significância de 5%?

Capítulo 11

Testes de comparação entre duas amostras

No Capítulo 10, introduzimos alguns conceitos básicos da metodologia dos testes estatísticos de hipóteses, ou testes de significância. Neste capítulo, discutiremos alguns testes bastante usados em pesquisa social, com ênfase nos chamados *testes t* de comparação entre duas médias. Iniciaremos com a apresentação de alguns problemas de pesquisa que envolvem testes estatísticos.

11.1 Testes de significância e delineamentos de pesquisa

Em geral, os testes estatísticos são usados para comparar diferentes grupos de elementos (pessoas, animais, etc.), com respeito a alguma variável de interesse (*variável resposta*). Esses grupos podem diferir quanto a diferentes tratamentos aplicados a seus elementos, ou a diferentes populações de onde os elementos foram extraídos.

Exemplo 11.1 Para comparar dois métodos, A e B, de ensinar matemática para crianças, podemos aplicar o método A num grupo de crianças e o método B em outro grupo. Para evitar a influência de fatores intervenientes, a composição prévia dos dois grupos deve ser feita de forma aleatória.[1] Ao longo do experimento, ambos os grupos devem ser tratados sob as mesmas condições, exceto quanto aos métodos de ensino

[1] A divisão aleatória pode ser feita por sorteio ou através de uma tabela de números aleatórios. Veja o Exercício 5, Capítulo 3.

em estudo. A comparação entre os dois grupos é realizada através de uma avaliação que mensure os conhecimentos de Matemática de cada criança (veja a Figura 11.1).

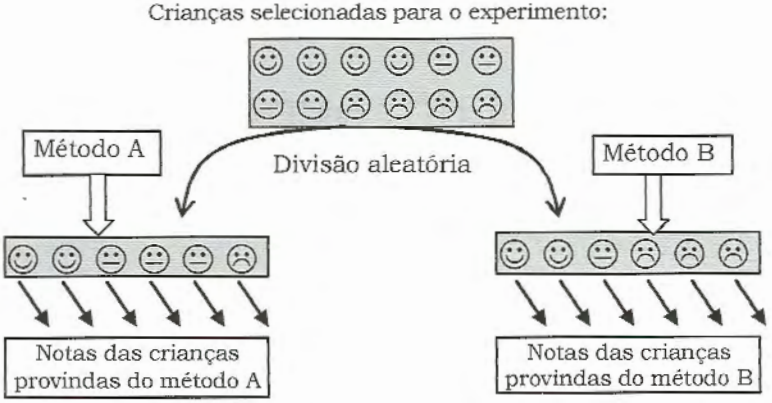


Figura 11.1 Esquema do planejamento de um experimento para comparar dois métodos de ensinar matemática para crianças.

■

"A aleatorização dos grupos é fundamental para resguardar a validade de um teste de significância" (FISHER, 1956, p.19). Entende-se por *aleatorização* não somente a divisão aleatória dos elementos nos grupos, mas também as condições idênticas em que esses grupos devem ser tratados, a não ser, é claro, pelos diferentes tratamentos em estudo. No Exemplo 11.1, devemos evitar qualquer interação entre as crianças dos dois grupos, qualquer variação devida aos instrutores, etc.

EXEMPLO 11.2 Para comparar o peso de recém-nascidos, em duas localidades, podemos extrair uma amostra aleatória de nascimentos em cada localidade, observando os pesos (veja a Figura 11.2).

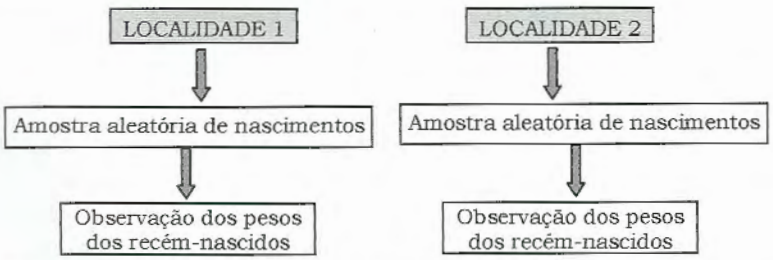


Figura 11.2 Esquema de um planejamento amostral, num estudo tipo levantamento, para comparar o peso de recém-nascidos em duas localidades.

■

Os testes estatísticos permitem avaliar se as diferenças observadas entre os dois grupos podem ser meramente justificadas por fatores casuais (H_0), ou se tais diferenças são reais (H_1).

Diferenças reais (*significativas*) podem ser causadas pelos diferentes tratamentos utilizados nos grupos em análise, como no Exemplo 11.1, ou pelas diferentes populações que geraram as amostras em estudo, como no Exemplo 11.2.

O Exemplo 11.3 mostra uma situação em que o objetivo central é comparar o comportamento de uma variável, observada sobre um conjunto de elementos, em dois momentos diferentes.

Exemplo 11.3 Com o objetivo de avaliar o efeito de um programa de treinamento sobre a produtividade dos funcionários de uma certa empresa, foi realizado um estudo em que se observou a produtividade de uma amostra de funcionários antes e depois do programa de treinamento (veja a Figura 11.3).

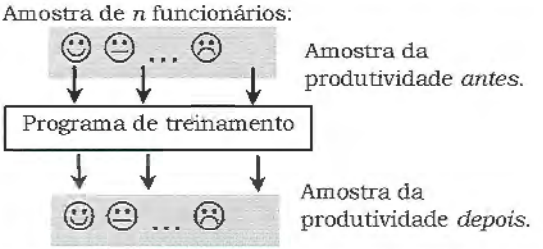


Figura 11.3 Esquema de um estudo, tipo *antes-e-depois*, para avaliar o efeito de um programa de treinamento na produtividade de funcionários de uma empresa.

■

O planejamento de pesquisa descrito no Exemplo 11.3 vai gerar *dados pareados*, pois cada funcionário estará associado a um par de medidas: uma *antes* e outra *depois* da aplicação do programa de treinamento. Por outro lado, os planejamentos descritos nos Exemplos 11.1 e 11.2 geram amostras *independentes*, já que as medidas são extraídas de grupos de elementos distintos e independentes.

O planejamento tipo *antes-e-depois* é apenas um exemplo de geração de dados pareados. Outro caso comum ocorre quando formamos pares de indivíduos relativamente similares, aplicando tratamentos diferentes nos indivíduos de cada par. Por exemplo, na comparação de dois métodos de ensino (Exemplo 11.1), podemos formar pares de indivíduos tão similares quanto possível em termos de inteligência e conhecimento prévio sobre assuntos correlacionados (ver Figura 11.4).

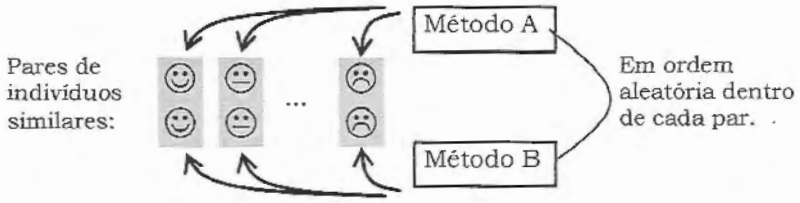


Figura 11.4 Planejamento de pesquisa alternativo para o Exemplo 11.1 – dados pareados.

Ao realizar o planejamento de uma pesquisa, é fundamental planejar, também, o procedimento estatístico que vai ser usado na análise dos dados. Particularmente, em pesquisas confirmatórias, isto é, naquelas em que temos hipóteses que desejamos colocar à prova, devemos realizar o planejamento da pesquisa preocupando-nos em verificar se a realização da pesquisa planejada vai gerar dados pareados ou amostras independentes, dados quantitativos ou categorizados, e assim por diante. Para cada situação, há um teste estatístico específico.

Um cuidado básico no planejamento (delineamento) de uma pesquisa é a perfeita coerência que deve haver entre a hipótese a ser testada e o planejamento e realização da pesquisa. Por exemplo, o planejamento proposto para o Exemplo 11.3 (procedimento *antes-e-depois*) somente é recomendado quando se tem segurança de que, no período entre as duas mensurações, o único fator que afeta sistematicamente os dados (valores de produtividade) é o fator em estudo (programa de treinamento). Caso contrário, é mais recomendado um delineamento como proposto no Exemplo 11.1 (amostras independentes).

Vamos apresentar alguns testes estatísticos que podem ser aplicados em problemas de comparação entre duas amostras, discutindo as situações adequadas para suas aplicações.

11.2 O TESTE DOS SINAIS

O teste dos sinais não é uma das técnicas estatísticas mais usadas em pesquisas sociais, mas será apresentado em primeiro lugar devido a sua simplicidade e por usar distribuições de probabilidades bastante discutidas em capítulos anteriores. Este teste é adequado quando:

- os dados são pareados e
- a variável em estudo é observada, ou analisada, de forma *qualitativa*, e com apenas duas categorias, tal como: *melhorou* ou *piorou*.

Voltemos a considerar o Exemplo 11.3, em que se quer verificar se um certo programa de treinamento aumenta a produtividade dos funcionários de uma certa empresa. Temos, então, as seguintes hipóteses:

H_0: a produtividade *não* se altera com o programa de treinamento;
H_1: a produtividade *aumenta* com o programa de treinamento.

Vamos supor que ao observar a produtividade de um funcionário, antes e depois da realização do programa de treinamento, é possível avaliar se *melhorou* ou *piorou*. Neste contexto, as hipóteses podem ser colocadas em termos do parâmetro π da distribuição binomial, como segue.

H_0: $\pi = 0{,}5$ e H_1: $\pi > 0{,}5$

onde π representa a probabilidade do funcionário aumentar a produtividade após o treinamento.

O teste é realizado com base numa amostra de n funcionários. Para cada funcionário é observada a sua produtividade *antes* e *depois* da aplicação do programa de treinamento, verificando se *melhorou* (sinal +) ou se *piorou* (sinal –). A estatística do teste é o *número Y de funcionários que aumentam a sua produtividade.*

Supondo que:

- todos os funcionários sejam observados sob as mesmas condições;
- não haja interação entre os funcionários que estão participando da pesquisa; e
- o único fator que esteja influenciando sistematicamente a produtividade dos funcionários, ao longo do estudo, seja o programa de treinamento,

a estatística Y tem distribuição binomial com parâmetros n e π (análogo ao exemplo da moeda do capítulo anterior). Assim, o valor p pode ser computado pela distribuição binomial ou, quando n for grande, pela distribuição normal.

Considere que $n = 10$ funcionários participaram da pesquisa descrita no Exemplo 11.3, gerando os resultados constantes na Tabela 11.1. O sinal + indica que o funcionário melhorou sua produtividade após o treinamento, e o sinal – indica que piorou.

Tabela 11.1 Avaliação qualitativa da produtividade de 10 funcionários, antes e depois de serem submetidos a um programa experimental de treinamento.

Funcionário	Avaliação da produtividade	Funcionário	Avaliação da produtividade
João	+	Joana	+
Maria	+	Flávio	+
José	–	Paulo	+
Pedro	+	Catarina	–
Rita	–	Felipe	+

Pela Tabela 11.1, temos: $Y = 7$. Assim, pela distribuição binomial (Tabela 2 do Apêndice), com $n = 10$ e $\pi = 0,5$, temos:

$$p = p(7) + p(8) + p(9) + p(10) = 0,1172 + 0,0439 + 0,0098 + 0,0010 = 0,1719.$$

Considerando o nível de significância de 5% ($\alpha = 0,05$), o teste dos sinais *não* pode rejeitar H_0 em favor de H_1 (pois, $p > \alpha$). Concluímos, então, que os dados *não* mostram evidência suficiente para garantir que o programa de treinamento melhora a produtividade de funcionários.

Num estudo tipo *antes-e-depois*, muitas vezes não é possível distinguir se um certo indivíduo *melhorou* ou *piorou*. Neste caso, é comum desprezar esses indivíduos da amostra (veja o Exercício 1d). Contudo, se houver um número grande de indivíduos nessa situação, a aplicação deste teste estatístico pode ficar prejudicada.

EXERCÍCIOS

1) Com o objetivo de avaliar se o desempenho de um certo candidato, numa apresentação em público, foi positivo, foi selecionada uma amostra de uma grande plateia, indagando de cada um, sua opinião sobre o candidato (se melhorou ou se piorou), antes e depois da apresentação.
   a) Apresente as hipóteses nula e alternativa.
   b) Se, numa amostra de 11 pessoas, 8 passaram a ter uma opinião mais favorável, enquanto 3 passaram a ter opinião menos favorável sobre o candidato, o que se pode afirmar? Use nível de significância de 5%.
   c) Se, numa amostra de 200 pessoas, 130 passaram a ter melhor impressão, enquanto 70 pioraram sua impressão sobre o candidato, o que se pode afirmar? Com que probabilidade de significância? Sugestão: use a aproximação normal (Seção 8.3).
   d) Considere que exista também a resposta *opinião inalterada*. Numa amostra de 100 pessoas, 60 passaram a ter opinião mais favorável, 30 passaram a ter opinião menos favorável e 10 mantiveram a mesma opinião. O que se pode afirmar, ao nível de significância de 5%? Sugestão: elimine da amostra as pessoas cujas opiniões ficaram inalteradas.

2) (SIEGEL, 1981, p. 80.) Um pesquisador está interessado em avaliar se determinado filme, sobre delinquência juvenil, contribui para modificar a opinião de uma comunidade sobre quão severa deve ser a punição em tais casos. Para tanto, ele extrai uma amostra aleatória de 100 indivíduos da comunidade e realiza um estudo tipo *antes-e-depois*. Pergunta a cada indivíduo da amostra se devem aplicar, nos casos de delinquência juvenil, punição mais forte ou mais fraca do que a que vem sendo aplicada correntemente. Em seguida, exibe o filme para estes 100 indivíduos e, após a exibição, repete a pergunta. Oitenta e cinco indivíduos mudaram de opinião, sendo que 59 deles modificaram sua opinião

de *mais* para *menos*, enquanto que 26 de *menos* para *mais*. Estes dados mostram evidência suficiente de que o filme produz um efeito sistemático nos indivíduos da comunidade em estudo? Com que probabilidade de significância?

11.3 O Teste *t* para dados pareados

O chamado *teste t* é apropriado para comparar dois conjuntos de dados *quantitativos*, em termos de seus *valores médios*. Nesta seção, trataremos do caso em que os dados são *pareados*.

Exemplo 11.4 Retomemos o problema do Exemplo 11.3, mas, agora, vamos supor que a variável produtividade possa ser mensurada *quantitativamente*, numa escala que varia de 20 a 40 pontos. Para aplicar o *teste t*, as hipóteses deverão ser formuladas em termos de valores médios, como segue:

H_0: a *produtividade média* dos funcionários *não se altera* com o programa de treinamento;

H_1: a *produtividade média* dos funcionários *aumenta* com o programa de treinamento.

Ou, ainda,

$H_0: \mu_{depois} = \mu_{antes}$ e $H_1: \mu_{depois} > \mu_{antes}$,

onde:

μ_{antes}: produtividade média dos funcionários antes do treinamento; e
μ_{depois}: produtividade média dos funcionários depois do treinamento.

Para colocar H_0 à prova, vamos observar os $n = 10$ funcionários, antes e depois de receberem o programa de treinamento (duas amostras pareadas de valores de produtividade). Os dados estão na Tabela 11.2.

Tabela 11.2 Valor da produtividade de cada funcionário, antes e depois de um programa experimental de treinamento.

Funcionário	Produtividade		
	Antes X_1	Depois X_2	Diferença $D = X_2 - X_1$
João	22	25	3
Maria	21	28	7
José	28	26	-2
Pedro	30	36	6
Rita	33	32	-1
Joana	33	39	6
Flávio	26	28	2
Paulo	24	33	9
Catarina	31	30	-1
Felipe	22	27	5

A última coluna da Tabela 11.2 mostra a diferença entre os valores de produtividade antes e depois. Esses incrementos (ou reduções) de produtividade estão também apresentados na Figura 11.5, sob forma de um diagrama de pontos.

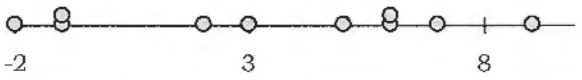


Figura 11.5 Diagrama de pontos das diferenças de produtividade.

Observamos no diagrama de pontos da amostra que houve uma tendência de ocorrer diferenças positivas (valores de produtividade *depois* maiores, em geral, do que os valores de produtividade *antes*). A realização do *teste t* permite verificar se esta tendência não poderia ser explicada, apenas, por efeitos casuais.

■

ESTATÍSTICA DO TESTE

A estatística do teste baseia-se nos valores observados da variável *D*, definida pela diferença de valores em cada par. Num estudo tipo *antes-e-depois*:

$$D = (\textit{medida depois}) - (\textit{medida antes})$$

Se a hipótese nula for correta, devemos esperar que os valores desta variável estejam em torno de zero ou, ainda, que a média destas diferenças, $\overline{D}$, esteja próxima de zero. Usaremos, como estatística do teste, uma função de $\overline{D}$, conhecida como *estatística t para dados pareados*, que é definida por:

$$t = \frac{\overline{D} \cdot \sqrt{n}}{S_D}$$

onde

n : tamanho das amostras, que, neste caso, corresponde ao número de pares observados;
$\overline{D}$: média das diferenças internas dos pares; e
S_D : desvio padrão das diferenças internas dos pares.

Exemplo 11.4 (continuação) Diferenças D (última coluna da Tabela 11.2):

3, 7, -2, 6, -1, 6, 2, 9, -1, 5

Então:

$$n = 10 \qquad \overline{D} = \frac{\sum D}{n} = \frac{34}{10} = 3{,}4$$

$$S_D = \sqrt{\frac{\sum D^2 - n \cdot \overline{D}^2}{n-1}} = \sqrt{\frac{246 - (10)\cdot(3{,}4)^2}{10-1}} = 3{,}81$$

E, portanto,

$$t = \frac{\overline{D} \cdot \sqrt{n}}{S_D} = \frac{3{,}4 \cdot \sqrt{10}}{3{,}81} = 2{,}82$$

■

O fato de a estatística do teste ser função de n é bem razoável, já que, quanto maior o tamanho da amostra, mais conhecimento se tem sobre o fenômeno em estudo e, consequentemente, um certo afastamento entre $\overline{D}$ e *zero* tem menor probabilidade de ser explicado meramente pelo acaso. A estatística t também é função do desvio padrão S_D, que é uma medida do grau de heterogeneidade daquilo que estamos estudando. Quanto maior esta heterogeneidade, maiores devem ser as diferenças observadas entre as duas medidas para evidenciar uma diferença média real (ou significativa) entre elas.

Distribuição do teste

Quando o valor calculado da estatística t estiver próximo de *zero*, H_0 poderá ser aceita. Por outro lado, se t estiver longe de *zero*, H_0 deverá ser rejeitada, em favor de H_1. É necessário, porém, ter uma distribuição de referência para especificarmos o que significa *próximo* ou *longe* de zero. Esta distribuição de referência existe sob a seguinte suposição.

Suposição básica para a aplicação do teste: Teoricamente, devemos supor que a variável D (diferença entre as duas mensurações) segue uma distribuição normal. Contudo, se o número de pares for razoavelmente grande ($n \geq 30$, por exemplo), o teste ainda permanece válido, mesmo que a variável D não tenha distribuição normal.

Na prática, recomendamos fazer histogramas de frequências ou diagramas de pontos das duas amostras para verificar se não existe algum ponto discrepante ou forte assimetria, o que poderia comprometer a

validade deste teste estatístico. Alternativamente, esta análise exploratória pode ser feita com os valores da variável *D*, como foi apresentado na Figura 11.5, onde não parece haver ponto discrepante ou forte assimetria.

Distribuição de referência: Sob H_0, e considerando a suposição acima descrita, a estatística *t* tem distribuição t de Student *com* $gl = n - 1$ graus de liberdade (veja Figura 11.6).

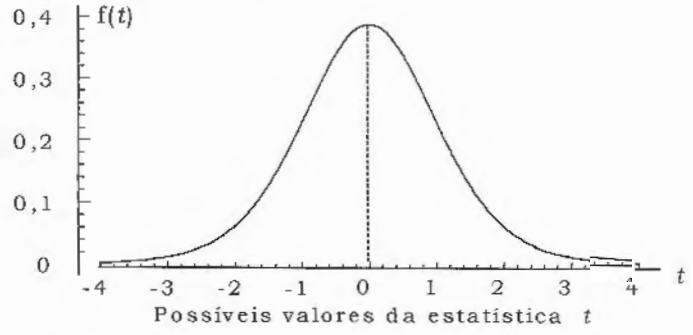


Figura 11.6 Distribuição de referência para o teste *t* do Exemplo 11.5. A Distribuição t de Student com *gl* = 9 graus de liberdade.

A Figura 11.6 mostra a distribuição dos possíveis valores da estatística *t*, na suposição de não haver diferença real entre as duas mensurações (H_0) – somente variações casuais em torno de zero.

PROBABILIDADE DE SIGNIFICÂNCIA

Depois de observar os dados e calcular o valor da estatística *t*, podemos obter o valor *p* pela distribuição t de Student (Tabela 5 do apêndice), conforme é mostrado na continuação do Exemplo 11.4.

EXEMPLO 11.4 (CONTINUAÇÃO) Para testar H_0: $\mu_{depois} = \mu_{antes}$ *versus* H_1: $\mu_{depois} > \mu_{antes}$, observamos uma amostra de $n = 10$ funcionários, que produziu o valor $t = 2{,}82$. Como $n = 10$, temos $gl = n - 1 = 9$ graus de liberdade. Tomemos, então, a linha de $gl = 9$ da Tabela 5 do apêndice, como mostra a Figura 11.7. Por esta tabela, obtemos a área relativa a um valor maior ou igual a $t = 2{,}82$. Como o teste é unilateral, esta área já corresponde à probabilidade de significância *p* descrita pelos dados da amostra.

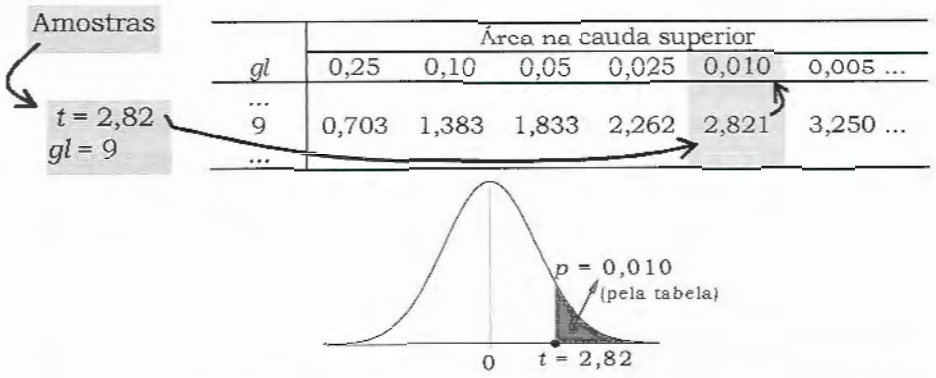


Figura 11.7 Uso da distribuição t de Student com $gl = 9$ para a obtenção da probabilidade de significância num teste unilateral, com $n = 10$ e $t = 2,82$.

Observando a linha correspondente a $gl = 9$, verificamos, na tabela, que o valor $t = 2,82$ (calculado com base na amostra) está próximo do valor tabulado 2,821. Logo, como ilustra a Figura 11.7, a probabilidade de significância é, aproximadamente, $p = 0,010$.

Considerando o nível de significância de 5% ($\alpha = 0,05$), o teste conclui que os dados mostram evidência suficiente de que H_0 *é falsa* (pois, $p = 0,010$ e, portanto, menor que o nível de significância adotado $\alpha = 0,05$), detectando, então, que houve um aumento real da produtividade entre as duas mensurações. Se admitirmos que não houve qualquer outro fator, além do programa de treinamento, atuando de forma sistemática entre as duas mensurações, podemos concluir que o programa de treinamento tende a aumentar a produtividade dos funcionários.

■

O leitor pode ter observado que os dados do Exemplo 11.3 correspondem aos dados do Exemplo 11.4, se estes fossem classificados em apenas duas categorias: *melhorou* (+) ou *piorou* (–). Mas as aplicações dos testes dos sinais e *t* levaram a conclusões diferentes. Isto pode ocorrer pelo fato do teste dos sinais usar apenas uma avaliação *qualitativa* das diferenças, enquanto que o teste *t* usa melhor a informação contida nos dados, trabalhando com as *quantidades*. O teste *t* é um teste *mais poderoso* do que o teste dos sinais, no sentido de ter maior probabilidade de detectar diferenças, quando elas realmente existem. Contudo, a validade do teste *t* está condicionada à suposição de a variável em estudo ter distribuição normal, especialmente se a amostra for pequena.

TESTES BILATERAIS

No Exemplo 11.4, realizamos um teste unilateral, pois a hipótese alternativa foi formulada com o sinal ">" (H_1: $\mu_{depois} > \mu_{antes}$). Quando o teste é bilateral, isto é, a hipótese alternativa tem o sinal "$\neq$", o procedimento é análogo, mas o valor de área da tabela deverá ser *dobrado*, para que o valor *p* corresponda às áreas das duas caudas da distribuição.

EXEMPLO 11.5 Desejamos verificar se uma certa alteração no turno de trabalho produz algum efeito, positivo ou negativo, na produtividade dos funcionários. Para isto, realizamos um estudo experimental, alterando o turno de trabalho de uma amostra de $n = 10$ funcionários da empresa. Temos as seguintes hipóteses:

H_0: $\mu_{depois} = \mu_{antes}$ e H_1: $\mu_{depois} \neq \mu_{antes}$

onde:

μ_{antes}: produtividade média dos funcionários da empresa no horário habitual; e

μ_{depois}: produtividade média dos funcionários da empresa com alteração no turno de trabalho.

Por simplicidade, suponha que os resultados foram os mesmos do Exemplo 11.4, apresentados na Tabela 11.2, resultando, como já vimos, em $t = 2{,}82$, com $gl = 9$. A obtenção da probabilidade de significância é análoga ao caso anterior, considerando, porém, ambos os lados da curva. Assim, $p = 2\times(0{,}010) = 0{,}020$. Portanto, ao nível de significância de 5%, o teste rejeita H_0, em favor de H_1.

■

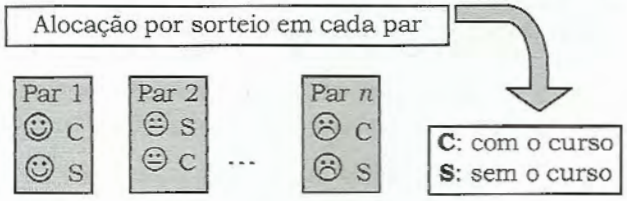
OUTRAS FORMAS DE PAREAMENTO

O plano de pesquisa de observar a variável resposta sobre os mesmos elementos, antes e depois de aplicar um certo tratamento, pareceu adequado no problema de avaliar o efeito de um programa de treinamento sobre a produtividade de funcionários. Contudo, se o programa de treinamento for relativamente longo, de tal forma que, nesse período, outros fatores puderem agir de forma sistemática sobre a produtividade, o estudo torna-se inócuo, pois diferenças reais entre as duas mensurações podem ser tanto devidas ao programa de treinamento, como devidas a fatores intervenientes.

Um planejamento mais adequado para a situação colocada consiste em observar dois grupos de funcionários, sendo que apenas um dos grupos recebe o programa de treinamento. Após a realização do treinamento, comparam-se os valores de produtividade entre os dois grupos.[2]

Uma maneira de constituir grupos de elementos comparáveis consiste em construir pares de elementos aproximadamente semelhantes. Os elementos de cada par são separados e, cada um, submetido a uma das condições (tratamentos) que se deseja comparar, formando os dois grupos. A observação do efeito dos tratamentos pode ser feita, em cada par, pela variável *D*, definida como a diferença entre os elementos do par (ver Figura 11.4).

Exemplo 11.6 Para avaliar o efeito de um curso sobre alimentação e controle de peso, em pessoas obesas, planeja-se realizar uma pesquisa com pares de pessoas relativamente similares. Os pares serão constituídos por pessoas de mesmo sexo, faixa de peso, faixa etária, além de outras características pertinentes. Em cada par, uma das pessoas, selecionada aleatoriamente, deverá participar do curso, e a outra não. Depois de três meses, é medida a variação de peso das pessoas de ambos os grupos. Esquematicamente:



Este procedimento deverá gerar um conjunto de dados pareados e quantitativos (pois a variável resposta, *variação de peso*, é quantitativa). Assim, podemos aplicar o teste *t* de forma análoga ao que fizemos no Exemplo 11.4.

Exercícios

3) Seja o problema do Exemplo 11.6.
   a) Apresente as hipóteses nula e alternativa.
   b) Considerando que a realização da pesquisa produziu os dados constantes na tabela seguinte, qual é a conclusão?

[2] Alternativamente, poder-se-ia comparar as variações de produtividade entre os dois grupos. Neste caso, torna-se necessário, também, medir a produtividade de todos os funcionários (ambos os grupos) antes de iniciar o programa de treinamento.

Par de pessoas obesas participantes do estudo	Variação do peso, em kg, ao longo de três meses[1]	
	com o curso	sem o curso
1	-4	2
2	-2	3
3	-3	-1
4	1	-2
5	0	5
6	2	2
7	-5	-1
8	-3	-3
9	1	2
10	0	4

[1] Valores positivos indicam ganho de peso, e valores negativos, perda de peso.

4) Para avaliar o efeito de um brinde nas vendas de determinado produto, planeja-se comparar as vendas em lojas que vendem o produto com o brinde, com as vendas em lojas que não oferecem o brinde. Para reduzir o efeito de variações devidas a outros fatores, as lojas foram grupadas em pares de lojas, sendo que as lojas de um mesmo par sejam tão similares quanto possível, em termos do volume de vendas, localidade, identidade de preços, etc. Em cada par de lojas, uma passou a oferecer o brinde, e a outra não.
   a) Apresente as hipóteses nula e alternativa.
   b) Os resultados das vendas, em quantidade de unidades vendidas, foram os seguintes:

Par de lojas	Vendas	
	sem brinde	com brinde
1	33	43
2	43	39
3	26	33
4	19	32
5	37	43
6	27	46

   Os dados mostram evidência suficiente para se afirmar que a oferta do brinde aumenta as vendas? Use nível de significância de 5%.

5) Para resolver o mesmo problema do exercício anterior, decidiu-se fazer um planejamento do tipo *antes-e-depois*. Observou-se a venda mensal do produto em questão nas 12 lojas. Depois, passou-se a oferecer um brinde e voltou-se a avaliar a venda mensal desse produto nas 12 lojas. Os incrementos (ou reduções) nas vendas foram os seguintes:

   7 10 5 -2 9 0 3 -4 8 9 1 3

   a) Os dados mostram evidência suficiente para se afirmar que a oferta do brinde aumenta as vendas? Use nível de significância de 5%.
   b) No problema em discussão, aponte as vantagens e desvantagens deste planejamento de pesquisa, em relação ao apresentado no Exercício 4.
   c) Apresente um terceiro planejamento de pesquisa para este problema, tentando aproveitar as vantagens dos dois procedimentos apresentados.

6) Para avaliar o governo perante os empresários, um instituto de pesquisa selecionou uma amostra aleatória de 64 empresários, indagando a cada um

sua aprovação com o governo, numa escala de 0 a 10. Foi realizada uma pesquisa logo após a posse do governo, e outra após seis meses, mas com a mesma amostra de empresários. A primeira amostra apontou uma média de 8,4 e a segunda 6,8 (diferença média de 1,6). O desvio padrão da diferença foi 2,0. Os dados mostram evidência suficiente para afirmar que na população de empresários houve redução na aprovação ao governo? Use $\alpha = 0{,}01$.

7) Considerando os dados do anexo do Capítulo 2, podemos afirmar que existe diferença significativa entre: (a) *satisfação dos alunos quanto à didática dos professores* e (b) *satisfação dos alunos quanto aos laboratórios e recursos materiais*? Use $\alpha = 0{,}01$. Em qual dos dois itens os alunos estão, em média, mais satisfeitos?

11.4 O teste *t* para amostras independentes

A formação de pares de elementos similares nem sempre é viável. Uma forma alternativa é considerar duas amostras independentes, como mostra o exemplo seguinte.

Exemplo 11.7 Retomemos o problema de comparar dois métodos, A e B, de ensinar matemática para crianças. As hipóteses podem ser:

H_0: em média, os dois métodos produzem os *mesmos* resultados; e
H_1: em média, os dois métodos produzem resultados *diferentes*.

Para realizar o teste, precisamos de uma amostra de crianças submetidas ao método A de ensino, e outra amostra de crianças submetidas ao método B, conforme planejamento discutido no Exemplo 11.1. Ao término dos estudos, todas as crianças devem efetuar uma mesma avaliação para medir o grau de aprendizagem. Em termos do planejamento proposto, podemos escrever:

$H_0: \mu_1 = \mu_2$ e $H_1: \mu_1 \neq \mu_2$,

onde:

μ_1: nota média (ou esperada) de crianças que sejam submetidas ao método A de ensino; e
μ_2: nota média (ou esperada) de crianças que sejam submetidas ao método B de ensino.

A Tabela 11.3 mostra os resultados do experimento descrito no Exemplo 11.7, considerando que ambos os grupos foram compostos por dez crianças. A Figura 11.8 apresenta o diagrama de pontos dos resultados da avaliação, segundo o método de ensino.

Tabela 11.3 Notas na avaliação, considerando o método de ensino

Método A de ensino					Método B de ensino				
45	51	50	62	43	45	35	43	59	48
42	53	50	48	55	45	41	43	49	39

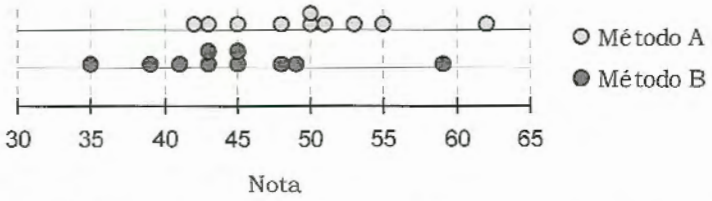


Figura 11.8 Diagrama de pontos das notas obtidas pelas crianças, segundo o método de ensino

ESTATÍSTICA DO TESTE

A estatística do teste toma como base a diferença entre as médias das duas amostras, $\overline{X}_1 - \overline{X}_2$, mas leva também em consideração o número de elementos em cada amostra e a variabilidade interna dessas amostras. Quanto maiores as amostras, maior a evidência de uma diferença real. Pense no caso extremo de apenas uma criança em cada grupo, apontando uma diferença de duas unidades numa escala de 0 a 10 – *não dá para dizer muita coisa!* Mas com 100 crianças em cada grupo, apontando uma diferença de duas unidades, leva-nos a induzir que os métodos produzem resultados diferentes. Por outro lado, se há muita variabilidade entre os elementos de cada amostra, uma possível diferença fica nebulosa. Veja a Figura 11.9.

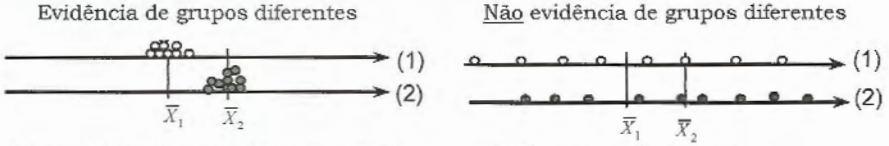


Figura 11.9 A importância de se considerar a variância interna dos grupos

Considerando o mesmo número n de elementos em cada amostra, a variância agregada, S_a^2, é obtida pela média aritmética das variâncias de cada grupo, S_1^2 e S_2^2, ou seja:[3]

$$S_a^2 = \frac{S_1^2 + S_2^2}{2}$$

[3] Lembramos ao leitor que a variância (S^2) é o desvio padrão (S) ao quadrado.

E a estatística do teste é dada por:

$$t = (\overline{X}_1 - \overline{X}_2) \cdot \sqrt{\frac{n}{2 \cdot S_a^2}}$$

onde:

n : tamanho da amostra em cada grupo;
$\overline{X}_1$: média da amostra 1;
$\overline{X}_2$: média da amostra 2;
S_1^2 : variância da amostra 1;
S_2^2 : variância da amostra 2; e
S_a^2 : variância agregada das duas amostras.

Exemplo 11.7 (continuação) Calculando as médias e as variâncias dos dados da Tabela 11.3:

Amostra 1: $n = 10$, $\overline{X}_1 = 49{,}90$ e $S_1^2 = 35{,}66$
Amostra 2: $n = 10$, $\overline{X}_2 = 44{,}70$ e $S_2^2 = 42{,}23$

Variância agregada:

$$S_a^2 = \frac{S_1^2 + S_2^2}{2} = \frac{35{,}66 + 42{,}23}{2} = \frac{77{,}89}{2} = 38{,}95$$

Estatística do teste:

$$t = (\overline{X}_1 - \overline{X}_2) \cdot \sqrt{\frac{n}{2 \cdot S_a^2}} = (49{,}90 - 44{,}70) \cdot \sqrt{\frac{10}{2 \cdot (38{,}95)}} = (5{,}2) \cdot \sqrt{0{,}1284} = (5{,}2) \cdot (0{,}3583)$$

Portanto: $t = 1{,}86$.

■

Suposições para a aplicação do teste:

1) os dois conjuntos de dados proveem de distribuições normais e
2) têm a mesma variância.[4]

Na prática, não é fácil verificar a veracidade destas suposições. Aconselhamos, contudo, construir histogramas de frequências ou diagramas de pontos para cada amostra. Esses gráficos permitem avaliar se existem fortes violações das suposições, tais como a presença de pontos discrepantes, distribuições com formas assimétricas ou, ainda, uma

[4] Se as amostras forem razoavelmente grandes (digamos, $gl = 2n - 2 \geq 30$) a suposição (1) pode ser relaxada. Quanto à suposição (2), só vai haver problemas sérios se as variâncias das duas populações forem demasiadamente diferentes.

distribuição bem mais dispersa do que a outra. No exemplo em discussão, construímos diagramas de pontos para as duas amostras (Figura 11.8), os quais mostram que as amostras em análise parecem compatíveis com as suposições do teste.

Distribuição de referência. Considerando que as suposições do teste estejam satisfeitas, se as médias populacionais forem iguais (H_0 verdadeira), então a estatística t tem distribuição t de Student com $gl = 2n - 2$ graus de liberdade.

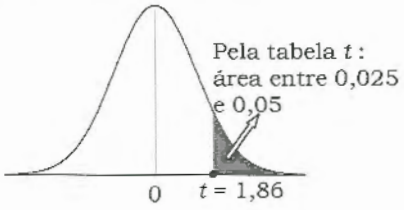
EXEMPLO 11.7 (CONTINUAÇÃO) O esquema seguinte ilustra o uso da Tabela 5 do apêndice para se obter a probabilidade de significância associada ao valor calculado $t=1,86$. No caso, tem-se $gl = 2n - 2 = 2(10) - 2 = 18$.

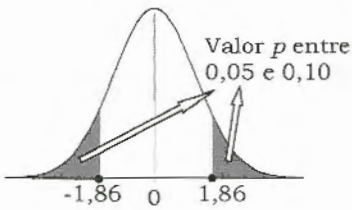
Amostras

$t = 1,86$
$gl = 18$

	Área na cauda superior					
gl	0,25	0,10	0,05	0,025	0,010	0,005 ...
...						
18	0,688	1,330	1,734	2,101	2,552	2,878 ...
...						

Os dados levaram ao valor $t = 1,86$, apontando para uma área na cauda superior da curva entre 0,025 e 0,05. Mas, como o teste é bilateral (H_1: $\mu_1 \neq \mu_2$), a área deve ser dobrada para se ter o valor p correto. Veja o esquema a seguir:





Portanto: $0,05 < p < 0,10$.

Ao nível de significância de 5%, concluímos que os dados não comprovam uma diferença entre os dois métodos de ensinar matemática. Existe uma probabilidade razoável, superior a 5%, de as diferenças observadas nos dados experimentais serem provenientes de fatores casuais.

■

Amostras de tamanhos diferentes

Quando as amostras têm tamanhos diferentes, a variância agregada é calculada por:

$$S_a^2 = \frac{(n_1 - 1) \cdot S_1^2 + (n_2 - 1) \cdot S_2^2}{gl}$$

onde:

n_1 : tamanho da amostra 1;
n_2 : tamanho da amostra 2;
S_1^2 : variância da amostra 1;
S_2^2 : variância da amostra 2; e
$gl = n_1 + n_2 - 2$: número de graus de liberdade das duas amostras agregadas.

A estatística do teste é dada por:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S_a \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

onde:

$\overline{X}_1$: média da amostra 1;
$\overline{X}_2$: média da amostra 2; e
S_a : desvio padrão agregado (raiz quadrada da variância agregada).

Exemplo 11.8 Queremos verificar, em alunos do ensino médio que já experimentaram algum tipo de droga, se a idade com que o fizeram pela primeira vez é diferente entre homens e mulheres.[5] Em especial, queremos testar as hipóteses:

$$H_0: \mu_1 = \mu_2 \quad \text{e} \quad H_1: \mu_1 \neq \mu_2$$

sendo μ_1 e μ_2 definidos na população de pessoas que já experimentaram droga, como:

μ_1: média de idade em que os homens experimentam droga; e
μ_2: média de idade em que as mulheres experimentam droga.

A pesquisa foi feita com 56 alunos (32 do sexo masculino e 24 do sexo feminino) que já experimentaram droga. Amostras e cálculos:

[5] Este trabalho foi realizado pelas alunas Kátia Vieira e Roseana Rotta na disciplina de Estatística, semestre 1999/1, Curso de Psicologia da UFSC. A população foi definida como sendo os alunos das escolas municipais de São José – SC.

Sexo	Idade em que experimentou pela 1ª vez	Média	Variância
Masc.	09 12 10 12 11 09 08 12 13 09 13 08 17 09 09 08 09 08 14 08 08 08 08 13 10 10 15 13 13 12 14 08	10,625	6,371
Fem.	14 15 08 13 16 12 14 17 14 10 13 12 13 14 10 15 12 17 16 12 15 13 14 14	13,458	4,781

Graus de liberdade: $gl = n_1 + n_2 - 2 = 24 + 31 - 2 = 54$

Variância agregada das duas amostras:

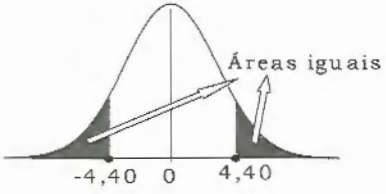
$$S_a^2 = \frac{(n_1 - 1)\cdot S_1^2 + (n_2 - 1)\cdot S_2^2}{gl} = \frac{(31)\cdot(6{,}371) + (23)\cdot(4{,}781)}{54} = 5{,}694$$

Desvio padrão agregado: $S_a = \sqrt{5{,}694} = 2{,}386$

Estatística do teste:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S_a \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{10{,}625 - 13{,}458}{(2{,}386)\cdot\sqrt{\frac{1}{24} + \frac{1}{32}}} = \frac{-2{,}833}{(2{,}386)\cdot(0{,}270)} = -4{,}40$$

Como a Tabela 5 relaciona valores positivos de t com áreas na cauda superior da curva e, também, a distribuição t é simétrica em torno de zero, devemos procurar a área relacionada com $t = 4{,}40$, como mostra o esquema a seguir:



Entrando na tabela com $gl = 60$ (o mais próximo do gl verdadeiro, igual a 54) e valor de $t = 4{,}40$, verificamos pela Tabela 5 que a área na cauda superior é inferior a 0,0005. Como o teste é bilateral, temos que o valor p é inferior a 0,001 (o dobro da área na cauda superior). Assim, o teste rejeita H_0 ao nível de significância de 0,05, pois, $p < 0{,}001 < 0{,}05 = \alpha$. Concluímos, então, que na população em estudo, os homens tendem a experimentar drogas com menor idade do que as mulheres.

■

Usando o computador

Como já discutimos anteriormente, hoje em dia a parte de cálculos da análise estatística tornou-se muito simples com o auxílio do computador. Existem, no mercado, diversos pacotes computacionais de estatística (*SAS, SPSS, STATISTICA, S-PLUS,* etc.) que fazem os diversos métodos discutidos na literatura, com uma interface *amigável.* Até mesmo as planilhas eletrônicas estão incorporando técnicas básicas de estatística. Na Figura 11.10 é apresentada uma saída do *Microsoft Excel*®, com a aplicação do teste *t* aos dados do Exemplo 11.8.[6]

Teste t: duas amostras presumindo variâncias equivalentes

	Meninos	Meninas
Média	10,62500	13,45833
Variância	6,37097	4,78080
Observações	32	24
Variância agrupada	5,69367	
Hipótese da diferença de média	0	
gl	54	
Estat t	-4,39732	
P(T<=t) unicaudal	0,000026	
t crítico unicaudal	1,67357	
P(T<=t) bicaudal	0,000052	
t crítico bicaudal	2,00488	

Figura 11.10 Teste t realizado pelo Excel® (Exemplo 11.8).

As três primeiras linhas da tabela de saída são medidas descritivas de cada amostra e, na quarta linha, tem-se a variância agregada das duas amostras. A "hipótese da diferença de médias" igual a zero (quinta linha) indica que a hipótese nula do teste afirma que as duas médias são iguais. Na sexta e sétima linhas, têm-se os graus de liberdade e o valor da estatística *t*. Os resultados apresentados nas últimas quatro linhas dependem se estamos fazendo um teste unilateral (*unicaudal*) ou bilateral (*bicaudal*). Como no nosso exemplo o teste é bilateral, leremos apenas as duas últimas linhas. Em "P(T<=t)" é dada a probabilidade de significância ($p = 0{,}000052$) e em "*t* crítico" é dado o menor valor de *t* para o teste rejeitar H_0, ao nível de significância de 5%. Usando a abordagem que vínhamos trabalhando (através do valor *p*), concluímos que o teste rejeita H_0.

[6] No *Microsoft Excel*, várias técnicas estatísticas podem ser feitas acionando no menu principal "ferramentas", "suplementos" e solicitando que se instalem as "ferramentas de análise". Clicar em "ferramentas" e "análise de dados". Para realizar o teste *t* discutido nesta seção (teste *t* para amostras independentes), escolher "Teste t: duas amostras presumindo variâncias equivalentes". Na janela que se abre, preencher os dados de entrada das duas variáveis (duas amostras), arrastando o cursor sobre as posições da planilha onde estão os dados.

Exercícios

8) Com a finalidade de verificar se o nível nutricional da mãe afeta o peso do recém-nascido, foram observadas duas amostras de nascimentos. A primeira foi extraída de uma maternidade particular (Localidade 1), onde as mães são, em geral, bem nutridas. A outra amostra foi tirada de uma maternidade pública, numa região extremamente pobre (Localidade 2), onde se acredita que as mães não são bem nutridas.

Resultados dos pesos, em kg, de recém-nascidos, em duas localidades

Localidade	Tamanho da amostra	Média (kg)	Desvio padrão (kg)
1	50	3,1	1,6
2	50	2,7	1,4

a) Os dados mostram evidência suficiente de que as crianças da Localidade 1 nascem, em média, com peso superior do que as crianças da Localidade 2? Use $\alpha = 0,05$.

b) Podemos afirmar com segurança que esta diferença no peso médio dos recém-nascidos é realmente devida ao nível nutricional da mãe?

9) Com o objetivo de comparar duas dietas para engordar frangos, realizou-se um experimento, em que 19 frangos, todos com um mês de vida, foram divididos aleatoriamente em dois grupos. No primeiro grupo, com 12 frangos, foi usada a dieta A, enquanto no segundo grupo, os 7 frangos foram tratados com a dieta B. No final de um mês, foram encontrados os seguintes resultados de ganho de peso, em gramas:

Grupo	Nº de frangos	Média (g)	Desvio padrão (g)
1	12	110	21
2	7	100	20

Os dados mostram evidência suficiente para se afirmar que as dietas produzem efeitos diferentes? Com que probabilidade de significância?

10) O objetivo é verificar se existe diferença significativa entre alunos bolsistas e não bolsistas, com respeito ao tempo médio para a conclusão dos créditos. Para isto, foi extraída uma amostra aleatória de cada grupo de alunos e observados os tempos para conclusão dos créditos, em meses:

Bolsistas					Não bolsistas					
62	24	30	34	54	56	34	60	62	42	63
					69	66	44	54	50	61

Faça o teste com $\alpha = 0,05$.

11) Numa pesquisa sobre clima organizacional nos departamentos da UFSC, professores respondem a um questionário, em que, num dos itens, o respondente atribui uma nota de 1 (um) a 5 (cinco) sobre a *clareza organizacional de seu departamento*. A tabela seguinte apresenta algumas estatísticas desta variável para os centros: Tecnológico (CTC) e Sócio-Econômico (CSE).

Centro	Tamanho da amostra	Média	Desvio padrão
CTC	79	2,67	1,06
CSE	49	2,81	1,24

Os dados mostram evidência suficiente para sugerir que os níveis médios da clareza organizacional dos departamentos são diferentes para os dois centros de ensino?

12) Num levantamento por amostragem, verificou-se o nível de renda familiar em três localidades de um certo bairro (anexo do Capítulo 4). Testar se existe diferença significativa entre essas localidades, comparando-as duas a duas.[7] Use $\alpha = 0,01$. A tabela seguinte mostra alguns resultados intermediários.

Algumas medidas descritivas da distribuição de renda de uma amostra de famílias do Bairro Saco Grande II, Florianópolis – SC, 1988

Localidade	Nº de famílias na amostra	Média (sal. mín.)	Desvio padrão (sal. mín.)
Monte Verde	40	8,10	4,28
Pq. da Figueira	42	5,83	2,57
Encosta do Morro	37	5,02	4,52

11.5 Tamanho das amostras

No planejamento de um estudo comparativo, surge a questão de qual o tamanho n da amostra em cada grupo. Para responder a esta questão, vamos relembrar alguns conceitos de testes estatísticos. Quando o teste rejeita a hipótese de igualdade entre os grupos (H_0), concluindo que existem diferenças significativas entre eles, podemos estar cometendo o chamado Erro Tipo I: *rejeitar H_0 quando verdadeira.* Os testes são construídos com a probabilidade deste erro fixada num nível bastante baixo, designada por α (nível de significância do teste). Nas ciências sociais, é comum usar $\alpha = 0,05$. Por outro lado, quando o teste aceita H_0, pode ocorrer o chamado Erro Tipo II: *aceitar H_0 quando falsa.* A probabilidade de se cometer este erro é designada por β. É desejável que, quando a diferença real entre os grupos for grande em termos práticos, a probabilidade β seja pequena; e para que isto aconteça, a quantidade n de elementos em cada grupo deve ser suficientemente grande.

A discussão que segue restringe-se a um teste bilateral para comparar duas amostras independentes em termos de médias, conforme discutido na Seção 11.4. Sejam μ_1 e μ_2 as médias das duas populações em estudo e seja:

[7] Para realizar a comparação entre mais de dois grupos, existem técnicas estatísticas mais apropriadas, conhecidas pelo nome de *Análise de variância*. Veja, por exemplo, em Barbetta, Reis e Bornia (1981).

$$\delta = \frac{|\mu_1 - \mu_2|}{\sigma}$$

A quantidade δ é a diferença de magnitude entre as verdadeiras médias em unidades de desvios padrões (σ) das populações em estudo. Supomos aqui que as duas populações tenham o mesmo desvio padrão.

Para avaliarmos o número n de elementos em cada grupo, o pesquisador precisa ser capaz de fornecer o valor mínimo de δ que leva a consequências práticas. Em geral, o pesquisador tem maior facilidade de raciocinar em termos da unidade em que se está medindo a variável em análise, mas, neste caso, é necessário termos uma avaliação de σ.

A Figura 11.11 indica o mínimo n para que uma diferença δ seja detectada pelo teste estatístico, com probabilidade 0,80 (β = 0,20) e com probabilidade 0,90 (β = 0,10).

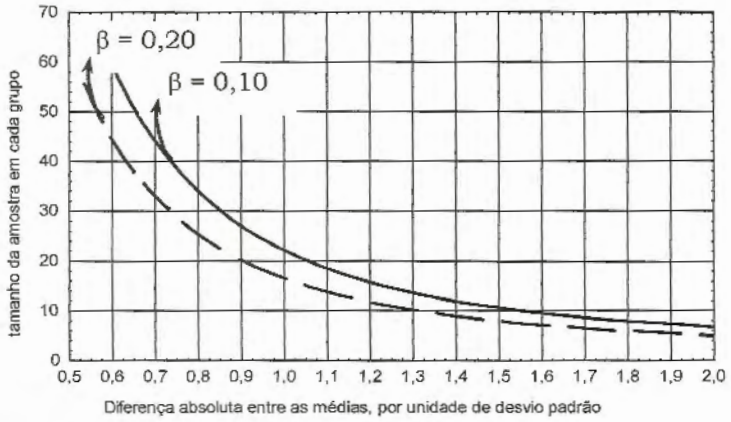
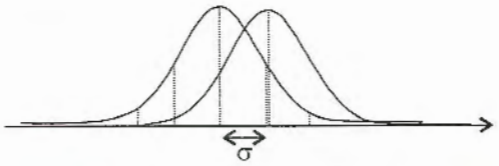


Figura 11.11 Tamanho mínimo da amostra, n, em cada grupo, em função da distância $\delta = |\mu_1 - \mu_2|/\sigma$ que se deseja detectar no teste estatístico.

Como exemplo, seja o problema de comparar dois métodos de ensinar matemática para crianças. Dois grupos de crianças devem ser formados, a fim de que os dois métodos sejam aplicados (um método em cada grupo). No final do estudo, o aprendizado de cada criança será avaliado numa escala de 0 a 10. Suponha que os pesquisadores consideram relevante uma diferença de 1,5 pontos entre as médias e, com base em estudos anteriores, o desvio padrão nesta escala não deve passar de duas unidades. Logo, $\delta = {}^{1,5}/_{2} = 0,75$. Pelo gráfico da Figura 11.11, o número mínimo de crianças em cada grupo deve ser de, aproximadamente, $n = 37$ para β = 0,10, ou $n = 28$ para β = 0,20.

Exercícios

13) Com o objetivo de comparar dois métodos de ensino, planeja-se um experimento com dois grupos de crianças (divididas aleatoriamente), sendo que em cada um dos grupos será aplicado um método de ensino. Quantas crianças devem ter em cada grupo, para garantir que um teste t bilateral para amostras independentes, ao nível de significância de 5%, detecte uma diferença de *um* desvio padrão, com 90% de probabilidade? Supondo distribuição normal, a diferença mínima que se quer detectar está representada na figura a seguir:



11.6 Comentários adicionais

Na Seção 11.3 descrevemos o teste t para dados pareados, e na Seção 11.4 o teste t para amostras independentes. A escolha do teste depende do planejamento da pesquisa, o qual pode gerar duas amostras de observações pareadas ou duas amostras de observações independentes. Mas o planejamento da pesquisa deve ser realizado da maneira mais adequada para o problema em questão. Em geral, quando é possível formar pares, tem-se maior controle sobre a variabilidade aleatória e, consequentemente, tem-se um projeto de pesquisa melhor. Por exemplo, no problema de se comparar dois tipos de materiais em termos do desgaste na sola de tênis de criança. Podemos planejar um experimento em que um grupo de crianças usa tênis com solas feitas com o material A e outro grupo usa tênis com solas feitas com o material B. Para cada criança, decidimos por sorteio qual material vai ser usado (*aleatorização*). Depois de algum tempo, medimos o desgaste das solas de todas as crianças do experimento e comparamos as médias das duas amostras através do teste t para amostras independentes.

Um projeto experimental alternativo é fabricar, para o estudo, pares de tênis com os diferentes tipos de sola, isto é, com um dos pés (alternando direito e esquerdo) com material A e o outro pé com material B. As crianças do experimento usam os dois tipos de materiais, fazendo com que a

comparação seja feita em cada criança, destacando uma possível diferença entre os tipos de materiais. Neste segundo planejamento, a comparação entre os materiais deve ser feita pelo teste *t* para dados pareados. A Figura 11.12 ilustra a diferença entre usar pares e usar duas amostras independentes na análise dos dados.

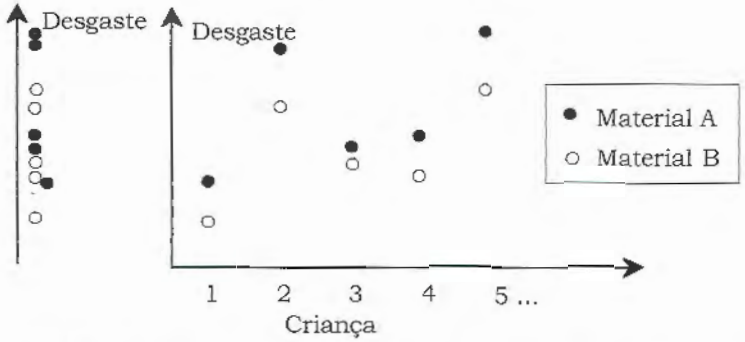


Figura 11.12 Um conjunto de dados visto de forma pareada (à direita) e de forma independente (à esquerda).

Analisando a Figura 11.12, fica evidente que, ao olhar os dados de forma pareada, tem-se mais informação sobre uma possível diferença entre os dois tipos de material. Observando as amostras de forma independente, as diferenças entre os dois tipos de material ficam ofuscadas pelas diferenças entre as crianças.

A aplicação de testes *t* pode ser feita em estudos experimentais ou em estudos de levantamento. No exemplo precedente, temos um estudo experimental, pois o pesquisador determina o material a ser aplicado em cada pé da criança, seja no primeiro ou no segundo caso. Se o teste rejeitar H_0, além de concluirmos que existe diferença significativa entre os dois grupos de valores, também concluímos que esta diferença deve-se ao material usado na sola do tênis (o único fator agindo sistematicamente e de forma diferenciada nos dois grupos). Assim, a aplicação de testes estatísticos em estudos experimentais permite verificar hipóteses de *causa e efeito*.

Por outro lado, se quisermos comparar o peso de recém-nascidos em duas localidades, podemos fazer um levantamento por amostragem, analisando os nascimentos nessas localidades. Neste caso, as duas amostras já estão naturalmente divididas pela localidade em que reside a mãe da criança. Com a aplicação do teste *t*, podemos detectar uma diferença significativa entre as duas localidades. Mas a inferência sobre a causa da diferença é mais difícil do que num estudo experimental, pois

podem existir diversos fatores, tais como etnia, condições socioeconômicas, hábitos de alimentação, etc., agindo de forma interativa e possivelmente diferenciada nas duas localidades (veja o Exercício 8).

Outro aspecto que merece comentários é a implicação prática de uma diferença *estatisticamente significativa*. Uma diferença significativa é uma diferença que não deve ter ocorrido meramente por acaso, mas não, necessariamente, é uma diferença relevante em termos práticos. Quando se analisam amostras grandes, os testes podem concluir que pequenas diferenças são significativas. Resta a análise prática para verificar se essas diferenças, estimadas pelos dados, são relevantes.

Existe uma grande quantidade de testes estatísticos para comparação entre duas amostras. Neste capítulo, demos ênfase aos *testes t* por serem os mais usados. Contudo, em muitas situações, as suposições desses testes podem estar sendo violadas. Quando isto ocorrer, devemos procurar técnicas alternativas, em especial os chamados *testes não paramétricos*, que não supõem uma determinada distribuição de probabilidades como geradora dos dados.[8] O teste dos sinais é um exemplo de teste não paramétrico, assim como o qui-quadrado, que será estudado no capítulo seguinte.

O Quadro 11.1 mostra alguns testes para comparação de duas amostras, segundo o tipo de variável e condição das amostras.

Quadro 11.1 Alguns testes para comparação de duas amostras

Amostras	Variável	
	Qualitativa	Quantitativa
Pareadas	Teste dos sinais (Seção 11.2)	Teste *t* pareado (Seção 11.3)
Independentes	Teste qui-quadrado (Seção 12.1)	Teste *t* amostras independentes (Seção 11.4)

Exercícios complementares

14) Uma cervejaria estuda a possibilidade de alterar o rótulo de uma de suas marcas, usando formas e cores mais vivas. Para avaliar se existe vantagem em alterar o rótulo, a empresa levou a cabo uma pesquisa de *marketing*. Enlatou a cerveja com o rótulo tradicional e com o rótulo novo. A pesquisa foi feita em oito estabelecimentos comerciais. Em quatro deles, extraídos por

[8] Os testes *t* supõem que os dados provenham de distribuições normais e as populações tenham, aproximadamente, a mesma variância.

sorteio, colocou-se o produto com o rótulo novo e, nos outros quatro, manteve-se o produto com o rótulo tradicional. Após um mês, avaliou-se a quantidade vendida em cada estabelecimento. Os estabelecimentos que usaram o rótulo tradicional tiveram os seguintes resultados nas vendas (em milhares de unidades): 6, 5, 2, 2. Os estabelecimentos que usaram o rótulo novo tiveram os seguintes resultados nas vendas (em milhares de unidades): 4, 9, 5, 6. Os dados mostram evidência suficiente de que a média de vendas é superior com o rótulo novo? Aplique um teste estatístico apropriado, ao nível de significância de 5%.

15) Para o mesmo problema da questão anterior, outro instituto de pesquisa, que tem uma equipe com melhor preparação em estatística, elaborou um projeto um pouco diferente. Com seis estabelecimentos comerciais dispostos a colaborar com a pesquisa, colocaram-se as duas embalagens (de rótulo tradicional e de rótulo novo) da mesma cerveja. Tomou-se o cuidado para que em cada estabelecimento, a apresentação das duas embalagens do produto fosse feita de forma idêntica. Os resultados das vendas mensais (em milhares de unidades), foram os seguintes:

Estabelecimento:	1	2	3	4	5	6
Rótulo tradicional:	16	12	28	32	19	25
Rótulo novo:	20	11	33	40	21	31

Os dados mostram evidência suficiente de que a média de vendas é superior com o rótulo novo? Use nível de significância de 5%.

16) Com respeito à questão anterior, suponha que os gerentes dos estabelecimentos comerciais se recusaram a fornecer os valores das vendas, mas informaram com qual rótulo as vendas foram maiores. Nos estabelecimentos 1, 3, 4, 5 e 6 as vendas foram maiores com o rótulo novo, e no estabelecimento dois as vendas foram maiores com o rótulo tradicional. Esses dados são suficientes para afirmar que a maioria dos estabelecimentos vende mais cerveja com o rótulo novo? Use nível de significância de 5%.

17) Com o objetivo de avaliar o efeito de uma merenda escolar reforçada, foi realizado um estudo com dois grupos de crianças que tinham princípios de desnutrição. Fizeram parte do estudo sete pares de crianças. Em cada par, as crianças tinham peso e idade similares. As crianças de cada par foram divididas em dois grupos, sendo um tratado com merenda "reforçada" (Grupo A) e o outro com merenda convencional (Grupo B). Os dados a seguir apresentam o ganho de peso, em kg, durante seis meses.

Grupo	Par de crianças						
	1	2	3	4	5	6	7
A	6	5	8	2	5	4	4
B	2	4	5	3	4	5	5

Esses dados mostram evidência suficiente para garantir que crianças tratadas com a merenda reforçada ganham, em média, mais peso do que crianças tratadas com merenda convencional? Justifique sua resposta através de um teste estatístico adequado, ao nível de significância de 10%.

18) Um estudo sobre a identidade social dos professores com o departamento a que pertencem, mostrou os seguintes resultados (quanto maior o escore maior identidade social com o departamento):

Deptº de Arquitetura: amostra de 24 professores, média de 40,8 e desvio padrão de 5,9 pontos.

Deptº de Psicologia: amostra de 19 professores, média de 42,5 e desvio padrão de 5,4 pontos.

Esses dados mostram evidência suficiente de que, em média, a identidade social com o departamento é diferente quando comparamos os departamentos de Arquitetura e Psicologia? Explique.

19) Para avaliar o governo perante os empresários, um instituto de pesquisa realisou duas pesquisas: a primeira, logo após a posse do governo, com uma amostra aleatória de 200 empresários, em que a nota média foi de 7,0 pontos, com desvio padrão de 2,0 pontos; a segunda, após seis meses, com outra amostra aleatória de 200 empresários, que mostrou aprovação média de 6,0 pontos, com desvio padrão de 3,0 pontos. Os dados mostram evidência suficiente para afirmar que, na população de empresários, houve redução na aprovação ao governo? Use $\alpha = 0{,}01$.

Parte V

Relação entre variáveis

Como medir e testar a significância da associação entre duas variáveis qualitativas

Como estudar a correlação entre duas variáveis quantitativas

Como construir modelos para o relacionamento entre duas variáveis

Capítulo 12

Análise de dados categorizados

Grande parte das variáveis estudadas nas Ciências Humanas e Sociais não é mensurada numericamente, mas somente permitem alocar cada elemento em categorias preestabelecidas. A observação dos elementos da amostra resulta em *dados categorizados*. Por exemplo, ao observar a variável *sexo* (*gênero*), cada indivíduo pesquisado deve ser alocado na categoria *masculino* ou na categoria *feminino*. Lembramos que as variáveis devem estar bem definidas, de maneira que cada elemento pesquisado se encaixe em uma (e apenas em uma) categoria.

Comparação entre amostras

O teste qui-quadrado, que será estudado neste capítulo, poderá ser usado em problemas de pesquisas com amostras *independentes*, análogos aos discutidos no capítulo anterior, porém com a variável resposta *qualitativa* (*categórica*), como, por exemplo, na comparação de métodos de ensino para vestibulandos, em que a variável resposta é o resultado no vestibular (para cada aluno: *aprovado* ou *reprovado*). Outro exemplo: na comparação das populações de homens e mulheres quanto ao tabagismo (*fumante* ou *não fumante*).

Análise de associação

Um dos grandes propósitos em pesquisas nas Ciências Sociais é verificar se duas ou mais variáveis se apresentam *associadas*.

Existe *associação* entre duas variáveis se o conhecimento de uma altera a probabilidade de algum resultado da outra.

Podemos dizer que existe associação entre o *clima* e *a propensão de uma pessoa ir à praia*, porque é maior a probabilidade de a pessoa ir à praia num dia quente e ensolarado do que num dia frio e chuvoso. Ou seja, o conhecimento do clima altera a probabilidade de a pessoa ir à praia, o que caracteriza uma associação.[1]

Neste capítulo, estudaremos como *testar* uma possível associação entre duas variáveis qualitativas, com base numa amostra de observações. Veremos, também, maneiras de *medir* o grau de associação descrito pela amostra.

12.1 O TESTE DE ASSOCIAÇÃO QUI-QUADRADO

O *teste qui-quadrado* é o teste estatístico mais antigo e um dos mais usados em pesquisa social. É um método que permite testar a significância da associação entre duas variáveis qualitativas, como também comparar (no sentido de teste de significância) duas ou mais amostras, quando os resultados da variável resposta estão dispostos em categorias.

EXEMPLO 12.1 Para estudar a associação entre sexo (*masculino* ou *feminino*) e tabagismo (*fumante* ou *não fumante*), numa certa população, foi observada uma amostra aleatória de 300 pessoas adultas dessa população, fazendo-se a classificação segundo o sexo e tabagismo. Os dados estão apresentados na Tabela 12.1.

Tabela 12.1 Distribuição de 300 pessoas, classificadas segundo o sexo e tabagismo

Tabagismo	Sexo		Total
	Masculino	Feminino	
Fumante (%)	92 (46,0)	38 (38,0)	130 (43,3)
Não fumante (%)	108 (54,0)	62 (62,0)	170 (56,7)
Total (%)	200 (100,0)	100 (100,0)	300 (100,0)

A Tabela 12.1 é uma tabela de contingência, de dimensão 2x2, mostrando os resultados de uma amostra de 300 indivíduos, classificados, simultaneamente, com respeito às variáveis *sexo* e *tabagismo*. O objetivo é verificar se os dados da amostra mostram evidência suficiente para afirmarmos que, na população em estudo, existe associação entre *sexo* e *tabagismo*.

[1] A existência de associação entre X e Y não implica, necessariamente, que X causa Y, ou que Y causa X.

Agora, considere que o projeto de obtenção de dados tivesse sido um pouco diferente: duas populações (a de homens e a de mulheres) e uma única variável resposta: tabagismo (*fumante* ou *não fumante*). Poderíamos ter interesse em testar se existe diferença significativa entre a proporção de homens fumantes e a proporção de mulheres fumantes. Formalmente, teríamos as seguintes hipóteses:

$$H_0: \pi_h = \pi_m \quad \text{e} \quad H_1: \pi_h \neq \pi_m$$

onde π_h é a proporção de homens fumantes e π_m é a proporção de mulheres fumantes, nas populações em estudo.[2]

Desconsiderando a questão do planejamento da pesquisa (uma ou duas populações), se $\pi_h = \pi_m$, então o conhecimento do sexo não fornece qualquer conhecimento sobre o fato de o indivíduo ser ou não fumante. Neste contexto, a hipótese nula pode ser escrita como:

H_0: *Sexo* e *tabagismo* são variáveis *independentes*, na população em estudo.

Por outro lado, se $\pi_h \neq \pi_m$, então o conhecimento do sexo aumenta (ou diminui) a chance de o indivíduo ser fumante. Logo,

H_1: Existe *associação* entre as variáveis *sexo* e *tabagismo*, na população em estudo.

■

Exemplo 12.2 Com o objetivo de verificar se três localidades são diferentes em termos do *nível de instrução*, foram selecionadas amostras aleatórias de indivíduos adultos nessas localidades, fazendo-se a classificação segundo o nível de instrução. Os resultados estão apresentados na Tabela 12.2.

Tabela 12.2 Distribuição de frequências do nível de instrução, segundo a localidade da residência.

Nível de instrução	Localidade		
	Monte Verde	Parque da Figueira	Encosta do Morro
Nenhum (%)	6 (15,0)	14 (32,6)	18 (48,7)
Fundamental (%)	11 (27,5)	14 (32,6)	13 (35,1)
Médio ou superior (%)	23 (57,5)	15 (34,8)	6 (16,2)
Total (%)	40 (100,0)	43 (100,0)	37 (100,0)

[2] Para o problema específico de testar duas proporções, também pode ser aplicado o chamado *teste Z de diferença entre duas proporções*, o qual usa a distribuição normal como referência e permite a abordagem unilateral. Para maiores detalhes, ver, por exemplo, Stevenson (1981, p. 282) ou Triola (2005, p. 336).

Aprendemos, no Capítulo 4, a interpretar uma tabela em termos descritivos, ou seja, tirar informações dos dados tabulados, sem se preocupar com generalizações. Contudo, se os dados são de amostras, podemos *testar* se as diferenças são significativas, isto é, se os dados mostram evidência suficiente para inferirmos que existem diferenças também nas populações de onde eles foram extraídos. Formalmente, podemos testar as seguintes hipóteses:

H_0: As distribuições de frequências do nível de instrução *são iguais* nas três localidades;

H_1: As distribuições de frequências do nível de instrução *não* são iguais nas três localidades.

Se considerarmos que as três localidades formam categorias da variável *localidade da residência*, podemos colocar as hipóteses em termos de independência (H_0) e associação (H_1).[3]

■

Dadas duas variáveis qualitativas, as hipóteses do *teste qui-quadrado* podem ser formuladas como:

H_0: As duas variáveis são *independentes*.

H_1: Existe *associação* entre as duas variáveis.

ESTATÍSTICA DO TESTE

Chamaremos de *célula* a cada cruzamento de linha e coluna de uma tabela de contingência.

A estatística do teste, que designaremos por χ^2 (qui-quadrado), é uma espécie de medida de distância entre as frequências observadas e as frequências que esperaríamos encontrar em cada célula, na suposição de as variáveis serem independentes (H_0 verdadeira). Ilustraremos a obtenção das frequências esperadas e da estatística χ^2, usando os dados da Tabela 12.1.

EXEMPLO 12.1 (CONTINUAÇÃO) Para obter as frequências esperadas, seja a distribuição percentual de fumantes e não fumantes em toda a amostra (43,3% de fumantes e 56,7% de não fumantes). Se *tabagismo* e *sexo* forem variáveis

[3] Muitos autores preferem considerar a presente situação como um teste de *homogeneidade* entre as amostras das diferentes localidades, já que no presente contexto a *localidade da residência* não é propriamente uma variável, mas sim uma referência às populações (ou aos subgrupos da população) em estudo. Porém, o teste qui-quadrado pode ser aplicado da mesma maneira.

independentes (H_0 verdadeira), devemos esperar que estas percentagens se mantenham, tanto no estrato dos *homens*, como no estrato das *mulheres*. Como foram observados 200 homens, devemos esperar em torno de:

43,3% de 200 homens fumantes [(0,433)×(200) = 86,6)] e
56,7% de 200 homens não fumantes [(0,567)×(200) = 113,4].

De forma análoga, podemos obter as frequências esperadas no estrato das mulheres.

■

O cálculo das *frequências esperadas* pode ser simplificado com a aplicação da seguinte fórmula, aplicada a cada célula da tabela de contingência:

$$E = \frac{(\textit{total da linha}) \times (\textit{total da coluna})}{(\textit{total geral})}$$

A estatística do teste qui-quadrado é definida por

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

onde: a soma se estende a todas as células da tabela de contingência;
O representa a frequência observada na célula; e
E representa a frequência esperada na célula.

Exemplo 12.1 (continuação) Cálculo das frequências esperadas:

Tabagismo	Sexo		
	Masculino	Feminino	Total
Fumante	$E = \frac{130 \times 200}{300} = 86{,}67$	$E = \frac{130 \times 100}{300} = 43{,}33$	130
Não fumante	$E = \frac{170 \times 200}{300} = 113{,}33$	$E = \frac{170 \times 100}{300} = 56{,}67$	170
Total	200	100	300

Cálculo das parcelas da estatística qui-quadrado:

Tabagismo	Sexo	
	masculino	feminino
Fumante	$\frac{(92-86{,}67)^2}{86{,}67} = 0{,}328$	$\frac{(38-43{,}33)^2}{43{,}33} = 0{,}656$
Não fumante	$\frac{(108-113{,}33)^2}{113{,}33} = 0{,}251$	$\frac{(62-56{,}67)^2}{56{,}67} = 0{,}501$

Assim, temos o valor da estatística qui-quadrado:

$$\chi^2 = 0{,}328 + 0{,}656 + 0{,}251 + 0{,}501 = 1{,}74$$

■

Quando as variáveis são *independentes* (H_0 verdadeira), as frequências observadas tendem a ficar perto das frequências esperadas: *apenas variações casuais!* Neste caso, o valor de χ^2 deve ser pequeno. Em outras palavras, um valor pequeno de χ^2 sugere que as variáveis podem ser independentes. Por outro lado, um valor grande na estatística χ^2, sinaliza que as diferenças entre as frequências observadas e frequências esperadas não devem ser meramente casuais, ou seja, deve haver *associação* entre as duas variáveis.

DISTRIBUIÇÃO DE REFERÊNCIA

Precisamos de uma distribuição de referência que permita julgar se um determinado valor da estatística χ^2 pode ser considerado grande o suficiente para rejeitar H_0, em favor de H_1. Suposições básicas para usar a chamada *distribuição qui-quadrado* como referência:

1) que os dados estejam dispostos numa tabela de contingência propriamente dita, isto é, cada elemento observado é alocado *numa e apenas numa* célula; e
2) que as amostras sejam grandes.

A verificação da adequação dos tamanhos das amostras é usualmente feita em termos das frequências esperadas. A maioria dos autores considera adequada a aplicação do teste qui-quadrado quando *todas* as frequências esperadas forem maiores ou iguais a 5 (cinco).[4] No exemplo em discussão, as frequências esperadas foram: 86,67, 43,33, 113,33 e 56,67. Portanto, todas superiores a 5, o que permite a realização do teste qui-quadrado.

> Supondo H_0 verdadeira e as condições (1) e (2), então os possíveis valores da estatística χ^2 seguem a chamada *distribuição qui-quadrado* com $gl=(\ell-1)\cdot(c-1)$ *graus de liberdade*, onde ℓ é o número de linhas e c é o número de colunas da tabela.

No Exemplo 12.1, ambas as variáveis têm duas categorias (tabela 2x2), então $\ell = 2$, $c = 2$ e, portanto, $gl = (2-1)\cdot(2-1) = 1$. Logo, se H_0 for

[4] Quando ocorrer alguma frequência esperada menor do que cinco, pode-se aplicar o chamado teste exato de Fisher. Veja, por exemplo, Levin (1985, p. 221).

verdadeira, os possíveis valores da estatística χ^2 devem seguir uma distribuição qui-quadrado com $gl = 1$ grau de liberdade. A forma da distribuição qui-quadrado torna-se menos assimétrica à medida que cresce o número de graus de liberdade (veja a Figura 12.1).

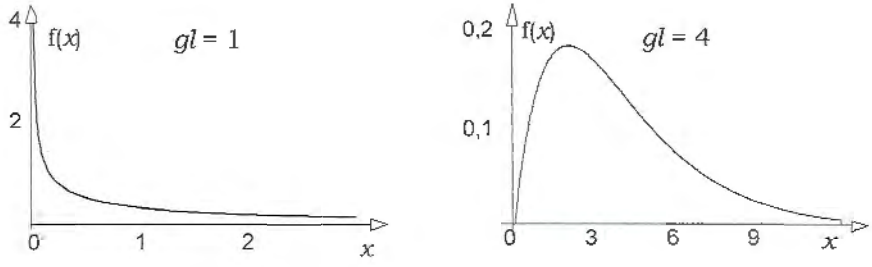


Figura 12.1 Distribuições qui-quadrado com $gl = 1$ e $gl = 4$.

Probabilidade de significância

Supondo que as duas variáveis sejam independentes (H_0 verdadeira), o valor p é a probabilidade de a estatística qui-quadrado acusar um valor maior ou igual do que o valor do χ^2, calculado com base na amostra (ver Figura 12.2).

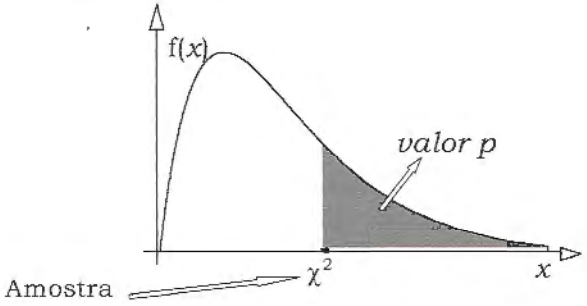


Figura 12.2 A probabilidade de significância (*valor p*) como uma área sob a curva da distribuição qui-quadrado.

Quando os dados levam a um χ^2 grande (e, em consequência, um valor p pequeno), o teste rejeita H_0, em favor de H_1. Por outro lado, quando os dados observados levam a um χ^2 pequeno (e, em consequência, um valor p grande), o teste não rejeita H_0, porque o valor calculado de χ^2 está condizente com a distribuição dos possíveis valores de qui-quadrado, construída à luz de H_0.

Conforme apresentado no Capítulo 10, adotado um nível de significância α, a decisão do teste estatístico é:

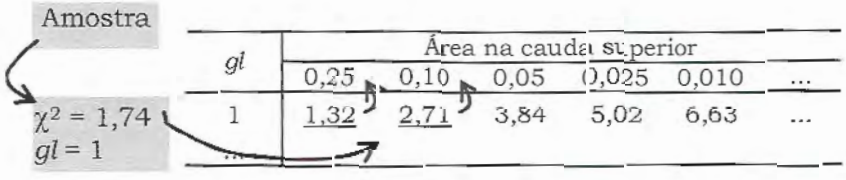
$p > \alpha \rightarrow$ aceita H_0

$p \leq \alpha \rightarrow$ rejeita H_0, em favor de H_1

Tabela da distribuição qui-quadrado

Depois de calculado o valor da estatística c^2, podemos obter a probabilidade de significância p, usando uma tabela da distribuição qui-quadrado (Tabela 6 do apêndice). A continuação do Exemplo 12.1 ilustra o uso dessa tabela.

Exemplo 12.1 (continuação) Usando a Tabela 6 do apêndice, entramos na linha correspondente a $gl = 1$. Verificamos que o valor calculado $\chi^2 = 1,74$ está entre os valores 1,32 e 2,71 da tabela, os quais estão associados às áreas na cauda superior iguais a 0,25 e 0,10, respectivamente, conforme ilustra o seguinte esquema:



Logo, para o valor calculado $\chi^2 = 1,74$, temos o valor p entre 0,10 e 0,25. Usando o nível usual de significância de 5% ($\alpha = 0,05$), o teste aceita H_0 (pois, $p > \alpha$). Concluímos, então, que os dados não mostram evidência de associação entre *sexo* e *tabagismo*, na população em estudo. Em outras palavras, a diferença verificada na amostra entre a proporção de homens fumantes e a proporção de mulheres fumantes pode ser explicada, meramente, por variações casuais da amostragem.

■

Correção de continuidade em tabelas 2 x 2

Já comentamos que a distribuição qui-quadrado, usada como distribuição de referência para a estatística χ^2, só é válida para amostras grandes. Em tabelas de dimensão 2 x 2, especialmente quando as amostras não forem muito grandes (por exemplo, quando existir alguma frequência

esperada entre 5 e 10), recomendamos aplicar a chamada *correção de continuidade de Yates*, que consiste em reduzir 0,5 unidade nas diferenças absolutas entre as frequências observadas e esperadas. Assim, para tabelas de contingência 2 x 2,

$$\chi^2 = \sum \frac{(|O-E|-0{,}5)^2}{E}$$

Ou seja, em cada célula, depois de calcular a diferença entre O e E, devemos desprezar o sinal (+ ou -) e reduzir 0,5 unidade. Em seguida, elevamos ao quadrado, e dividimos pela frequência esperada da célula.

Vamos refazer o cálculo do χ^2 do Exemplo 12.1, usando a correção de continuidade. Primeiramente, faremos o cálculo das parcelas do χ^2, referentes a cada célula:

Tabagismo	Sexo	
	Masculino	Feminino
Fumante	$\frac{(\lvert 92-86{,}67\rvert-0{,}5)^2}{86{,}67} = 0{,}269$	$\frac{(\lvert 38-43{,}33\rvert-0{,}5)^2}{43{,}33} = 0{,}538$
Não fumante	$\frac{(\lvert 108-113{,}33\rvert-0{,}5)^2}{113{,}33} = 0{,}206$	$\frac{(\lvert 62-56{,}67\rvert-0{,}5)^2}{56{,}67} = 0{,}412$

Resultando em: $\chi^2 = 0{,}269 + 0{,}538 + 0{,}206 + 0{,}412 = 1{,}43$.

Usando a Tabela 6 com $gl = 1$, encontramos a probabilidade de significância na mesma faixa do caso anterior, isto é, $0{,}10 < p < 0{,}25$.

■

Quando as amostras não forem muito grandes, o uso da correção de continuidade pode levar a resultados bastante diferentes (veja o Exercício 1). É justamente neste caso que a correção é mais recomendada.

Uma fórmula mais rápida para o cálculo do χ^2 em tabelas 2 x 2

Em tabelas 2 x 2, representadas segundo o esquema abaixo, podemos calcular a estatística χ^2, com correção de continuidade, da seguinte forma:

a	b	$a+b$
c	d	$c+d$
$a+c$	$b+d$	n

$$\chi^2 = \frac{n \cdot \left(|ad-bc| - \frac{n}{2}\right)^2}{(a+b)\cdot(c+d)\cdot(a+c)\cdot(b+d)}$$

Vamos ilustrar o uso desta fórmula com os dados da Tabela 12.1:

$a = 92$	$b = 38$	$a + b = 130$
$c = 108$	$d = 62$	$c + d = 170$
$a + c = 200$	$b + d = 100$	$n = 300$

Assim,

$$\chi^2 = \frac{300 \times \left(\left|92 \times 62 - 38 \times 108 - \frac{300}{2}\right|\right)^2}{130 \times 170 \times 200 \times 100} = \frac{300 \times [1.600 - 150]^2}{442.000.000} = \frac{300 \times (2.102.500)}{442.000.000} = 1,43$$

Para calcular a estatística χ^2 sem a correção de continuidade, basta excluir a fração $n/2$ do numerador.

Aplicação do teste qui-quadrado em tabelas de grande dimensão

Exemplo 12.3 (BOX; HUNTER; HUNTER, 1978, p. 145.) Num estudo exploratório está se examinando a recuperação funcional de pacientes submetidos a um certo ato cirúrgico, em cinco hospitais de uma cidade. Os hospitais A, B, C e D são hospitais comuns, enquanto o Hospital E é um hospital de referência, o qual recebe os casos mais graves. A Tabela 12.3 mostra os resultados de um levantamento por amostragem, realizado nos cinco hospitais.

Tabela 12.3 Resultados (frequências e percentagens) da recuperação funcional de pacientes, submetidos a um certo procedimento cirúrgico, em cinco hospitais.

Recuperação funcional	Hospital				
	A	B	C	D	E
Nenhuma (%)	13 (27,7)	5 (16,1)	8 (10,1)	21 (16,4)	43 (52,4)
Parcial (%)	18 (38,3)	10 (32,3)	36 (45,6)	56 (43,8)	29 (35,4)
Completa (%)	16 (34,0)	16 (51,6)	35 (44,3)	51 (39,8)	10 (12,2)

Com o objetivo de verificar se realmente existe associação entre *hospital* e *recuperação do paciente*, vamos realizar o teste qui-quadrado. A Tabela 12.4 mostra as frequências esperadas e as parcelas de cada célula no cálculo da estatística χ^2.

Tabela 12.4 Resultados do procedimento cirúrgico: frequências observadas (centro), frequências esperadas (canto superior direito) e parcelas do χ^2 (canto inferior esquerdo).

Recuperação funcional	Hospital					Total
	A	B	C	D	E	
Nenhuma	11,53 13 0,19	7,60 5 0,89	19,37 8 6,67	31,39 21 3,44	20,11 43 26,05	90
Parcial	19,08 18 0,06	12,59 10 0,53	32,07 36 0,48	51,94 56 0,31	33,39 29 0,55	149
Completa	16,39 16 0,01	10,81 16 2,49	27,55 35 2,02	44,64 51 0,91	28,60 10 12,10	128
Total	47	31	79	128	82	367

Somando os valores das parcelas do χ^2, temos:

$$\chi^2 = 56{,}7$$

com

$$gl = (\ell - 1)\cdot(c - 1) = (3 - 1)\cdot(5 - 1) = 8$$

Pela Tabela 6 do apêndice, verificamos que a probabilidade de significância p é inferior a 0,0005. Então, para qualquer nível usual de significância (por exemplo, $\alpha = 0{,}05$), o teste detecta associação entre *recuperação funcional de pacientes* e *hospital* (pois, $p < \alpha$). Em outras palavras, o teste qui-quadrado mostrou que os hospitais em estudo são diferentes quanto à recuperação funcional de seus pacientes.

■

Muitas vezes, ao analisar uma tabela de grande dimensão, temos também o interesse em estudar partes desta tabela para entendermos melhor uma eventual associação entre duas variáveis. Podemos comparar grupos de categorias agregadas segundo algum critério e, posteriormente, estudar separadamente as categorias que estavam agrupadas.

Exemplo 12.3 (continuação) Observando as parcelas do χ^2 (canto inferior esquerdo das células da Tabela 12.4), verificamos que as maiores contribuições partiram do Hospital E que é um hospital de referência e recebe os casos mais graves. Podemos, então, fazer uma análise estatística para verificar se a significância foi em razão de diferenças entre os hospitais comuns e o hospital de referência, somente entre os hospitais comuns, ou ambos os casos.

A Tabela 12.5 agrega todos os hospitais comuns (A, B, C e D) para confrontar com o hospital de referência E. O valor das frequências observadas na coluna dos hospitais comuns corresponde à soma das frequências observadas dos hospitais A, B, C e D da Tabela 12.3. As frequências esperadas e as parcelas do χ^2 foram calculadas novamente.

Tabela 12.5 Comparação do hospital de referência com os demais. Frequências observadas (centro), frequências esperadas (canto superior direito) e parcelas do χ^2 (canto inferior esquerdo).

Recuperação funcional	Hospitais comuns (A + B + C + D)	Hospital de referência (E)	Total
Nenhuma	69,89 47 7,50	20,11 43 26,05	90
Parcial	115,71 120 0,16	33,29 29 0,55	149
Completa	99,40 118 3,48	28,60 10 12,10	128
Total	285	82	367

Temos: $\chi^2 = 49{,}8$ e $gl = 2$. Usando a Tabela 6, verificamos que $p < 0{,}001$, mostrando haver diferença significativa entre os hospitais comuns e o hospital de referência. Finalmente, a Tabela 12.6 analisa os hospitais comuns entre si. As frequências observadas dessa tabela correspondem às frequências observadas da Tabela 12.3, eliminando o Hospital E.

Tabela 12.6 Comparação entre os hospitais comuns. Frequências observadas (centro), frequências esperadas (canto superior direito) e parcelas do χ^2 (canto inferior esquerdo).

Recuperação funcional	Hospital				Total
	A	B	C	D	
Nenhuma	7,75 13 3,55	5,11 5 0,00	13,03 8 1,94	21,11 21 0,00	47
Parcial	19,79 18 0,16	13,05 10 0,71	33,26 36 0,23	53,89 56 0,08	120
Completa	19,46 16 0,61	12,84 16 0,78	32,71 35 0,16	53,00 51 0,08	118
Total	47	31	79	128	285

Temos: $\chi^2 = 8{,}4$, $gl = 6$ e, portanto, $0{,}10 < p < 0{,}25$. Considerando o nível de significância de 5% ($\alpha = 0{,}05$), ou até mesmo de 10% ($\alpha = 0{,}10$), o teste não detecta associação. Assim, podemos dizer que não há diferença significativa entre os hospitais comuns.

■

Uso do computador

Considerando o anexo do Capítulo 4, vamos verificar se existe associação significativa entre o *local da residência* e a *utilização de programas de alimentação popular*. A Figura 12.3 mostra uma saída do pacote computacional SPSS®.[5]

Programa de alimentação popular * Local da residência Crosstabulation

		Local da residência			
		Monte Verde	Parque da Figueira	Morro	Total
Programa de alimentação popular	não usa	18	12	12	42
	usa	22	31	25	78
Total		40	43	37	120

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2,816[a]	2	0,245
Likelihood Ratio	2,791	2	0,248
Linear-by-Linear Association	1,388	1	0,239
N of Valid Cases	120		

[a] 0 cells (,0%) have expected count less than 5. The minimum expected count is 12,95.

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Cramer's V	0,153	0,245
	Contingency Coefficient	0,151	0,245
N of Valid Cases		120	

Figura 12.3 Tabela de contingência, teste qui-quadrado e medidas de associação pelo SPSS.

Com os dados brutos é construída uma tabela de contingência. O teste qui-quadrado (*Pearson Chi-square*) é apresentado na segunda tabela com os resultados $\chi^2 = 2{,}816$, $gl = 2$ e $p = 0{,}245$ (segunda tabela da Figura 12.3), mostrando não haver associação (aceitando H_0). Também são

[1] Ver www.spss.com

mostrados os resultados de outras abordagens do teste qui-quadrado, que não serão discutidos neste texto. No rodapé desta segunda tabela, diz-se que não há frequência esperada inferior a cinco, condição para a validade do teste. Finalmente, a terceira tabela apresenta algumas medidas de associação, que serão discutidas na próxima seção.

EXERCÍCIOS

1) Seja a seguinte amostra:

Classificação de uma amostra de 38 indivíduos, quanto à *ansiedade* e *tabagismo*

Fumante	Ansioso	
	sim	não
sim	15	7
não	6	10

a) Calcule a estatística χ^2 sem usar a correção de continuidade.
b) Calcule a estatística χ^2 usando a correção de continuidade.
c) Você pode dizer que existe associação entre *tabagismo* e *ansiedade*, ao nível de significância de 10%?

2) (LEVIN, 1985, p. 266) Dois grupos de estudantes fizeram exames finais de estatística. Somente um grupo recebeu preparação formal para o exame; o outro leu o texto recomendado, mas nunca compareceu às aulas. Enquanto 22 dos 30 membros do primeiro grupo (os *frequentadores*) passaram no exame, apenas 10 dos 28 do segundo grupo (os *ausentes*) lograram aprovação. Os dados mostram evidência suficiente para afirmar que existe associação entre *frequência às aulas* e *aprovação no exame final*? Use $\alpha = 0{,}05$.

3) a) Faça um teste qui-quadrado com os dados da Tabela 12.2 para verificar se existe diferença significativa entre as distribuições do nível de instrução nas três localidades. Use $\alpha = 0{,}01$.
b) Verifique se existe diferença significativa na distribuição do nível de instrução entre a Encosta do Morro e os conjuntos residenciais Monte Verde e Pq. da Figueira (agregados).
c) Verifique se existe diferença significativa na distribuição do nível de instrução entre os dois conjuntos residenciais.

4) Usando os dados do anexo do Capítulo 4, verifique se existe associação entre:
a) uso de programas de alimentação popular e localidade da residência;
b) uso de programas de alimentação popular e nível de instrução do chefe da casa.[6]

[6] Como já comentamos, a presença de associação entre duas variáveis não implica a existência de uma relação de *causa e efeito* entre elas. No Exercício 4.b, por exemplo, se houver associação entre *uso de programas de alimentação popular* e *nível de instrução do chefe da casa*, então esta pode ser devida a uma terceira variável: *renda familiar*, que por estar associada às duas variáveis em estudo, pode induzir uma associação entre elas.

12.2 Medidas de associação

Como vimos, a aplicação do teste qui-quadrado permite verificar se existe associação entre duas variáveis, com base em um conjunto de observações. É um processo de inferência, em que se parte dos dados para se tirar conclusões sobre o universo de onde os dados foram extraídos. Em muitas situações, porém, o interesse está restrito em descrever adequadamente a amostra, sem extrapolar para um universo maior. Neste contexto, ao invés de um teste estatístico, é mais interessante estudar o nível de associação descrito pela própria amostra.

Nesta seção, apresentaremos alguns coeficientes que têm por objetivo *medir a força da associação* entre duas variáveis categorizadas. Enfatizamos que essas medidas são descritivas, isto é, referem-se apenas aos dados observados. Porém, o cálculo dos coeficientes de associação também pode ser realizado após a aplicação de um teste estatístico, se este detecta associação. Neste caso, um coeficiente de associação fornece uma *estimativa* do grau de associação entre as duas variáveis.

Exemplo 12.4 Vamos contrapor duas amostras (A e B), classificadas segundo o sexo (*homem* ou *mulher*) e tabagismo (*fumante* ou *não fumante*).

Amostra A

Tabagismo	Sexo	
	homem	mulher
fumante	80 (40%)	40 (40%)
não fumante	120 (60%)	60 (60%)
Total	200 (100%)	100 (100%)

Amostra B

Tabagismo	Sexo	
	homem	mulher
fumante	200 (100%)	0 (0%)
não fumante	0 (0%)	100 (100%)
Total	200 (100%)	100 (100%)

Na amostra A, os dados indicam uma situação de completa *independência*, pois o conhecimento do sexo do respondente não fornece qualquer informação sobre a variável *tabagismo* (veja que a percentagem de homens fumantes é igual à percentagem de mulheres fumantes). Por outro lado, a amostra B ilustra um caso de *associação perfeita*, já que os fumantes são todos homens e todos os não fumantes são mulheres.

■

Um coeficiente de associação, aplicado a uma tabela de contingência, produz um valor numérico que descreve se os dados se aproximam mais de uma situação de independência ou de uma situação de associação perfeita. Ou seja, descreve o *quanto* os dados das duas variáveis se mostram associados.

A própria estatística χ^2, desenvolvida na seção anterior, pode ser usada como uma medida de associação. Efetuando o cálculo desta estatística sobre os dados das amostras A e B, sem a correção de continuidade, encontramos os seguintes valores: $\chi^2 = 0$ (amostra A) e $\chi^2 = 300$ (amostra B). Mas a interpretação da estatística χ^2 como um coeficiente de associação não é muito simples, pois o seu valor máximo (associação perfeita) varia de acordo com a dimensão da tabela e o número de elementos observados.

COEFICIENTE DE CONTINGÊNCIA

Um coeficiente muito usado para medir o grau de associação em uma tabela de contingência é o chamado *coeficiente de contingência*, definido com base na estatística χ^2 e do número n de elementos, da seguinte forma:[7]

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}}$$

Mas o valor máximo de C também depende da dimensão da tabela. Para facilitar a interpretação, usaremos uma modificação deste coeficiente. Chamaremos de k o menor valor entre ℓ (número de linhas da tabela) e c (número de colunas da tabela). Por exemplo, numa tabela de dimensão 2x2, temos $k = 2$. Numa tabela 3x5, temos $k = 3$. O chamado *coeficiente de contingência modificado* é dado por:

$$C^* = \sqrt{\frac{k \cdot \chi^2}{(k-1) \cdot (n + \chi^2)}}$$

O valor de C^* sempre estará no intervalo de 0 (zero) a 1 (um). Será 0 somente quando houver independência. Será 1 somente quando houver associação perfeita. Valores de C^* próximos de 1 descrevem uma *associação forte*, enquanto valores de C^* próximos de 0 indicam *associação fraca*. Os valores de C^* em torno de 0,5 podem ser interpretados como *associação moderada*.

EXEMPLO 12.4 (CONTINUAÇÃO) Temos na amostra A: $n = 300$, $k = 2$ e $\chi^2 = 0$. Então:

$$C^* = \sqrt{\frac{(2) \cdot (0)}{(2-1) \cdot (0 + 300)}} = 0 \rightarrow \textit{Independência!}$$

[7] Para calcular o coeficiente de contingência é conveniente calcular o χ^2 sem a correção de continuidade.

Temos na amostra B: $n = 300$, $k = 2$ e $\chi^2 = 300$. Então:

$$C^* = \sqrt{\frac{(2)\cdot(300)}{(2-1)\cdot(300+300)}} = 1 \rightarrow \textit{Associação perfeita!}$$

■

Exemplo 12.5 Vamos medir o grau de associação entre *hospital* e *recuperação funcional de pacientes*, descrito pelos dados da Tabela 12.4. Foram observados $n = 367$ pacientes, classificados numa tabela 3x5. Assim, $k = 3$ e, como vimos anteriormente, $\chi^2 = 56{,}7$. Então:

$$C^* = \sqrt{\frac{3\cdot(56{,}7)}{2\cdot(367+56{,}7)}} = 0{,}45$$

Logo, concluímos que a amostra descreve uma associação moderada entre *hospital* e *recuperação funcional de pacientes*.

Outros coeficientes de associação

O coeficiente de contingência é apenas uma opção dentre várias propostas de coeficientes de associação. Em tabelas 2x2, é usual o chamado coeficiente *phi*:

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

O coeficiente *phi* tem a vantagem de ser bastante simples e seu resultado sempre estará entre 0 e 1, permitindo interpretação similar ao coeficiente de contingência modificado. Mas é específico para tabelas 2x2. Uma generalização do coeficiente *phi* para tabelas de dimensão maiores é o chamado V de Cramér, definido por:

$$V = \sqrt{\frac{\chi^2}{n\cdot(k-1)}}$$

onde k é o menor valor entre ℓ (número de linhas da tabela) e c (número de colunas da tabela). Ver saída computacional apresentada na seção anterior.

Dados ordinais categorizados

Muitas vezes, as categorias de uma variável qualitativa formam uma ordenação (crescente ou decrescente). Isto ocorre, por exemplo, nos dois seguintes itens de um questionário (em ambos os itens as categorias estão numa ordem crescente):

a) Qual é o seu nível de instrução?
( ) nenhum
( ) fundamental
( ) médio
( ) superior

b) Qual é a sua opinião sobre o novo projeto educacional de seu município?
( ) totalmente contrário
( ) contrário
( ) indiferente ou sem opinião
( ) favorável
( ) completamente favorável

Ao estudarmos a associação entre duas variáveis ordinais, podemos não só ter interesse no grau de associação, mas também no seu sentido (*positiva* ou *negativa*). Preferimos, neste contexto, usar o termo *correlação* no lugar de *associação*. Dizemos que existe *correlação positiva* quando, na medida em que o nível de uma variável aumenta, cresce a chance de ocorrer níveis mais elevados na outra variável; *correlação negativa* ocorre quando, ao aumentar o nível de uma variável, diminui a chance de ocorrer níveis mais elevados na outra variável.

O coeficiente de correlação que apresentaremos aqui se baseia nos conceitos de *concordância* e *discordância*. Dizemos que dois indivíduos são concordantes se eles se posicionam em posições concordantes nas duas variáveis. São discordantes, se eles trocam de posição ao mudar de variável. Veja a seguinte situação:

João é alto e pesado;
Maria é baixa e leve

Podemos dizer que João e Maria formam um par *concordante*, pois, ao mudar de João para Maria, ambas as variáveis mudam para níveis inferiores (estatura: *alta* → *baixa*; peso: *pesado* → *leve*). E de Maria para João, ambas as variáveis mudam para níveis superiores (estatura: *baixa* → *alta*; peso: *leve* → *pesado*). Já na situação seguinte:

Pedro é baixo e pesado;
José é alto e leve

temos um par discordante, pois, ao passar do Pedro para o José, a estatura aumenta, enquanto que o peso diminui (estatura: *baixa* → *alta*; peso: *pesado* → *leve*).

Um conjunto de dados que tem, relativamente, muitos pares concordantes pode ser interpretado como tendo *correlação positiva*. Por

outro lado, um conjunto de dados que tem, relativamente, muitos pares discordantes, pode ser interpretado como tendo *correlação negativa.*

Vejamos, através de um exemplo, como contar o número n_c de pares concordantes e o número n_d de pares discordantes, num conjunto de observações de duas variáveis ordinais, apresentado numa tabela de contingência. O procedimento que apresentaremos vale para tabelas de qualquer dimensão, desde que as categorias das duas variáveis estejam dispostas numa mesma ordem (crescente ou decrescente).

Exemplo 12.6 Estudo da correlação entre *nível de instrução* e *posição com relação ao aborto* (Tabela 12.7).

Tabela 12.7 Classificação de 1.425 indivíduos, segundo o nível de instrução e a posição a respeito do aborto.

Nível de instrução	Posição com relação ao aborto		
	desaprova	indiferente	aprova
baixo	209	101	237
médio	151	126	426
alto	16	21	138

Fonte: Agresti (1984, p.157).

Como as categorias das duas variáveis da Tabela 12.7 já estão dispostas numa mesma ordem (ambas estão em ordem crescente), passamos a contar o número de concordâncias e o número de discordâncias, conforme o esquema a seguir:

Número de pares concordantes: n_c =

209	x	x
x	126	426
x	21	138

x	101	x
x	x	426
x	x	138

= 209·(126+426+21+138) + 101·(426+138) +

x	x	x
151	x	x
x	21	138

x	x	x
x	126	x
x	x	138

+ 151· (21+138) + 126· (138)

Portanto: n_c = 246.960

Número de pares discordantes: n_d =

x	x	237
151	126	x
16	21	x

x	101	x
151	x	x
16	x	x

= 237·(151+126+16+21) + 101·(151+16) +

x	x	x
x	x	426
16	21	x

x	x	x
x	126	x
16	x	x

+ 426· (16+21) + 126· (16)

n_d = 109.063

■

Coeficiente γ de Goodman e Kruskal

O coeficiente γ é definido por:

$$\gamma = \frac{n_c - n_d}{n_c + n_d}$$

O valor de γ estará sempre entre -1 e +1. Será +1 quando só houver concordâncias, e será -1 quando só houver discordâncias. Quando γ estiver em torno de zero, indica que o número de concordâncias e o número de discordâncias são aproximadamente iguais (ausência de correlação). Quanto mais próximo de +1 estiver γ, mais o número de concordâncias estará superando o número de discordâncias (correlação positiva forte). Simetricamente, quanto mais próximo de -1 estiver γ, mais o número de discordâncias estará superando o número de concordâncias (correlação negativa forte).

Exemplo 12.6 (continuação) Calculamos $n_c = 246.960$ e $n_d = 109.063$. Assim,

$$\gamma = \frac{246.960 - 109.063}{246.960 + 109.063} = 0{,}39$$

Concluímos, então, que a amostra apresenta uma correlação positiva moderada entre *nível de instrução* e *aceitação do aborto*. Ou seja, em termos dos indivíduos observados, existe uma leve tendência de quanto maior o *nível de instrução*, maior a *aceitação do aborto*.

■

Uso do computador

Considerando o anexo do Capítulo 4, buscou-se verificar uma possível associação entre o nível de instrução e a renda familiar. A Figura 12.4 mostra uma saída do pacote computacional SPSS®.

O resultado do teste qui-quadrado de Pearson ($\chi^2 = 16{,}28$, $gl = 4$ e $p = 0{,}003$) leva à rejeição de H_0, isto é, mostra haver associação entre *renda* e *nível de instrução*. O coeficiente γ, em torno de 0,5, indica uma *correlação positiva moderada* entre essas variáveis. Embora neste texto não comentamos a respeito de inferências sobre o coeficiente γ, podemos notar que a última tabela mostra o resultado de um teste estatístico (H_0: correlação nula na população e H_1: correlação não nula na população). Como o valor p é menor que um milésimo (última coluna), podemos concluir que o teste detecta a existência de correlação na população de onde foram extraídos os dados.

Classes de renda * Nível de instrução Crosstabulation

			Nível de instrução			Total
			nenhum	ensino fundamental	ensino médio	
Classes de renda	até 4,9 sal. mín.	Count	24	18	10	52
		% within Nível de instrução	64,9%	47,4%	22,7%	43,7%
	de 5 a 9,9 sal. mín.	Count	11	14	22	47
		% within Nível de instrução	29,7%	36,8%	50,0%	39,5%
	10 ou mais sal. mín.	Count	2	6	12	20
		% within Nível de instrução	5,4%	15,8%	27,3%	16,8%
Total		Count	37	38	44	119
		% within Nível de instrução	100,0%	100,0%	100,0%	100,0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	16,282(a)	4	0,003
Likelihood Ratio	17,302	4	0,002
Linear-by-Linear Association	15,473	1	0,000
N of Valid Cases	119		

[a] 0 cells (,0%) have expected count less than 5. The minimum expected count is 6,22.

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Gamma	0,495	0,101	4,556	0,000
N of Valid Cases		119			

[a] Not assuming the null hypothesis.
[b] Using the asymptotic standard error assuming the null hypothesis.

Figura 12.4 Saída do pacote computacional SPSS.

Cabe observar que houve um caso inválido (falta de resposta), ou seja, a análise foi realizada com 119 famílias e não com as 120 famílias amostradas.

Exercícios

5) Sejam os dados da Tabela 12.1, calcule e interprete:
   a) o coeficiente de contingência modificado;
   b) o coeficiente *phi*.

6) Para os dados da Tabela 12.2, calcule e interprete:
   a) o coeficiente de contingência modificado;
   b) o coeficiente *V* de Cramér.

7) Noventa crianças foram classificadas segundo suas habilidades em matemática e música, resultando nos seguintes dados.

Habilidade para música	Habilidade para matemática		
	alta	média	baixa
alta	20	10	7
média	12	10	8
baixa	6	7	10

Calcule o coeficiente γ e interprete.

8) Considere os dados do anexo do Capítulo 4.
   a) Calcule o coeficiente C^* para as variáveis *localidade da residência* e *uso de programas de alimentação popular*. Interprete.
   b) As localidades Monte Verde, Parque da Figueira e Encosta do Morro estão em ordem decrescente, em termos da qualidade das construções habitacionais. Usando esta informação, calcule o coeficiente γ entre *localidade da residência* e *uso de programas de alimentação popular*. Interprete.

9) Considerando os dados do anexo do Capítulo 2, calcule o coeficiente γ entre *satisfação com a didática dos professores* e *satisfação geral com o curso*. Interprete.

EXERCÍCIOS COMPLEMENTARES

10) A tabela que segue apresenta uma classificação de pessoas segundo o nível de instrução e colaboração com a coleta seletiva de lixo.[8] Verifique se existe associação significativa entre estas duas variáveis.

Nível de instrução	Colabora com a coleta seletiva de lixo	
	sim	não
nenhum ou fundamental	22	13
médio	33	34
superior	39	36

11) Os dados abaixo se referem ao tipo de escola em que o aluno realizou o ensino médio (0 = *pública* e 1 = *particular*) e o resultado no vestibular (0 = *não passou* e 1 = *passou*).

aluno	escola	vestib.	aluno	escola	vestib.	aluno	escola	vestib.
1	1	1	11	0	0	21	1	0
2	1	1	12	0	1	22	0	0
3	1	0	13	0	0	23	0	0
4	0	0	14	0	1	24	0	0
5	0	1	15	1	1	25	1	0
6	1	1	16	1	0	26	0	0
7	0	0	17	0	0	27	0	0
8	1	1	18	1	1	28	1	1
9	1	0	19	0	0	29	0	1
10	0	0	20	0	0	30	1	1

[8] É parte de uma pesquisa realizada em Florianópolis – SC, em 1999, pelos acadêmicos João Fáveri e Ângela Queiroz, do Curso de Psicologia da UFSC, semestre 1999/1.

Construa uma distribuição de frequências conjunta para as variáveis *tipo de escola* e *resultado no vestibular*. Apresente essa distribuição numa tabela de dupla entrada. Os dados sugerem associação? Explique através de um teste estatístico apropriado com $\alpha = 0{,}10$.

12) Para verificar se em estudantes universitários existe associação entre três áreas de estudo (*humanas*, *biológicas* ou *exatas*) e a aprovação em relação ao exame de final de curso proposto pelo governo (*favorável* ou *contrário*), foram observados 120 estudantes aleatoriamente. Dos 40 estudantes da área de humanas, 10 disseram ser favoráveis (e os restantes, contrários). Dos 30 estudantes da área biológica, 10 eram favoráveis (e os restantes, contrários). E dos 50 da área de exatas, 20 eram favoráveis (e os restantes, contrários). Pode-se dizer que existe associação entre essas duas variáveis? Faça um teste estatístico apropriado ao nível de significância de 5%.

13) Considere que você tenha um conjunto de dados de seus clientes, contendo as seguintes características:
   - Sexo (*masculino* ou *feminino*);
   - Local da residência (*na própria cidade* ou *em outra cidade*);
   - Nível de satisfação (escala de 0 a 10) e
   - Valor mensal das compras (média dos últimos 3 meses, em R$).

   Que técnicas estatísticas você usaria para:

   a) verificar se existe relação entre *sexo* e *local da residência* do cliente;
   b) verificar se o valor das compras tende a ser diferente para homens e mulheres;
   c) verificar se há relação do *nível de satisfação* com o *local de residência* do cliente.

Capítulo 13

Correlação e regressão

Neste capítulo, vamos dar sequência ao estudo de associação entre duas variáveis, mas agora, supondo que ambas sejam mensuradas *quantitativamente*. Usaremos, neste caso, o termo *correlação* no lugar de *associação*.

Variáveis correlacionadas

Dizemos que duas variáveis, X e Y, são *positivamente correlacionadas* quando elas *caminham num mesmo sentido*, ou seja, elementos com valores pequenos de X tendem a ter valores pequenos de Y, e elementos com valores grandes de X tendem a ter valores grandes de Y. São *negativamente correlacionadas* quando elas *caminham em sentidos opostos*, ou seja, elementos com valores pequenos de X tendem a ter valores grandes de Y, e elementos com valores grandes de X tendem a ter valores pequenos de Y.

As variáveis *peso* e *altura* apresentam-se, em geral, *correlacionadas positivamente*, pois os indivíduos altos tendem a ser mais pesados, enquanto a maioria dos indivíduos baixos é leve. Por outro lado, no Brasil, as variáveis *renda familiar* e *número de elementos da família* costumam apresentar-se *correlacionadas negativamente*, pois as famílias de baixa renda, em geral, tendem a ter mais filhos do que as de alta renda.

Ilustraremos o estudo de correlações entre duas variáveis, usando os dados da Tabela 13.1, relativos a alguns indicadores sociais de uma amostra de municípios brasileiros.[1]

[1] Estamos usando uma amostra bastante pequena para ilustrar as técnicas. Um estudo dessas variáveis pode ser feito com toda a população de municípios, já que esses dados estão disponíveis no Censo Demográfico de 2000 ou no Atlas de Desenvolvimento Humano, (www.pnud.org.br/atlas).

Tabela 13.1 Alguns dados, baseados no Censo Demográfico de 2000, de uma amostra aleatória de municípios brasileiros

Município	DistCap	EspVida	MortInf	Alfab	Renda
Araruna (PR)	365	67,99	23,19	86,23	188,29
Nova Redenção (BA)	278	61,19	56,56	63,00	74,79
Monção (MA)	150	59,58	63,32	63,64	66,96
Porto Rico do Maranhão (MA)	78	58,96	66,05	79,33	65,34
Campo Erê (SC)	468	68,10	31,71	83,38	173,38
Lagoa do Piauí (PI)	40	63,65	47,08	65,81	60,00
São José das Palmeiras (PR)	486	71,01	16,62	77,54	150,67
Paraíba do Sul (RJ)	83	71,36	15,69	89,28	264,55
Malhada dos Bois (SE)	65	64,46	44,18	69,95	80,69
Jandaíra (BA)	175	62,45	51,57	59,72	58,68
Vespasiano (MG)	14	68,68	32,81	90,43	196,51
Ipaba (MG)	167	67,42	37,04	81,82	125,75

Fonte: Atlas de Desenvolvimento Humano (www.pnud.org.br/atlas).
Descrição das variáveis:
DistCap: distância à capital da respectiva Unidade da Federação.
EspVida: esperança de vida ao nascer
MortInf: mortalidade (número médio de mortes em 1.000) até um ano de idade.
Alfab: taxa de alfabetização (percentagem da população adulta alfabetizada).
Renda: renda *per capita* do município (R$).

13.1 Diagramas de dispersão

Uma maneira de visualizarmos se duas variáveis apresentam-se correlacionadas é através do *diagrama de dispersão*, no qual os valores das variáveis são representados por pontos, num sistema cartesiano. Esta representação é feita sob forma de pares ordenados (x, y), onde x é um valor de uma variável e y é o correspondente valor da outra variável. A Figura 13.1 ilustra a construção de um diagrama de dispersão.

x	y
365	67,99
278	61,19
150	59,58

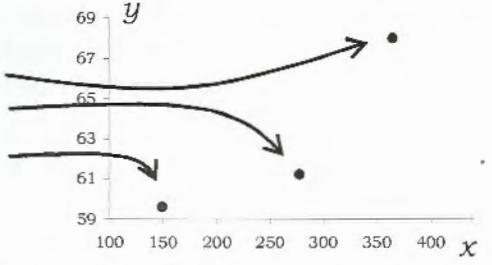


Figura 13.1 Construção de um diagrama de dispersão. Representação das três primeiras observações de X = *distância da capital* e Y = *esperança de vida ao nascer*, referente aos dados da Tabela 13.1

A Figura 13.2 mostra quatro diagramas de dispersão, relativos aos *cruzamentos* de algumas variáveis da Tabela 13.1. O leitor deve notar que cada par de observações refere-se ao mesmo elemento (município), ou seja, a análise baseia-se em *dados pareados*.

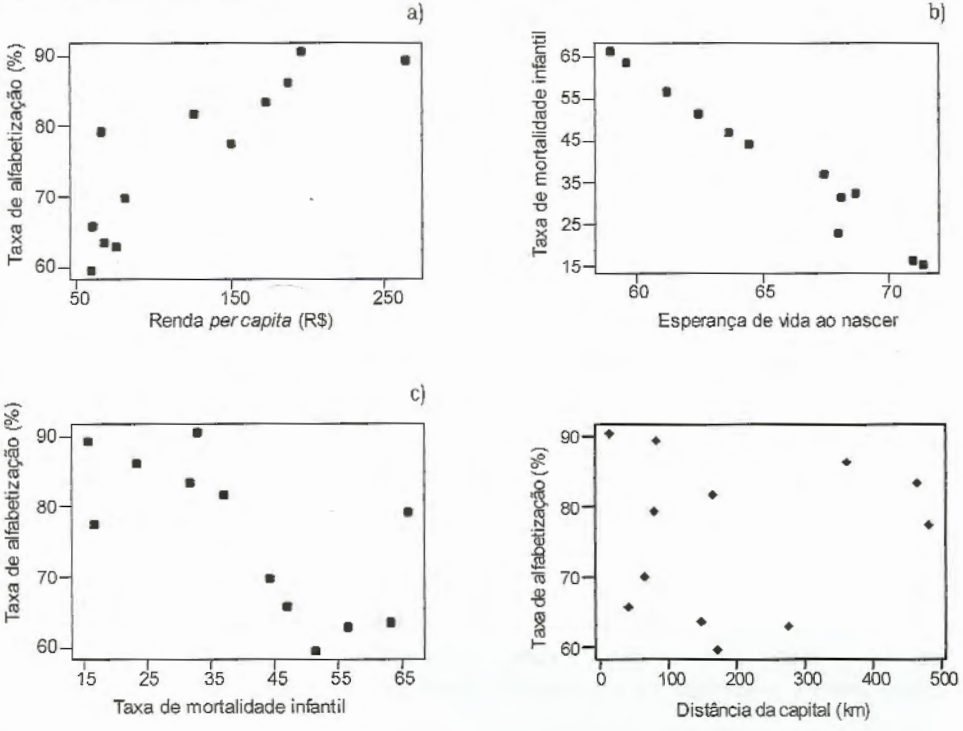


Figura 13.2 Alguns diagramas de dispersão, construídos com os dados da Tabela 13.1

O diagrama (a) da Figura 13.1 mostra uma situação de *correlação positiva*, porque os pontos estão em torno de uma linha imaginária *ascendente*. Em geral, valores pequenos de uma variável são também pequenos na outra, o mesmo acontecendo para valores grandes.

Os diagramas (b) e (c) mostram correlações negativas, porque, em ambos os casos, os pontos estão em torno de uma linha imaginária descendente. Valores pequenos de uma variável são, em geral, grandes na outra. Em (b) os pontos apresentam-se *mais próximos* de uma linha descendente do que em (c), o que caracteriza uma correlação *mais forte*.

Os dados de *distância da capital* e *taxa de alfabetização*, diagrama (d), não se apresentam correlacionados, pois valores pequenos (ou grandes) de uma variável estão associados tanto a valores pequenos

quanto a valores grandes da outra. Os pontos não se posicionam em torno de alguma linha ascendente ou descendente.

A Figura 13.3 mostra um conjunto de pontos aproximando-se mais de uma parábola do que de uma reta, ilustrando um caso de *correlação não linear*. As correlações não lineares são mais difíceis de serem interpretadas e não serão abordadas neste livro.

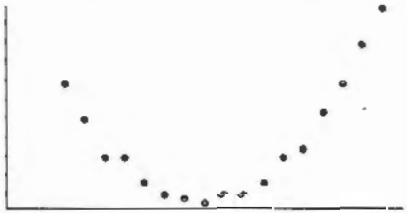


Figura 13.3 Diagrama de dispersão de um exemplo hipotético de correlação não linear

É importante ressaltar que o conceito de *correlação* refere-se a uma associação numérica entre duas variáveis, não implicando, necessariamente, uma relação de *causa e efeito*, ou mesmo numa estrutura com interesses práticos. Se observarmos, por exemplo, as variáveis *população da Argentina* e *venda de cerveja no Brasil* ao longo dos últimos anos, elas devem se apresentar correlacionadas positivamente, pois ambas estão aumentando com o tempo. Contudo, em termos práticos, esta correlação é espúria, não trazendo qualquer informação relevante.

A análise de dados para verificar correlações é usualmente feita em termos exploratórios, em que a verificação de uma correlação serve como um elemento auxiliar na análise do problema em estudo. Ou seja, o estudo da correlação numérica entre as observações de duas variáveis é geralmente um passo intermediário na análise de um problema.

13.2 O COEFICIENTE DE CORRELAÇÃO LINEAR DE PEARSON

No capítulo anterior, estudamos o *coeficiente de contingência*, que descreve, através de um único número, o grau de associação dos dados de duas variáveis categorizadas. Nesta seção, apresentaremos o chamado *coeficiente de correlação (linear) de Pearson*, apropriado para descrever a correlação linear dos dados de duas variáveis *quantitativas*.

Valores padronizados e o coeficiente r

O valor do coeficiente de correlação não deve depender da unidade de medida dos dados. Por exemplo, o coeficiente de correlação entre as variáveis *peso* e *altura* deve acusar o mesmo valor, independentemente se o peso for medido em *gramas* ou *quilogramas*, e a altura em *metros* ou *centímetros*.

Para evitar o efeito da unidade de medida, os dados devem ser padronizados da seguinte forma:

$$x' = \frac{x - \overline{X}}{S_x} \qquad y' = \frac{y - \overline{Y}}{S_y}$$

onde:

x': um valor padronizado;
x: um valor da variável X;
$\overline{X}$: média dos dados da variável X;
S_x: desvio padrão dos dados de X;

y': um valor padronizado;
y: um valor da variável Y;
$\overline{Y}$: média dos dados da variável Y e
S_y: desvio padrão dos dados de Y.

O coeficiente de correlação linear de Pearson, r, é definido pela seguinte expressão em termos dos valores padronizados:

$$r = \frac{\sum(x' \cdot y')}{n - 1}$$

onde:

n é o tamanho da amostra, isto é, o número de pares (x, y) e
$\sum(x' \cdot y')$ é a soma dos produtos $x' \cdot y'$ dos pares de valores padronizados, isto é, para cada par (x', y'), fazemos o produto $x' \cdot y'$ e, depois, somamos os resultados desses produtos.

Os exemplos 13.1 e 13.2 procuram justificar como a expressão de r mede a correlação. O primeiro ilustra uma situação de correlação positiva, e o segundo um caso de correlação negativa.

Exemplo 13.1 Cálculo dos valores padronizados e do coeficiente de correlação de Pearson de um conjunto de dados hipotéticos com correlação positiva (ver Tabela 13.2).

Tabela 13.2 Cálculos intermediários para se obter r (Exemplo 13.1).

	Valores originais		Valores padronizados		Produtos
	X	Y	X'	Y'	$X'\cdot Y'$
	2	4	-1,50	-1,75	2,63
	3	7	-1,00	-0,88	0,88
	4	9	-0,50	-0,29	0,15
	5	10	0,00	0,00	0,00
	5	11	0,00	0,29	0,00
	6	11	0,50	0,29	0,15
	7	13	1,00	0,88	0,88
	8	15	1,50	1,46	2,19
Soma:	40	80	0,00	0,00	6,87
Média:	5,00	10,00	0,00	0,00	
Desvio padrão:	2,00	3,42	1,00	1,00	

Observe que calculamos a média e o desvio padrão dos valores das variáveis X e Y. De cada valor, diminuímos a média e dividimos pelo desvio padrão. Por exemplo, para o primeiro valor de X, $x = 2$, calculamos o valor padronizado $x' = (x - 5)/2 = (2 - 5)/2 = -1{,}5$. Veja a mudança de escala com a padronização na Figura 13.4.

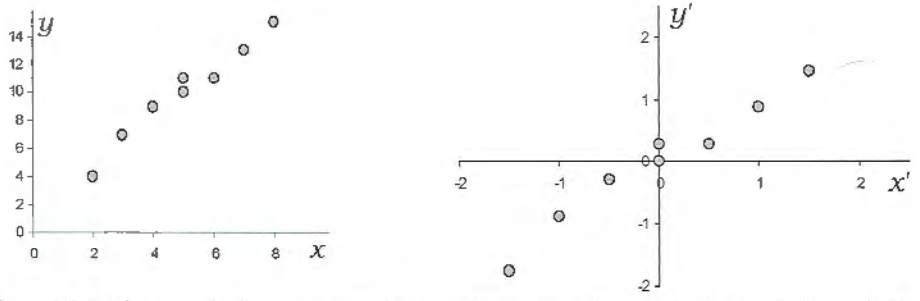


Figura 13.4 Diagrama de dispersão dos valores originais e dos valores padronizados do Exemplo 13.1

Quando estamos trabalhando com dados correlacionados positivamente, como no exemplo precedente, os pares (x', y') tendem a ter o mesmo sinal (+ ou -), especialmente para aqueles pontos longe da origem. Assim, a maioria dos produtos $x'\cdot y'$ resulta em valores positivos (ver Figura 13.4). Em consequência, o coeficiente r será positivo. Concluindo os cálculos da Tabela 13.2, temos:

$$r = \frac{\sum (x' \cdot y')}{n - 1} = \frac{6{,}87}{7} = 0{,}981$$

■

Exemplo 13.2 Cálculo dos valores padronizados e do coeficiente de correlação de Pearson de um conjunto de dados hipotéticos com correlação negativa (Tabela 13.3).

Tabela 13.3 Cálculos intermediários para se obter r (Exemplo 13.2).

	Valores originais		Valores padronizados		Produtos
	X	Y	X'	Y'	$X'\cdot Y'$
	2	16	-1,50	1,75	-2,63
	3	13	-1,00	0,88	-0,88
	4	11	-0,50	0,29	-0,15
	5	10	0,00	0,00	0,00
	5	9	0,00	-0,29	0,00
	6	9	0,50	-0,29	-0,15
	7	7	1,00	-0,88	-0,88
	8	5	1,50	-1,46	-2,19
Soma:	40	80	0,00	0,00	-6,87
Média:	5,00	10,00	0,00	0,00	
Desvio padrão:	2,00	3,42	1,00	1,00	

Complementando os cálculos da Tabela 13.3, temos o coeficiente:

$$r = \frac{\sum(x' \cdot y')}{n-1} = \frac{-6,87}{7} = -0,981$$

Neste exemplo o coeficiente é negativo, porque os pares (x', y') tiveram, em geral, sinais trocados, especialmente para aqueles pontos longe da origem (veja Figura 13.5). Isto tende a levar os produtos $x' \cdot y'$ a resultarem em valores negativos e, em consequência, gerar um coeficiente r negativo. A Figura 13.5 ilustra esta situação. Verificamos maior concentração de pontos nos quadrantes II e IV (onde x' e y' têm sinais trocados), acarretando num valor negativo para r.

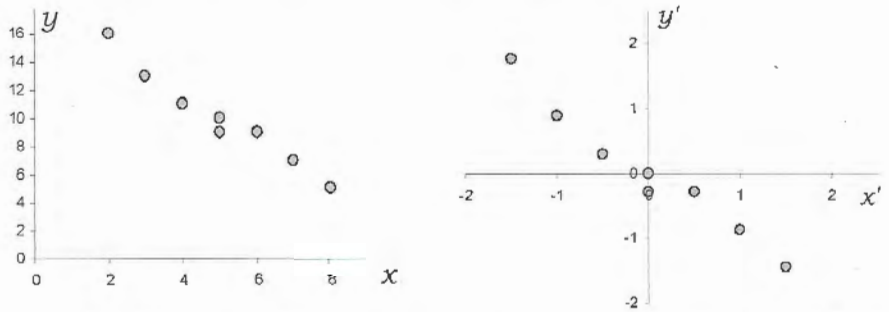


Figura 13.5 Diagrama de dispersão dos valores originais e dos valores padronizados do Exemplo 13.2.

Dos exemplos 13.1 e 13.2, verificamos que o sinal da soma dos produtos dos valores padronizados, $\sum(x' \cdot y')$, fará com que o coeficiente r tenha sinal compatível com o que vimos nos diagramas de dispersão (veja também a Figura 13.6). Para dados correlacionados positivamente, os pontos se concentrarão nos quadrantes I e III, com x' e y' de mesmo sinal (produtos positivos). Para dados correlacionados negativamente, os pontos ficarão nos quadrantes II e IV, fazendo com que x' e y' tenham sinais trocados (produtos negativos). Se os dados forem não correlacionados, os pontos se espalharão de forma aproximadamente igual em todos os quadrantes, fazendo com que tenhamos produtos positivos e negativos, acarretando numa soma próxima de zero.

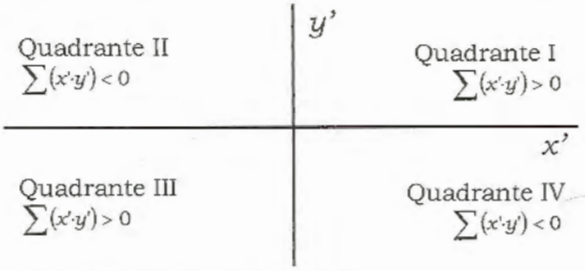


Figura 13.6 Justificativa do sinal de r.

Para qualquer conjunto de dados, o valor do coeficiente de correlação de Pearson, r, estará no intervalo de -1 a 1. Será *tão mais próximo* de 1 (ou -1) quanto mais *forte* for a correlação nos dados observados. Teremos $r = +1$ se os pontos estiverem exatamente sobre uma reta ascendente (*correlação positiva perfeita*). Por outro lado, teremos $r = -1$ se os pontos estiverem exatamente sobre uma reta descendente (*correlação negativa perfeita*). Quando não houver correlação nos dados, r acusará um valor próximo de 0 (zero). Veja a Figura 13.7.

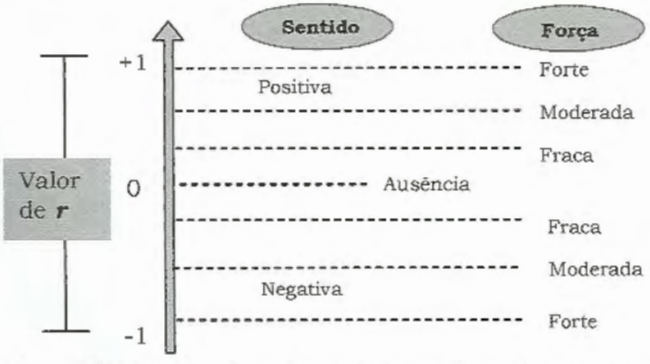


Figura 13.7 Sentido e força da correlação em função do valor de r.

Cálculo de r

O cálculo de r pela expressão apresentada no tópico anterior tem o inconveniente de incorporar erros de arredondamentos, pois normalmente os valores da média e desvio padrão não são inteiros. Neste contexto, sugerimos usar a seguinte fórmula alternativa, a qual é baseada nas observações originais:

$$r = \frac{n \cdot \Sigma(X \cdot Y) - (\Sigma X) \cdot (\Sigma Y)}{\sqrt{n \cdot \Sigma X^2 - (\Sigma X)^2}\ \sqrt{n \cdot \Sigma Y^2 - (\Sigma Y)^2}}$$

Para obter os somatórios, procedemos da seguinte maneira.

$\Sigma(X \cdot Y)$: fazemos os produtos $x \cdot y$, referentes a cada par de observações, e, depois, efetuamos a soma;
ΣX: somamos os valores da variável X;
ΣY: somamos os valores da variável Y;
ΣX^2: elevamos ao quadrado cada valor de X e, depois, efetuamos a soma; e
ΣY^2: elevamos ao quadrado cada valor de Y e, depois, efetuamos a soma.

Para ilustrar o uso da última expressão, vamos refazer o Exemplo 13.1. A Tabela 13.4 apresenta alguns cálculos intermediários.

Tabela 13.4 Cálculos intermediários para a obtenção de r

	Valores originais		Cálculos intermediários		
	X	Y	X^2	Y^2	$X \cdot Y$
	2	4	4	16	8
	3	7	9	49	21
	4	9	16	81	36
	5	10	25	100	50
	5	11	25	121	55
	6	11	36	121	66
	7	13	49	169	91
	8	15	64	225	120
Soma:	40	80	228	882	447

Sendo

$$r = \frac{n \cdot \Sigma(X \cdot Y) - (\Sigma X) \cdot (\Sigma Y)}{\sqrt{n \cdot \Sigma X^2 - (\Sigma X)^2}\ \sqrt{n \cdot \Sigma Y^2 - (\Sigma Y)^2}}$$

temos,

$$r = \frac{8 \cdot (447) - 40 \cdot (80)}{\sqrt{8 \cdot (228) - (40)^2}\ \sqrt{8 \cdot (882) - (80)^2}} =$$

$$= \frac{3.576 - 3.200}{\sqrt{1824 - 1.600}\ \sqrt{7.056 - 6.400}} =$$

$$= \frac{376}{\sqrt{224}\ \sqrt{656}} = \frac{376}{383,33} = 0,981$$

Encontramos o mesmo resultado obtido no tópico anterior, o que era de se esperar, pois as fórmulas são matematicamente equivalentes.

TESTE DE SIGNIFICÂNCIA SOBRE R

Quando os dados são provenientes de uma população, além de mensurar o grau de correlação observado nos dados, muitas vezes temos interesse em testar a existência de correlação entre duas variáveis, X e Y, na população. Isso é feito com base em uma amostra de observações pareadas (x, y). As hipóteses são:

H_0: as variáveis X e Y são *não correlacionadas*;
H_1: as variáveis X e Y são *correlacionadas*;

podendo, ainda, a hipótese alternativa indicar o sentido da correlação (teste unilateral), tal como,

H_1': X e Y são correlacionadas *positivamente* ou
H_1'': X e Y são correlacionadas *negativamente*.

O teste unilateral é aplicado nos casos em que já se espera que o coeficiente de correlação tenha determinado sinal (+ ou -).

Restringiremo-nos à verificação de correlação *linear* e vamos supor que os dados de X e de Y provenham de distribuições *normais*.[2] Podemos realizar o teste com auxílio da Tabela 7 do apêndice, que apresenta o valor absoluto mínimo de r para ser significativo (rejeitar H_0), para cada n.

EXEMPLO 13.3 Com o objetivo de verificar se existe correlação positiva entre *aptidão em matemática* e *aptidão em música*, foi selecionado um grupo de crianças de 8 a 10 anos de idade, que foram submetidas a dois testes de aptidão: um de matemática e outro de música. A ordem da aplicação dos testes em cada criança foi aleatória.

[2] Para se verificarem as suposições do teste de correlação, sugerimos construir: (1) um diagrama de pontos para os dados de cada variável para verificar se não existe forte evidência de desvio da distribuição normal; e (2) um diagrama de dispersão para verificar se os dados sugerem uma relação *não linear*.

Temos, então, as seguintes hipóteses, relativas às crianças da faixa etária de 8 a 10 anos, similares ao grupo de crianças que participaram do estudo:

H_0: não existe correlação entre *aptidão em matemática* e *aptidão em música.*

H_1: a *aptidão em matemática* e a *aptidão em música são* correlacionadas positivamente.[3]

Os resultados dos testes de aptidão foram os seguintes:

	Valores de aptidão em			Valores de aptidão em	
Criança	matemática	música	Criança	matemática	música
1	60	80	7	48	79
2	58	62	8	72	88
3	73	70	9	75	54
4	51	83	10	83	82
5	54	62	11	62	64
6	75	92	12	52	69

Efetuando-se o cálculo do coeficiente de correlação de Pearson, conforme visto anteriormente, temos: $r = 0,17$. Pela Tabela 7 do apêndice, verificamos que, ao nível de significância usual de 5%, o valor mínimo de r para a correlação ser significativa é de 0,497 (teste unilateral). Como o valor encontrado ($r = 0,17$) é menor que o valor tabelado (0,497), o teste aceita H_0. Em outras palavras, a correlação positiva fraca ($r = 0,17$), descrita pelos dados da amostra, não é suficiente para afirmarmos a existência de correlação positiva entre as duas variáveis, na população em estudo.

A Tabela 7 também pode ser usada para se ter uma avaliação da probabilidade de significância (valor p). No exemplo em questão, podemos verificar que o valor encontrado ($r = 0,17$) é inferior a todos os valores tabelados para $n = 12$, ou seja, $p > 0,10$ (teste unilateral). Assim, mesmo que estivéssemos fazendo o teste ao nível de significância de $\alpha = 10\%$, o teste ainda aceitaria H_0.

Uso do computador

A maioria dos pacotes computacionais de Estatística apresenta os resultados de uma análise de correlações em forma matricial. Na primeira linha e primeira coluna, são apresentadas as variáveis. Em cada cruza-

[3] Observe que o problema sugere um teste unilateral (hipótese alternativa afirmando *correlação positiva* e não somente *existência de correlação*). Cabe observar que as hipóteses estatísticas levam em conta o instrumento de mensuração das variáveis, isto é, supõe-se que os testes de aptidão estejam realmente medindo aquilo que se propõem.

mento, o coeficiente de correlação r do correspondente par de variáveis. Alguns pacotes apresentam também o número n de pares usado no cálculo de r e o valor p do teste bilateral sobre o correspondente coeficiente de correlação populacional. A Tabela 13.5 mostra uma saída computacional do SPSS® relativa aos dados da Tabela 13.1. Vemos, por exemplo, que o coeficiente de correlação entre DISTCAP e ESPVIDA é 0,337 (positiva fraca). Observando o correspondente valor p = 0,284, verificamos que não se pode dizer que existe correlação entre essas duas variáveis na população de municípios brasileiros.

Tabela 13.5 Saída computacional de uma análise de correlação pelo SPSS

		DISTCAP	ESPVIDA	MORTINF	ALF	RENDA
DISTCAP	Pearson Correlation	1	0,337	-0,400	0,087	0,205
	Sig. (2-tailed)	.	0,284	0,198	0,788	0,523
	N	12	12	12	12	12
ESPVIDA	Pearson Correlation	0,337	1	-0,983(**)	0,718(**)	0,865(**)
	Sig. (2-tailed)	0,284	.	0,000	0,009	0,000
	N	12	12	12	12	12
MORTINF	Pearson Correlation	-0,400	-0,983(**)	1	-0,684(*)	-0,860(**)
	Sig. (2-tailed)	0,198	0,000	.	0,014	0,000
	N	12	12	12	12	12
ALF	Pearson Correlation	0,087	0,718(**)	-0,684(*)	1	0,863(**)
	Sig. (2-tailed)	0,788	0,009	0,014	.	0,000
	N	12	12	12	12	12
RENDA	Pearson Correlation	0,205	0,865(**)	-0,860(**)	0,863(**)	1
	Sig. (2-tailed)	0,523	0,000	0,000	0,000	.
	N	12	12	12	12	12

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

Variáveis indicadoras

Algumas vezes, temos variáveis codificadas com 1 e 0, indicando a *presença* ou *ausência* de algum atributo, respectivamente. Por exemplo, a variável X pode ser indicadora de indivíduos fumantes (X = 1 para *fumantes* e X = 0 para *não fumantes*) e a variável Y indicadora de uma certa doença (Y = 1 para indivíduos *doentes* e Y = 0 para *sadios*). O cálculo

de r entre duas variáveis deste tipo pode indicar correlação positiva (fumante tem maior chance de ter a doença) ou negativa (fumante tem menor chance de ter a doença), além do grau (forte, moderada, fraca ou ausência). Mas não faz sentido o teste de significância discutido anteriormente, porque ele só é válido para variáveis com distribuição aproximadamente normal.

Quando 0 e 1 representam apenas rótulos de uma variável (por exemplo, 0 para feminino e 1 para masculino), é melhor considerar o coeficiente r sem sinal, indicando apenas o grau de *associação* descrita pelos dados. O coeficiente r para variáveis 0-1, em valor absoluto, é o coeficiente de associação *phi*, definido no capítulo anterior. Daí, para verificar sua significância, realizamos um teste qui-quadrado.

13.3 Correlação por postos

Quando os dados de alguma das variáveis em estudo mostram-se com distribuição muito assimétrica ou com valores discrepantes, a análise da correlação através do coeficiente r pode ficar comprometida. Uma alternativa é aplicar a abordagem não paramétrica do coeficiente de correlação r_s *de Spearman*, o qual se utiliza apenas da ordenação dos valores.

Sejam os dados da Tabela 13.6, relativos a um estudo correlacional entre *aptidão em matemática* e *aptidão em música*. Para facilitar, os valores de *aptidão em matemática* já estão ordenados em ordem crescente. Para cada variável, são atribuídos postos (*ranks*) da seguinte maneira: ao menor valor é atribuído o posto 1; ao segundo menor, posto 2; e assim por diante. Quando ocorre algum empate (repetição de algum valor), consideramos que isto tenha acontecido por deficiência do instrumento de medida e atribuímos postos sequenciais, mas, em seguida, calculamos a média dos postos dos valores empatados. Por exemplo, na variável *aptidão em matemática*, temos as crianças 6 e 9 com valores empatados em 75. Preliminarmente, uma recebe posto 10 e a outra, posto 11; depois, alocamos posto 10,5 (média entre 10 e 11) para ambas.

Tabela 13.6 Alocação de postos para o cálculo de r_s *de Spearman.*

Criança	Aptidão em matemática (X)	Posto de X	Aptidão em música (Y)	Posto de Y	D	D^2
7	48	1	79	7	-6	36
4	51	2	83	10	-8	64
12	52	3	69	5	-2	4
5	54	4	62	2,5[2]	1,5	2,25
2	58	5	62	2,5[2]	2,5	6,25
1	60	6	80	8	-2	4
11	62	7	64	4	3	9
8	72	8	88	11	-3	9
3	73	9	70	6	3	9
6	75	10,5[1]	92	12	-1,5	2,25
9	75	10,5[1]	54	1	9,5	90,25
10	83	12	82	9	3	9
					Soma	245

Notas: [1] Média dos postos 10 e 11, referentes ao valor empatado 75.
[2] Média dos postos 2 e 3, referentes ao valor empatado 62.

A sexta coluna da Tabela 13.6 apresenta as diferenças entre postos:

$$D = \text{Posto de } X - \text{Posto de } Y$$

Na última coluna temos as diferenças quadráticas entre postos, cuja soma denotamos por $\sum D^2$. O coeficiente de correlação de Spearman é definido por:[4]

$$r_s = 1 - \frac{6 \cdot \sum D^2}{n \cdot (n^2 - 1)}$$

Com os dados da Tabela 13.6, temos: $\sum D^2 = 245$. E o coeficiente r_s *de Spearman*:

$$r_s = 1 - \frac{6 \cdot \sum D^2}{n \cdot (n^2 - 1)} = 1 - \frac{6 \cdot (245)}{12 \cdot (12^2 - 1)} = 1 - 0{,}86 = 0{,}14$$

indicando uma correlação positiva muito fraca nos dados observados.[5]

Em função do tamanho n da amostra e do nível de significância α adotado, a Tabela 8 do apêndice apresenta os valores absolutos mínimos de r_s para que esse seja significativo. Em termos do exemplo em questão,

[4] O coeficiente r_s é o próprio coeficiente de correlação de Pearson, r, calculado sobre os postos de X e Y.

[5] Assim como o *r de Pearson*, o r_s *de Spearman* varia entre –1 e +1, com a mesma interpretação. Porém, os resultados de r e r_s não são matematicamente iguais por usarem metodologias diferentes de cálculo.

para $n = 12$ e nível de significância de 5%, temos o valor mínimo tabelado de 0,503 (teste unilateral). Como o valor encontrado ($r_s = 0,14$) é menor que o valor tabelado, o teste não acusa significância. Não é possível dizer que existe correlação positiva entre *aptidão em matemática* e *aptidão em música*, na população de onde os dados foram extraídos.

Exercícios

1) Considerando os dados da Tabela 13.1, construir um diagrama de dispersão para as variáveis *renda per capita* e *esperança de vida ao nascer*. Quais as informações observadas no gráfico?

2) Sejam X = *nota na prova do vestibular de matemática* e Y = *nota final na disciplina de cálculo*. Estas variáveis foram observadas em 20 alunos, ao final do primeiro período letivo de um curso de engenharia. Os dados são apresentados a seguir.

X	Y	X	Y	X	Y	X	Y	X	Y
39	65	43	78	21	52	64	82	65	88
57	92	47	89	28	73	75	98	47	71
34	56	52	75	35	50	30	50	28	52
40	70	70	50	80	90	32	58	67	88

   a) Construa um diagrama de dispersão e verifique se existe correlação entre os dados dessas duas variáveis.
   b) Existe algum aluno que *foge* ao comportamento geral dos demais (ponto discrepante)?
   c) Calcule o coeficiente r.
   d) Retire o valor discrepante detectado no item (b) e calcule novamente o coeficiente r. Verifique se é significativo ao nível de significância de 5%. Interprete.
   e) Calcule o coeficiente r_s com todos os valores e verifique se é significativo ao nível de significância de 5%.

3) Sejam os dados do anexo do Capítulo 2. Faça um diagrama de dispersão com os dados das variáveis: X = *satisfação do aluno com o curso* e Y = *desempenho do aluno*. Interprete.

4) Sejam os dados do anexo do Capítulo 4. Considerando apenas a Encosta do Morro, faça um diagrama de dispersão com os dados de: X = *renda familiar* e Y = *número de moradores no domicílio*. Interprete.

5) Faça o cálculo do coeficiente r com os dados do Exemplo 13.3 e confira o resultado encontrado.

6) Considerando as variáveis *taxa de alfabetização* e *taxa de mortalidade infantil*, (Tabela 13.1), calcule:
   a) o coeficiente de correlação de Pearson. Interprete o resultado obtido.
   b) o coeficiente de correlação de Spearman e verifique se é significativo ao nível de significância de 5%.

7) Com respeito aos 23 alunos de uma turma de estatística, foram observadas as variáveis: *número de faltas* e *nota final na disciplina*. Esses dados levaram à seguinte correlação, descrita pelo coeficiente de correlação de Pearson: $r = -0,56$. Comente as seguintes frases relativas à turma em estudo e ao coeficiente obtido.
   a) Como $r = -0,56$ (correlação negativa moderada), nenhum aluno com grande número de faltas tirou nota alta.
   b) Como as duas variáveis são correlacionadas, bastaria usar uma delas como critério de avaliação, pois uma acarreta a outra.
   c) Os dados mostraram uma leve tendência de que a nota final se relaciona inversamente com o número de faltas; então, os alunos *frequentadores* tiveram, em geral, melhores desempenhos nas avaliações do que os alunos que faltaram muito.

8) Numa amostra aleatória de $n = 212$ livros da Biblioteca Central da UFSC, encontramos $r = 0,207$ para as variáveis: *idade da edição* e *número de páginas do livro*.
   a) O que se pode dizer com base no valor deste coeficiente de correlação?
   b) Esta correlação pode ser explicada meramente por fatores casuais? Faça um teste estatístico apropriado, ao nível de significância de 5%.

13.4 REGRESSÃO LINEAR SIMPLES

O termo *regressão* surgiu com os trabalhos de Galton no final do século XIX. Esses trabalhos procuravam explicar certas características de um indivíduo a partir das características de seus pais. Galton acreditava que os filhos de pais excepcionais, com respeito à determinada característica, também possuíam essa característica, mas, em geral, numa intensidade menor do que a média de seus pais. Seus estudos baseavam-se em observações empíricas. Em um desses trabalhos ele relacionou centenas de alturas de indivíduos com as respectivas alturas médias de seus pais.

EXEMPLO 13.4 Vamos considerar uma parte dos dados coletados por Galton, por volta de 1885 (Tabela 13.7).

Tabela 13.7 Alturas de indivíduos (Y) e alturas médias de seus pais (X), medidas em centímetros.

X	Y	X	Y	X	Y	X	Y
164	166	164	168	166	166	166	168
166	171	166	173	169	166	169	168
169	171	169	173	171	166	171	168
171	171	171	173	171	176	173	168
173	171	173	176	173	178	176	171
176	173	176	176	178	176	178	178

Fonte: Stigler (1986, p. 286), com adaptações.

A Figura 13.8 representa as observações da Tabela 13.7 num diagrama de dispersão, indicando uma correlação positiva, como era de se esperar. Supondo que os dados *flutuem* em torno de alguma relação entre X e Y, a Figura 13.8 também ilustra dois modelos matemáticos para essa estrutura. A reta (A): $y = x$ indica que, *em média*, os filhos têm alturas iguais à altura média de seus pais; a reta (B) representa a hipótese de Galton, a qual postulava uma tendência de que filhos de pais altos teriam alturas inferiores às alturas médias de seus pais, enquanto os filhos de pais baixos teriam alturas superiores às alturas médias de seus pais.

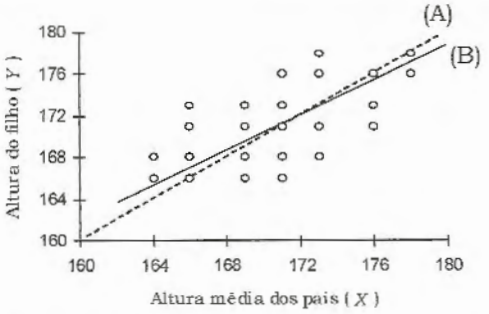


Figura 13.8 Diagrama de dispersão dos dados da Tabela 13.7 e ilustração de dois modelos matemáticos relacionando X e Y.

■

O Exemplo 13.4 se distingue dos exemplos anteriores por supor uma relação de causalidade entre X e Y, descrita em termos de uma equação matemática. É esta a diferença básica de um estudo de correlações e uma análise de regressão.

A análise de regressão é geralmente feita sob um referencial teórico que justifique a adoção de alguma relação matemática de causalidade.

O modelo da regressão linear simples

O modelo estatístico-matemático de regressão, em sua formulação mais simples, relaciona uma variável Y, chamada de variável *dependente* ou *resposta*, com uma variável X, denominada variável *explicativa* ou *independente*. Veja o quadro 13.1.

Quadro 13.1 Aplicações do modelo de regressão linear simples.

Variável independente (X)	→	Variável dependente (Y)
Renda	→	Consumo (R$)
Gasto com o controle da qualidade (R$)	→	Número de defeitos nos produtos
Memória RAM do computador (Gb)	→	Tempo de resposta do sistema (segundos)
Área construída do imóvel (m²)	→	Preço do imóvel (R$)

Assim como num estudo de correlações, a análise de regressão também toma por base um conjunto de observações pareadas (x, y), relativas às variáveis X e Y. Diremos que um dado valor y depende, em parte, do correspondente valor x. Por exemplo, a altura de um indivíduo (y) depende, em parte, da altura média de seus pais (x). Simplificaremos essa dependência por uma relação linear entre x e y, tal como:

$$y = \alpha + \beta x$$

Fixando valores para α e β, a equação $y = \alpha + \beta x$ é a equação de uma reta. Por exemplo, se $\alpha = 1$ e $\beta = 2$, a equação $y = 1 + 2x$ é uma certa reta, num par de eixos cartesianos. Para desenhar esta reta basta atribuir dois valores para X e calcular os correspondentes valores de Y. Digamos: $x = 0 \Rightarrow y = 1 + 2\times0 = 1$ e $x = 1 \Rightarrow y = 1 + 2\times1 = 3$. Com estes dois pontos, podemos traçar a reta da Figura 13.9.

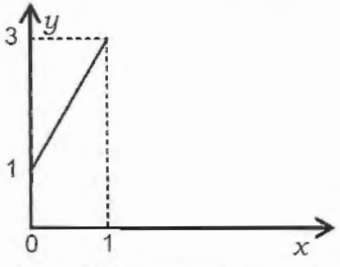


Figura 13.9 Representação gráfica da equação $y = 1 + 2x$.

Ao observarmos um conjunto de observações (x, y), verificamos que, em geral, os pontos não estão exatamente sobre uma reta, mas *flutuam* em torno de alguma reta imaginária. Então, um modelo para um par de observações pode ser:

$$y = \alpha + \beta x + \varepsilon$$

onde ε representa o *erro aleatório*, isto é, o efeito de uma infinidade de fatores que estão afetando a observação y de forma aleatória. Por exemplo, a altura de um indivíduo (y) não depende somente da altura média de seus pais (x), mas, também, de sua alimentação, do genótipo de seus ancestrais e de uma infinidade de outros fatores, tudo representado no modelo por ε.

No modelo $y = \alpha + \beta x + \varepsilon$, chamaremos de *parte estrutural* à parcela de y determinada por x, isto é, $\alpha + \beta x$. E o procedimento inicial da análise de regressão é encontrar estimativas para α e β, com base em uma amostra de observações (x, y).

Estimativas dos parâmetros α e β

A ideia básica da construção da parte estrutural do modelo, supostamente linear, é encontrar a reta que passe mais próximo possível dos pontos observados. Representaremos esta reta por:

$$\hat{y} = a + bx$$

e a chamaremos de *reta de regressão* ou *equação de regressão*. Veja a Figura 13.10.

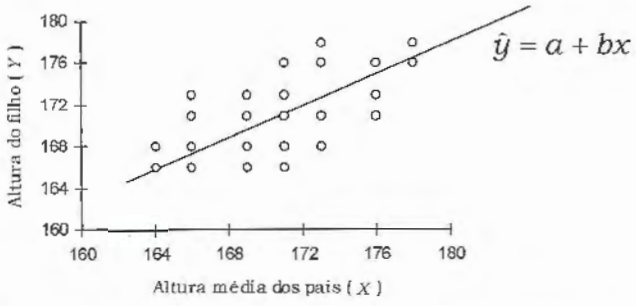


Figura 13.10 Representação da equação de regressão do Exemplo 13.4.

O chamado *método de mínimos quadrados* fornece as seguintes expressões para a equação de regressão:[6]

$$b = \frac{n \cdot \Sigma(X \cdot Y) - (\Sigma X) \cdot (\Sigma Y)}{n \cdot \Sigma X^2 - (\Sigma X)^2}$$

$$a = \frac{\Sigma Y - b \cdot \Sigma X}{n}$$

onde:

n é o número de pares (x, y) observados (tamanho da amostra);

$\Sigma(X \cdot Y)$ é somatório dos produtos $x \cdot y$ (primeiramente fazemos os produtos $x \cdot y$, relativos a todos os pares observados e, depois, efetuamos a soma);

ΣX é a soma dos valores de X;

ΣY é soma dos valores de Y; e

ΣX^2 é soma dos quadrados dos valores de X (primeiro elevamos os valores de X ao quadrado e, depois, efetuamos a soma).

[6] A obtenção da equação de regressão, pelo método de mínimos quadrados, consiste em fazer com que a soma quadrática dos efeitos aleatórios, $\Sigma\varepsilon^2$, seja a menor possível. A solução deste problema matemático gera as expressões de a e b. Veja, por exemplo, Wonnacott e Wonnacott (1991, p. 287).

EXEMPLO 13.5 Ilustraremos a construção da equação de regressão com parte das observações da *altura média dos pais* (X) e *altura do filho* (Y), extraídas da Tabela 13.7. A Tabela 13.8 mostra os cálculos dos somatórios.

Tabela 13.8 Parte das observações da Tabela 13.7 e cálculos intermediários para a construção da reta de regressão.

Dados		Cálculos intermediários	
X	Y	X^2	$X \cdot Y$
164	166	26.896	27.224
166	166	27.556	27.556
169	171	28.561	28.899
169	166	28.561	28.054
171	171	29.241	29.241
173	171	29.929	29.583
173	178	29.929	30.794
176	173	30.976	30.448
178	178	31.684	31.684
$\Sigma X = 1.539$	$\Sigma Y = 1.540$	$\Sigma X^2 = 263.333$	$\Sigma(X \cdot Y) = 263.483$

$$b = \frac{9 \cdot (263.483) - (1.539) \cdot (1.540)}{9 \cdot (263.333) - (1.539)^2} = \frac{1.287}{1.476} = 0{,}872$$

$$a = \frac{1.540 - (0{,}872).(1.539)}{9} = 22{,}00$$

Assim, temos a reta de regressão: $\hat{y} = 22 + (0{,}872)x$. Para traçar a reta no plano formado pelos eixos X e Y, basta atribuir dois valores para X e calcular os correspondentes valores de $\hat{y}$, pois, por dois pontos passa uma e apenas uma reta.[7] Veja a Figura 13.11.

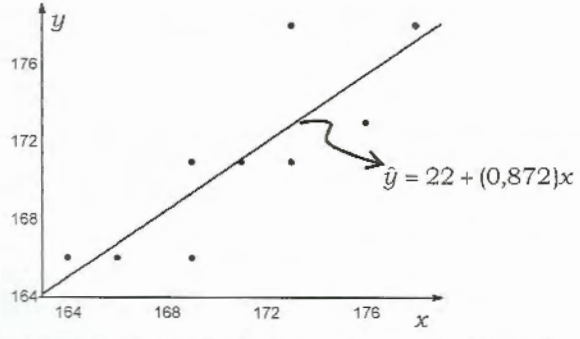


Figura 13.11 Diagrama de dispersão dos dados da Tabela 13.8 e a reta de regressão ajustada aos dados.

[1] Por exemplo, para um dado valor $x = 164 \Rightarrow \hat{y} = 22 + (0{,}872) \cdot (164) = 165{,}0$ e para $x = 178 \Rightarrow \hat{y} = 22 + (0{,}872) \cdot (178) = 177{,}2$. Marcamos os pontos (164; 165) e (178; 177,2) no plano formado pelos eixos X e Y, e traçamos a reta que passa por estes dois pontos.

Com a equação de regressão, podemos predizer a altura de um indivíduo ($\hat{y}$), com base na altura média de seus pais (x). Por exemplo, com uma altura média dos pais de $x = 175$ cm, temos uma predição da altura do filho de $\hat{y} = 22 + (0{,}872)\cdot(175) = 174$ cm.

■

O coeficiente b fornece uma estimativa da variação esperada de Y provocada pela variação de *uma* unidade em X. O sinal desse coeficiente indica o sentido (positivo ou negativo) da relação. No Exemplo 13.5, temos $b = 0{,}872$. Então, a cada centímetro a mais na altura média dos pais, esperamos um acréscimo de 0,872 cm na altura do filho.[8]

Variação explicada e não explicada

Ao ajustar uma equação de regressão aos dados, podemos estar interessados em verificar o quanto as variações da variável dependente, Y, podem ser explicadas por variações da variável independente, X, segundo o modelo especificado e para os dados da amostra.

Para cada valor x observado (ou estabelecido), temos o correspondente valor de Y, representado por y. Com o ajuste do modelo, temos também o *valor predito* por este: $\hat{y} = a + bx$. Por exemplo, para o oitavo indivíduo da amostra, temos $x = 176$ e o correspondente valor de Y ($y = 173$). Já o *valor predito* pela equação de regressão é $\hat{y} = 22 + (0{,}872)\cdot(176) = 175{,}47$ (ver Figura 13.12). A diferença entre o valor observado e o valor predito pelo modelo é chamada de *resíduo* – aquilo que a parte estrutural do modelo não consegue explicar.

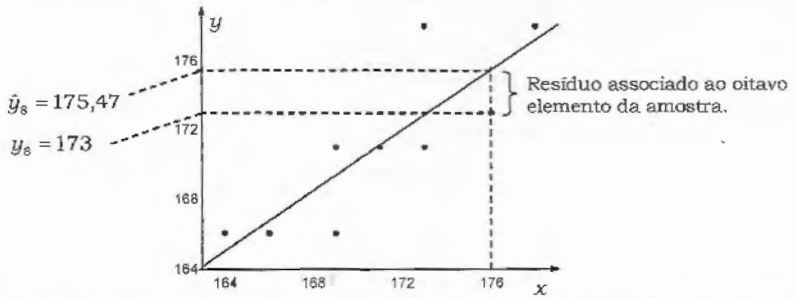


Figura 13.12 Valores observado e predito para $x = 176$ (oitavo elemento da amostra).

[8] A equação de regressão $\hat{y} = 22 + (0{,}872)x$ está compatível com a teoria de Galton, no sentido de que sua inclinação é inferior à da reta $y = x$. Contudo, os dados não estão provando a sua teoria, já que estamos analisando uma amostra extremamente pequena. A diferença da reta construída com base nos dados e a reta teórica, $y = x$, pode ser meramente casual. Para dar maior embasamento a essa discussão, pode ser feito um teste estatístico sobre os parâmetros do modelo. Ver, por exemplo, Chatterjee, Hadi e Price (2000).

Se desconsiderarmos a relação entre X e Y, então podemos predizer valores de Y, simplesmente, pela média aritmética de suas observações ($\overline{y}$). Naturalmente, nas situações em que X afeta Y, os resíduos em relação à média aritmética vão ser, em geral, maiores do que em relação à equação de regressão (Figura 13.13).

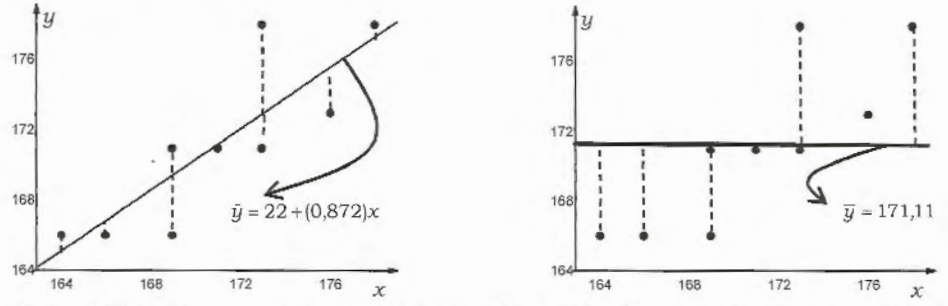


Figura 13.13 Resíduos em relação ao modelo de regressão ($y - \hat{y}$), e resíduos em relação à média aritmética ($y - \overline{y}$), com os dados do Exemplo 13.5.

A *soma de quadrados dos resíduos* é o numerador de uma estimativa para a variância da parte aleatória do modelo. Se a parte estrutural for avaliada apenas pela média aritmética, sem levar em conta qualquer relação entre Y e X, temos a *soma de quadrados total*:

$$SQT = \sum (y - \overline{y})^2$$

Se a parte estrutural levar em conta os diferentes valores de X pela equação de regressão, resta apenas aquilo que a equação de regressão não consegue explicar. É a chamada *soma de quadrados devida ao erro aleatório*, dada por:

$$SQE = \sum (y - \hat{y})^2$$

Observe na Figura 13.13 que SQE deve ser menor que SQT, especialmente quando a relação entre X e Y for forte. A diferença das duas somas de quadrados:

$$SQR = SQT - SQE$$

é conhecida como *soma de quadrados da regressão* e pode ser interpretada como a parte da variação de Y que a equação de regressão consegue explicar a mais do que simplesmente a média aritmética de Y.

O chamado *coeficiente de determinação*, dado por:

$$R^2 = \frac{SQR}{SQT} = \frac{\text{variação explicada}}{\text{variação total}}$$

pode ser interpretado como uma medida descritiva da proporção da variação de Y que pode ser explicada por X, segundo o modelo especificado. Em se tratando de regressão linear simples, pode-se mostrar, matematicamente, que o coeficiente de determinação R^2 é o quadrado do coeficiente de correlação *r de Pearson*, estudado na Seção 13.2.

Outra medida usada para avaliar o modelo é a variabilidade da parte aleatória, cujo desvio padrão pode ser estimado por:[9]

$$S_e = \sqrt{\frac{SQE}{n-2}}$$

Esta medida é usada para comparar modelos (quanto menor S_e, melhor é o ajuste do modelo aos dados).

Exemplo 13.5 (continuação) A Tabela 13.9 mostra o cálculo das somas de quadrados.

Tabela 13.9 Cálculo dos valores preditos, resíduos e somas de quadrados dos desvios (dados do Exemplo 13.5).

x	y	Média $\bar{y}$	$y-\bar{y}$	$(y-\bar{y})^2$	Preditos $\hat{y}$	Resíduos $y-\hat{y}$	$(y-\hat{y})^2$
164	166	171,11	-5,11	26,11	165,01	0,992	0,98
166	166		-5,11	26,11	166,75	-0,752	0,56
169	171		-0,11	0,01	169,37	1,632	2,66
169	166		-5,11	26,11	169,37	-3,368	11,36
171	171		-0,11	0,01	171,11	-0,112	0,01
173	171		-0,11	0,01	172,86	-1,856	3,46
173	178		6,89	47,47	172,86	5,144	26,42
176	173		1,89	3,57	175,47	-2,472	6,10
178	178		6,89	47,47	177,22	0,784	0,61
	Soma:		0	177		0	52

A Tabela 13.10 apresenta as somas de quadrados, sendo que *SQT* e *SQE* foram obtidas da Tabela 13.9, e *SQR* pela diferença das duas.

Tabela 13.10 Decomposição da variação de Y.

Fonte de variação	Somas de quadrados
Explicada por X pelo modelo de regressão (variação *explicada*)	$SQR = 125$
Devida ao erro aleatório (variação *não explicada*)	$SQE = 52$
Variação total	$SQT = 177$

[9] O subíndice e é para enfatizar que esta medida se refere ao erro aleatório.

Com os dados da Tabela 13.10,

$$R^2 = \frac{SQR}{SQT} = \frac{125}{177} = 0{,}706$$

Ou seja, dentre os nove indivíduos em estudo, as variações de suas alturas são explicadas, em parte, pela variação das alturas de seus pais ($R^2 \approx 70\%$ de explicação); e outra parte ($1 - R^2 \approx 30\%$) em razão de outros fatores.

O desvio padrão da parte aleatória (aquela que não pode ser explicada por variações das alturas dos pais) é:

$$S_e = \sqrt{\frac{SQE}{n-2}} = \sqrt{\frac{52}{7}} = 2{,}73$$

■

Uso do computador

Exemplo 13.6 O anexo deste capítulo contém dados relativos a cinquenta apartamentos da cidade de Criciúma – SC. Com o objetivo de construir um modelo para subsidiar a atualização dos valores dos tributos municipais, vamos realizar uma regressão entre valor (Y), em milhares de reais, e área privativa (X), em m^2. Usando o *Excel*, obtivemos os resultados apresentados na Figura 13.14:[10]

Estatísticas da regressão	
R múltiplo	0,881
R-quadrado	0,777
R-quadrado ajustado	0,772
Desvio padrão	43,3
Observações	50

ANOVA

	gl	*SQ*	*MQ*	*F*	Valor *p*
Regressão	1	313.285,6	313.285,6	166,93	0,0000
Resíduo	48	90.082,0	1.876,7		
Total	49	403.368,6			

	Coeficientes	Erro padrão	Estat. *t*	Valor *p*	Intervalo de confiança (95,0%)	
Interseção	-64,57	14,66	-4,40	0,000	-94,0	-35,1
Valor novo	1,67	0,13	12,92	0,000	1,4	1,9

Figura 13.14 Resultados de uma análise de regressão pelo Excel® (Exemplo 13.6).

[10] Para realizar a análise, no menu principal do Excel, clicar em Ferramentas, Análise de Dados e Regressão. Se, ao clicar em Ferramentas, não aparecer Análise de Dados, clique em Suplementos e assinale Ferramentas de Análise. Quanto aos resultados, os termos foram adequados à língua portuguesa e aos termos técnicos deste livro. Os valores foram formatados como *números*.

A primeira tabela da Figura 13.14 mostra algumas estatísticas e, em particular, o R^2 (*R-quadrado*) igual a 0,777. Este resultado indica que na amostra, cerca de 78% da variação do valor de venda do apartamento pode ser *explicada* por uma relação linear com a área privativa. Os demais 22% são a parcela da variação provocada por outros fatores não incluídos no modelo de regressão. Essa parte aleatória tem desvio padrão estimado de S_e = 43,3 mil reais.[11] Na primeira linha da tabela, tem-se o chamado coeficiente de correlação múltiplo, que, no caso de apenas uma variável independente, é o coeficiente r de Pearson (Seção 13.2).

A segunda tabela apresenta a análise de variância (ANOVA) do modelo. A coluna *SQ* apresenta as somas de quadrados. Mas o mais importante são os resultados de um teste estatístico para as hipóteses:

H_0: não existe relação linear entre X e Y; e
H_1: a relação linear entre X e Y é significativa (não é mero resultado do acaso).

O teste, conhecido como teste F do modelo, resultou em F = 166,93, com correspondente valor p = 0,0000. Como o valor p é extremamente pequeno, o teste estatístico rejeita H_0, indicando que a área privativa do apartamento (X) é significativa para *explicar* o seu preço (Y).

A terceira tabela fornece várias informações relevantes. A primeira coluna apresenta as estimativas dos coeficientes, de onde extraímos a equação de regressão:

$$\hat{y} = -64{,}57 + (1{,}67)x$$

ou seja, tendo a área privativa (x) podemos obter uma previsão para o preço do imóvel ($\hat{y}$). Por exemplo, um apartamento com área privativa de 100 m² tem seu valor predito pelo modelo de:

$$\hat{y} = -64{,}57 + (1{,}67) \cdot (100) = 102{,}43$$

ou seja, R$ 102.430,00.

Interpretando o coeficiente de x, temos que, a cada m² a mais de área, estima-se que o valor do apartamento aumenta em b = 1,67 mil reais.

A última tabela fornece os resultados de testes estatísticos sobre cada um dos parâmetros do modelo. Em particular, na regressão simples,

[11] Observe que, embora o R^2 indique um ajuste razoável, o desvio padrão mostra que este modelo ainda não é adequado na prática, pois, pela distribuição normal, é natural valores se afastarem da média em até dois desvios padrões. Ou seja, as predições baseadas no modelo podem predizer valores de venda com mais de 86 mil reais de diferença do valor efetivamente vendido.

o teste sobre o parâmetro b (inclinação) é equivalente ao teste F da análise de variância sobre o modelo. As duas últimas colunas dessa tabela apresentam um intervalo de 95% de confiança para os dois parâmetros do modelo (o intercepto a e a inclinação β), com a mesma interpretação dos intervalos de confiança discutidos no Capítulo 9.

EXERCÍCIO

9) Sejam os dados de *número de faltas* e *nota na prova* de uma turma de Estatística:

Número de faltas	8	2	5	0	1	4	10	2
Nota na prova	7	10	6	10	8	5	2	8

a) Qual deve ser a variável dependente, e qual a independente? (Escolha a que lhe faz mais sentido.)
b) Estabeleça a equação de regressão.
c) Faça um gráfico com os pontos observados e a reta de regressão.
d) Calcule o coeficiente R^2.
e) Calcule S_e.
f) Quais são as principais informações que podem ser obtidas pela presente análise?

10) Na década de 1970, em várias regiões, houve um movimento migratório que fez crescer bastante a população urbana nos municípios médios e grandes. Neste contexto, vamos tentar *explicar* o crescimento demográfico de um município em função de sua população urbana, considerando dados de doze importantes municípios catarinenses, no período em discussão.

Pop. urb. (em 1.000 hab.)	101	193	42	304	42	152	55	105	68	219	129	42
Taxa de crecimento dem.	3,2	4,6	2,8	6,5	2	1,9	2,9	5,3	2,7	3,1	3,1	1,2

a) Qual deve ser a variável dependente, e qual a independente?
b) Estabeleça a equação de regressão.
c) Faça um gráfico com os pontos observados e a reta de regressão.
d) Qual é a taxa de crescimento demográfico, predita pela equação de regressão, para um município de 300 mil habitantes?
e) Calcule o coeficiente R^2.
f) Quais são as principais informações que podem ser obtidas pela presente análise?

11) (Fazer com o auxílio do computador.) Considerando que a satisfação de um aluno com um curso universitário (Y) pode ser afetada pelo seu desempenho no curso (X), faça uma análise de regressão usando os dados do anexo do Capítulo 2. Interprete os resultados.

13.5 Análise dos resíduos e transformações

Na seção anterior, estabelecemos um modelo para um conjunto de observações (x, y), relativo às variáveis X e Y, da forma

$$y = \alpha + \beta x + \varepsilon$$

onde α e β são parâmetros a serem estimados com os dados e ε representa o *erro aleatório*. Ou seja, estamos assumindo que X causa Y através de uma relação linear, e toda variação em torno dessa relação deve-se ao efeito do erro aleatório. Além disso, para a validade dos intervalos de confiança e testes estatísticos discutidos no Exemplo 13.6, é necessário supor que as observações de Y sejam independentes, e o termo de erro tenha distribuição aproximadamente normal com média nula e variância constante. Apresentaremos um processo gráfico para verificar se estas suposições podem ser válidas e, caso contrário, o que pode ser feito para adequar o modelo.

Um primeiro gráfico pode ser feito antes da análise de regressão. É o diagrama de dispersão, conforme discutido na Seção 13.1. Por esse gráfico, podemos verificar se a função linear é adequada para representar a forma estrutural entre X e Y. Veja o gráfico à esquerda da Figura 13.15.

Após a estimação dos parâmetros do modelo, podemos calcular os *resíduos* do modelo ajustado aos dados. O resíduo é calculado para cada observação, e definido como a diferença entre o valor observado y e o valor *predito* $\hat{y}$. Ou seja,

$$\textit{resíduo} = y - \hat{y}$$

Um gráfico apresentando os pares $(x, \textit{resíduo})$ é bastante útil na avaliação do modelo de regressão. Veja o gráfico à direita da Figura 13.15.

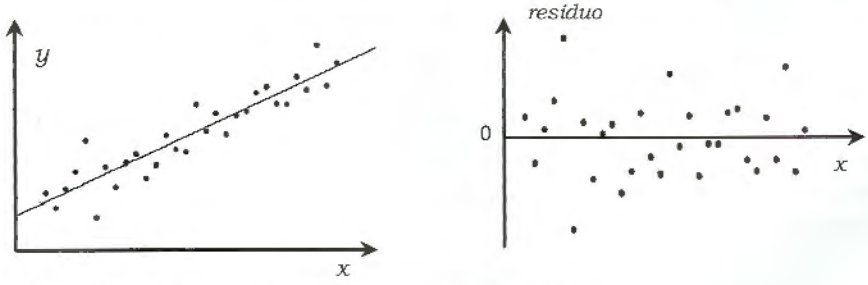


Figura 13.15 Gráficos para verificar a adequação do modelo.

Os gráficos da Figura 13.15 indicam uma situação em que as suposições do modelo estão aparentemente satisfeitas, pois os resíduos apresentam-se distribuídos de forma aleatória e razoavelmente simétrica em torno da reta de regressão. No gráfico dos resíduos, a reta de regressão corresponde à linha horizontal sobre o valor zero.

A Figura 13.16 apresenta uma situação em que temos um ponto discrepante. Esse ponto é visível nos dois gráficos, mas no gráfico dos resíduos ele aparece mais nitidamente. Seja:

$$\textit{resíduo padronizado} = \frac{y - \hat{y}}{S_e}$$

Supostamente, os resíduos padronizados devem seguir uma distribuição normal padrão, pelo menos aproximadamente. Então, em torno de 95%, os valores devem estar entre 2 ou -2 (Capítulo 8). Fora deste intervalo, são casos suspeitos de serem discrepantes. Assim, o uso de resíduos padronizados é melhor para detectar pontos discrepantes.

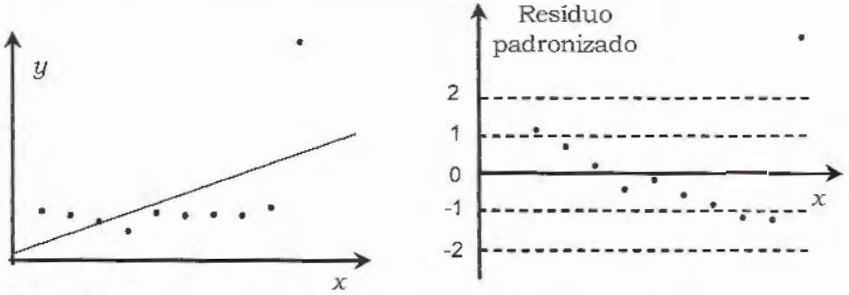


Figura 13.16 Gráficos indicando a presença de um valor discrepante.

A Figura 13.16 mostra como um ponto discrepante pode *forçar* uma inclinação na reta, sugerindo uma tendência não compatível com as demais observações. Esse problema surge, principalmente, quando se tem uma amostra pequena, e o ponto discrepante estiver numa das extremidades do intervalo de observação de X. É prudente, neste caso, buscar a razão da existência desse ponto discrepante. Se a sua causa for algum erro, alguma falha no experimento ou, ainda, puder ser considerada uma situação atípica, devemos efetuar nova análise sem a observação discrepante.

Quando se trata de um estudo experimental, a variável X costuma ser estabelecida. Por exemplo, num estudo para verificar a relação entre o tempo de cozimento (X) e a maciez (Y) de um alimento, podemos estabelecer diferentes tempos de cozimento e observar os resultados de

Y. Recomendamos variar *X* uniformemente sobre o intervalo de estudo. Por exemplo, se pretendemos fazer a análise entre 20 e 30 minutos de cozimento, podemos fazer ensaios com os tempos de cozimentos de 20, 21, 22, ..., 30 minutos.

Em estudos de levantamento, normalmente *X* e *Y* são observadas, sendo comum ocorrer uma distribuição assimétrica de valores de *X*. Por exemplo, considere o problema de se avaliar a relação entre renda (*X*) e consumo (*Y*) de indivíduos de certa região. A maioria dos indivíduos tem renda baixa e, consequentemente, tendem a consumir pouco, provocando distribuições assimétricas para *X* e *Y*. Assim, os dados devem se distribuir conforme mostra a Figura 13.17.

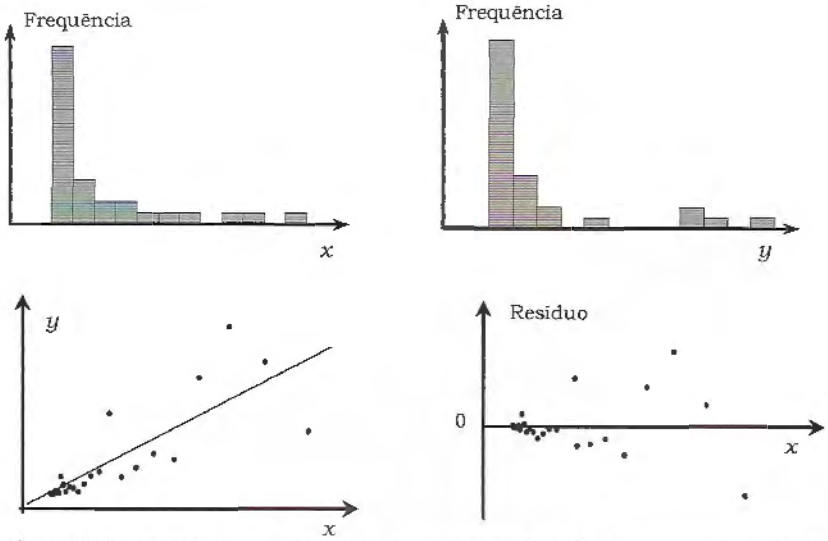


Figura 13.17 Gráficos indicando distribuições assimétricas de *X* e *Y*, além da variância de *Y* ser maior para valores maiores de *X* e *Y*.

Nesta situação, os valores grandes de *X* vão ter mais peso na determinação da inclinação da reta. Neste caso, recomendamos a aplicação da transformação logarítmica, tanto nos valores de *X* como nos valores de *Y*, estabelecendo o seguinte modelo:[12]

$$\log(y) = \alpha + \beta \cdot \log(x) + \varepsilon$$

[12] É comum usar o logaritmo natural ou na base 10. Outra transformação que se presta ao mesmo propósito é a raiz quadrada. Esta segunda transformação é usada nas situações em que a inadequação do modelo não aparece de forma tão forte como visto na Figura 13.17. Observamos que estas transformações são possíveis somente quando todos os valores são positivos.

A transformação logarítmica aumenta as distâncias entre os valores pequenos e reduz as distâncias entre os valores grandes, tornando distribuições assimétricas de cauda longa à direita em distribuições mais simétricas. Com isso, temos uma situação mais adequada para estabelecer a reta de regressão. Em termos computacionais, devemos:

a) calcular o logaritmo natural de cada valor x e de cada valor y;
b) aplicar a análise de regressão linear sobre os dados transformados [log(x), log(y)]; e
c) construir novamente o gráfico de resíduos para verificar a adequação das suposições neste novo modelo.

A Figura 13.18 apresenta uma situação que sugere relação *não linear*, com Y crescendo rapidamente para valores pequenos de X, e crescendo lentamente para valores grandes de X. É uma situação em que recomendamos uma transformação logarítmica (ou raiz quadrada) somente nos valores da variável X, ou seja, passamos a considerar o seguinte modelo para os dados:

$$y = \alpha + \beta \cdot \log(x) + \varepsilon$$

Note que esse modelo pode ser considerado linear em termos das variáveis log(x) e y (não mais entre x e y). Em termos computacionais, devemos:

a) calcular o logaritmo de cada valor x;
b) aplicar a análise de regressão linear sobre os dados [log(x), y]; e
c) construir novamente o gráfico de resíduos para verificar a adequação das suposições nesse novo modelo.

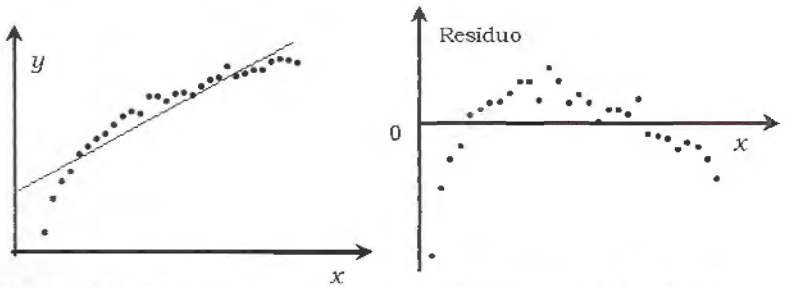


Figura 13.18 Gráficos indicando uma relação *não linear*, aparentemente logarítmica.

A Figura 13.19 apresenta uma situação com os seguintes problemas: (1) relação *não linear* para a parte estrutural do modelo e (2) aumento da variância à medida que X aumenta. Recomendamos uma

transformação logarítmica nos valores da variável Y, ajustando o seguinte modelo aos dados:

$$\log(y) = \alpha + \beta x + \varepsilon$$

Para ajustar o modelo, devemos:

a) calcular o logaritmo de cada valor y;

b) aplicar a análise de regressão linear sobre os dados $[x, \log(y)]$; e

c) construir novamente o gráfico de resíduos para verificar se o novo modelo é mais adequado aos dados.

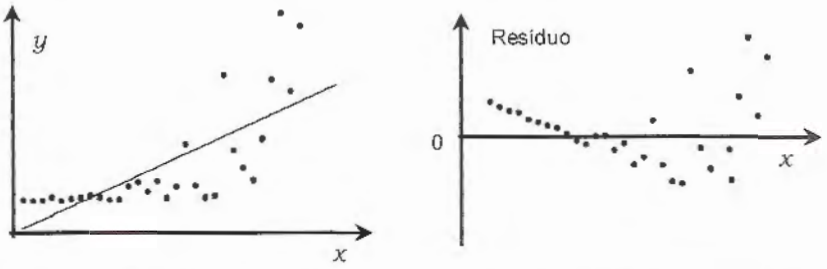


Figura 13.19 Gráficos indicando uma relação *não linear* – aparentemente exponencial – e variância não constante.

O uso de transformações auxilia o pesquisador a encontrar um modelo mais adequado para os dados, ainda que utilizando as expressões da regressão linear. A transformação logarítmica é muito usada por ter uma interpretação prática interessante, já que transforma variações percentuais de mesma magnitude em variações constantes. Por exemplo, se considerar um aumento absoluto no salário de R$ 100,00, o seu significado vai ser muito diferente para quem ganha R$ 100,00 e para quem ganha R$ 1.000,00. Por isso, é mais comum ouvir falar em aumentos percentuais de salários. Um aumento de 10% no salário representa um ganho de R$ 10,00 para quem ganha R$ 100,00 e um ganho de R$ 100,00 para quem ganha R$ 1.000,00. Na escala logarítmica, esses incrementos são iguais. Por esta razão, é comum usar a escala (ou transformação) logarítmica em variáveis econômicas ou medidas de tamanho em geral.

Exemplo 13.6 (continuação) Na seção anterior foi realizada uma regressão do valor de um imóvel (Y) com relação a sua área privativa (X), considerando uma amostra de cinquenta apartamentos, apresentada no anexo deste capítulo. A Figura 13.20 apresenta a reta de regressão e o gráfico dos resíduos desse modelo.

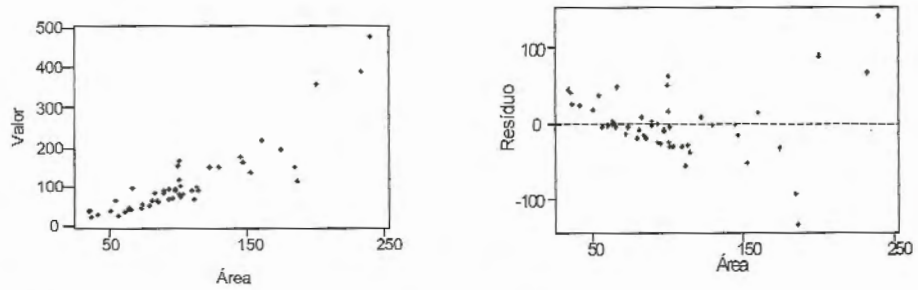


Figura 13.20 Gráficos de dispersão e dos resíduos (Exemplo 13.6).

Observamos na Figura 13.20 uma predominância de valores pequenos com respeito às duas variáveis. Isto era esperado porque são mais comuns apartamentos pequenos (área e preço pequenos) do que apartamentos grandes (área e preço grandes). Também podemos observar maior variabilidade nos apartamentos mais caros. Essas condições sugerem tentarmos uma transformação logarítmica em X e em Y. Assim, foi aplicado o logaritmo natural em cada um dos cinquenta valores de X e Y. Por exemplo, o primeiro apartamento da amostra tem $x = 96$ m^2 e $y =$ 69 mil reais. Aplicando o logaritmo natural, encontramos:

$$\log(x) = \log(96) = 4{,}56 \quad \text{e} \quad \log(y) = \log(69) = 4{,}23$$

A análise com os dados transformados produziu os gráficos de dispersão e de resíduos apresentados na Figura 13.21.

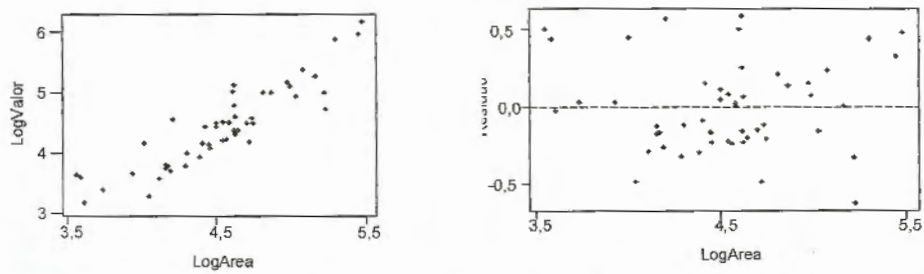


Figura 13.21 Gráficos de dispersão e dos resíduos (Exemplo 13.6), após transformações nas variáveis.

Verificamos pela Figura 13.21 que, após as transformações, as condições básicas do modelo estão aparentemente satisfeitas. A equação de regressão, obtida com apoio de um sistema computacional para análise estatística é:

$$\text{Predição de } \log(y) = -1{,}58 + (1{,}33) \cdot \log(x)$$

com $R^2 = 0{,}813$ e $S_e = 0{,}294$. Observar que o poder explicativo deste modelo é melhor que o anterior (81,3% contra 77,7%). Já o S_e não é comparável devido a transformação de escala.

Para predizer o valor de um apartamento com área privativa de 100 m², devemos, primeiramente, transformar este valor na escala logarítmica:

$$x = 100 \rightarrow \log(x) = 4{,}605$$

Aplicar o modelo de regressão:

$$\text{Predição de } \log(y) = -1{,}58 + (1{,}33) \cdot (4{,}605) = 4{,}545$$

Efetuar a transformação inversa do logaritmo:

$$\hat{y} = \exp\{4{,}545\} = 94{,}15$$

Assim, por este novo modelo, o apartamento valeria R$ 94.150,00.

■

13.6 Introdução à regressão múltipla

Em geral, uma variável dependente (ou resposta) Y depende de várias variáveis independentes ou explicativas (X_1, X_2, ..., X_k). Na análise de regressão múltipla, vamos construir um modelo estatístico-matemático para se estudar, objetivamente, a relação entre as variáveis independentes e a variável dependente e, com o modelo construído, conhecer a influência de cada variável independente, como também predizer a variável dependente em função do conhecimento das variáveis independentes. O Quadro 13.2 ilustra alguns exemplos.

Quadro 13.2 Aplicações do modelo de regressão múltipla.

Variáveis independentes (X_1, X_2, ..., X_k)	→	Variável dependente (Y)
X_1 = altura do pai (cm) X_2 = altura da mãe (cm) X_3 = sexo (1 = homem, 0 = mulher)	→	Y = altura de um indivíduo (cm)
X_1 = renda (R$) X_2 = poupança (R$) X_3 = taxa de juros (%)	→	Y = Consumo (R$)
X_1 = área construída do imóvel (m²) X_2 = idade (anos) X_3 = localização	→	Y = preço do imóvel (R$)
X_1 = memória RAM (Gb) X_2 = sistema operacional X_3 = tipo de processador	→	Y = tempo de resposta do sistema computacional (segundos)

Para estabelecer o modelo clássico de regressão múltipla, consideraremos que Y seja uma variável quantitativa contínua e X_1, X_2, ..., X_k sejam variáveis quantitativas ou indicadoras de certos atributos. A variável indicada deve ter valor 1 quando o atributo está presente; e 0 quando não está presente. Por exemplo, a variável X_3 = *localização do imóvel* pode ter valor 1 quando o imóvel estiver numa área valorizada, e 0 quando estiver numa área pouco valorizada. Também será considerado que Y é uma variável aleatória, isto é, somente será conhecida após a observação do elemento (indivíduo, imóvel, etc.), enquanto X_1, X_2, ..., X_k também podem provir de observação ou serem estabelecidas *a priori*.

A análise de regressão múltipla parte de um conjunto de observações $(x_1, x_2, ..., x_k, y)$, relativas às variáveis X_1, X_2, ..., X_k e Y. Diremos que um dado valor y depende dos correspondentes valores $x_1, x_2, ..., x_k$, mas também de uma infinidade de outros fatores não incluídos no modelo, que serão representados por ε (*erro aleatório*). Mais especificamente, supomos o seguinte modelo para as observações:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$$

onde α, β_1, β_2, ... , β_k são parâmetros a serem estimados com os dados e ε representa o *erro aleatório*, cujo desvio padrão também pode ser estimado pelos dados. As suposições são análogas às suposições da regressão simples, acrescentando que as variáveis independentes X_1, X_2, ..., X_k não devem ter correlações altas entre si.

EXEMPLO 13.7 Voltando à questão de construir um modelo para o valor de um apartamento (Y) com os dados do anexo deste capítulo. Sejam as variáveis independentes:

X_1 = área comum do apartamento (m^2);
X_2 = idade (anos);
X_3 = consumo de energia elétrica do morador (Kw/mês) e
X_4 = localização (1= área valorizada; 0 = área pouco valorizada).

Como discutimos no Exemplo 13.6, as variáveis Y e X_1 serão analisadas na escala logarítmica. A variável X_3 está sendo usada como uma *proxi* do padrão de vida do morador do apartamento e, por sua vez, da qualidade do apartamento. Temos o seguinte modelo teórico para os dados:

$$\log(y) = \alpha + \beta_1 \log(x_1) + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

Usando o *Excel*®, obtivemos os resultados apresentados na Figura 13.22.[13]

[13] Nos resultados, os termos foram adequados à língua portuguesa e aos termos técnicos deste livro. Os valores foram formatados como *números*.

Estatísticas da regressão	
R múltiplo	0,943
R-quadrado	0,889
R-quadrado ajustado	0,879
Desvio padrão	0,234
Observações	50

ANOVA

	gl	*SQ*	*MQ*	*F*	Valor *p*
Regressão	4	19,702	4,926	89,863	0,000
Resíduo	45	2,467	0,055		
Total	49	22,169			

	Coeficientes	Erro padrão	Estat. *t*	Valor *p*	Intervalo de confiança (95,0%)	
Interseção	-1,208	0,376	3,210	0,002	-1,966	-0,450
LogArea	1,195	0,084	14,242	0,000	1,026	1,364
Idade	-0,025	0,005	-4,623	0,000	-0,036	-0,014
Energia	0,0024	0,0016	1,5214	0,135	-0,001	0,0057
Local	0,076	0,076	1,010	0,318	-0,076	0,229

Figura 13.22 Resultados de uma análise de regressão pelo Excel® (Exemplo 13.7).

Observamos, na primeira tabela da Figura 13.22, o valor de R^2 (*R-quadrado*) igual a 0,889 e S_e = 0,234. Comparando com os resultados do Exemplo 13.6 (R^2 = 0,813 e S_e = 0,294), vemos melhora no modelo com a inclusão das variáveis: idade, gasto de energia elétrica e localização. O valor R^2 = 0,889, indica que quase 90% da variação do logaritmo do valor de um apartamento pode ser *explicado* por uma relação linear que envolve o logaritmo da área comum (X_1), idade (X_2), consumo de energia elétrica do morador (X_3) e dois níveis de localização (X_4).

A segunda tabela (ANOVA) fornece o resultado de um teste estatístico da seguinte hipótese nula:

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

ou seja, por esta hipótese, o conjunto de variáveis independentes em estudo não tem poder *explicativo* sobre a variável dependente.[14] Este teste, conhecido como *teste F do modelo*, resultou na estatística F = 89,863, com correspondente valor p extremamente pequeno (menor que um milésimo). Assim, o teste estatístico rejeita H_0, indicando que as variáveis independentes escolhidas são significativas para *explicar* a variável dependente.

A terceira tabela fornece as estimativas dos coeficientes, incluindo intervalos de confiança e testes estatísticos para cada coeficiente. A primeira coluna apresenta as estimativas dos coeficientes, de onde podemos extrair a seguinte equação:

[14] Cabe observar que o teste estatístico refere-se à população, ou seja, quando se tem uma amostra muito pequena, podemos obter um valor alto de R^2 e o teste aceitar H_0.

$$\text{Predição de } \log(y) = -1{,}208 + 1{,}195 \cdot \log(x_1) - 0{,}025x_2 + 0{,}0024x_3 + 0{,}076x_4$$

Assim, tendo a área do apartamento (x_1), a idade (x_2), o consumo de energia elétrica (x_3) e a localização (x_4) podemos obter uma predição de seu valor. Por exemplo, um apartamento com 100 m², que tenha 5 anos de uso, morador consumindo 200 Kw e localização em área valorizada, temos:

$$\text{Predição de } \log(y) = -1{,}208 + 1{,}195 \cdot \log(100) - (0{,}025) \cdot 5 + (0{,}0024) \cdot 200 + (0{,}076) \cdot 1$$

ou: Predição de $\log(y) = 4{,}726$. Portanto: $\hat{y} = \exp(4{,}726) = 112{,}84$

ou, seja, valor estimado de R$ 112.840,00.

Devemos observar que os sinais dos coeficientes do modelo construído estão coerentes. Coeficiente de X_1 positivo, isto é, quanto maior o apartamento, maior deverá ser o seu valor; coeficiente de X_2 negativo (quanto mais velho, menor o valor); coeficiente de X_3 positivo (quanto maior o consumo de energia do morador, maior o valor); e coeficiente de X_4 positivo (em área valorizada, maior o valor).

A última tabela também fornece os resultados de testes estatísticos para cada variável. Pelos valores *p*, verificamos que as variáveis *energia* e *local* são não significativas e, portanto, poderiam ser excluídas do modelo sem que os indicadores de qualidade do ajuste (R^2 e S_e) piorem demasiadamente. Isso não significa que a localização não seja relevante para explicar o valor do imóvel, mas seu efeito já pode estar parcialmente incluído nas outras variáveis independentes.

■

Para verificar a adequação de um modelo de regressão múltipla, podemos calcular os *resíduos* e, com base neles, fazer uma análise gráfica similar a que foi feita em regressão simples.

EXERCÍCIOS COMPLEMENTARES

12) Para verificar se existe correlação entre X = *tamanho da ninhada* e Y = *número de brincadeiras filhote-mãe*, em *hamsters* dourados, observaram-se o relacionamento de um filhote com sua mãe, em cada uma das 20 ninhadas de mesmo tempo de vida, durante uma hora. Anotaram-se, para cada ninhada, os valores das variáveis X e Y e calculou-se o valor do coeficiente r nessa amostra: $r = -0{,}20$. Podemos concluir que realmente existe correlação entre X e Y, ao nível de significância de 5%?

13) Para cada um dos itens abaixo, calcule um coeficiente de associação (ou de correlação) e interprete. Escolha o coeficiente de acordo com a forma de medida das variáveis.
   a) Para avaliar o relacionamento entre *renda familiar* (em unidades de salários mínimos) e *número de filhos* nas seis famílias de uma pequela localidade,

observaram-se os seguintes valores de renda familiar: 1, 2, 4, 8, 12 e 20; e os respectivos números de filhos: 4, 5, 5, 3, 2 e 2.

b) Para avaliar o relacionamento entre *peso* e *altura* de um grupo de 10 indivíduos, fez-se a classificação cruzada, apresentada na tabela abaixo:

peso	altura		
	baixa	mediana	alta
baixo	2	1	1
mediano	0	2	0
alto	1	1	2

c) Para avaliar o relacionamento entre *sexo* e *altura*, num grupo de 100 pessoas adultas, observou-se que das 40 mulheres, 30 eram baixas e 10 eram altas, enquanto, dos 60 homens, observaram-se 40 altos e 20 baixos.

14) Com o objetivo de verificar se numa certa região existe correlação entre o *nível de escolaridade médio dos pais* e o *nível de escolaridade dos filhos*, observou-se uma amostra aleatória de 8 indivíduos adultos, verificando o número de anos que estes frequentaram (e tiveram aprovação) em escolas regulares (Y) e o número médio de anos que os seus pais frequentaram (e tiveram aprovação) em escolas regulares (X). Os resultados da amostra são apresentados abaixo:

X	0	0	2	3	4	4	5	7
Y	2	3	2	5	9	8	8	15

a) Calcule o coeficiente de correlação de Pearson.
b) Em termos do resultado do item (a), o que se pode dizer sobre a correlação entre o número de anos que os 8 indivíduos frequentaram escolas regulares (Y) e o número médio de anos que os seus pais frequentaram escolas regulares?
c) Estabeleça a reta de regressão de Y em relação a X.
d) Apresente o diagrama de dispersão acompanhado da reta de regressão.

15) Um administrador de uma grande sorveteria anotou por um longo período de tempo a *temperatura média diária*, em °C (X), e o *volume de vendas diária de sorvete*, em kg (Y). Com os dados, estabeleceu uma equação de regressão, resultando em:

$$y = 0{,}5 + 1{,}8x, \text{ com } R^2 = 0{,}80$$

Pergunta-se:
a) Qual é o consumo esperado de sorvete num dia de 27°C?
b) Qual é o incremento esperado nas vendas de sorvete a cada 1°C de aumento da temperatura?

16) A tabela, a seguir, relaciona os pesos (em centenas de kg) e as taxas de rendimento de combustível em rodovia (km/litro), numa amostra de 10 carros de passeio novos.

Peso	12	13	14	14	16	18	19	22	24	26
Rendimento	16	14	14	13	11	12	09	09	08	06

a) Calcule o coeficiente de correlação de Pearson.
b) Considerando o resultado do item (a), como você avalia o relacionamento entre *peso* e *rendimento*, na amostra?
c) Para estabelecer uma equação de regressão, qual deve ser a variável dependente e qual deve ser a variável independente? Justifique a sua resposta.

d) Estabeleça a equação de regressão, considerando a resposta do item (c).
e) Apresente o diagrama de dispersão e a reta de regressão obtida em (d).
f) Você considera adequado o ajuste do modelo de regressão do item (d)? Dê uma medida desta adequação interpretando-a.
g) Qual é o rendimento esperado para um carro de 2.000 kg? Use o modelo do item (d). Lembrete: os dados de peso na tabela estão em *centenas* de kg.
h) Você considera seu estudo capaz de predizer o rendimento esperado de um veículo com peso de 7.000 kg? Justifique sua resposta.

ANEXO

Dados de apartamentos de Criciúma – SC. Variáveis: valor (em milhares de reais), área privativa (m^2), idade (anos), consumo mensal de energia elétrica (Kw) e local (1 = região mais valorizada; 0 = região menos valorizada).

Valo	Área	Idade	Energia	Local	Valor	Área	Idade	Energia	Local
69	96	14	170	1	98	114	4	170	1
176	145	8	144	1	120	101	4	192	1
195	175	2	147	1	51	80	14	170	1
80	101	4	160	1	90	115	2	128	0
390	233	2	220	1	65	55	2	118	0
360	201	6	228	1	90	98	12	143	1
80	104	2	160	1	219	161	6	175	1
45	64	14	118	0	167	101	4	192	1
153	100	2	174	1	63	85	12	172	0
66	112	17	181	1	150	123	4	154	1
90	90	2	144	1	36	61	12	163	0
114	187	28	146	0	139	153	8	144	1
165	147	4	183	0	39	51	18	135	0
101	102	2	160	1	24	37	14	163	1
150	185	8	144	1	84	83	16	147	1
75	102	6	180	0	96	67	2	118	0
38	35	6	144	1	65	82	4	147	0
68	94	28	146	0	30	42	2	160	0
90	110	14	158	0	41	66	12	154	1
60	86	6	146	0	476	240	2	183	1
55	74	10	147	0	43	64	18	184	0
92	98	4	160	0	27	57	14	143	0
84	90	4	147	1	44	65	12	147	0
92	94	12	187	0	44	73	12	128	0

Fonte: Amostra extraída dos dados da dissertação de mestrado ZANCAN, Evelise C. *Metodologia para Avaliação em Massa de Imóveis para Efeito de Cobrança de Tributos Municipais – Caso de Apartamentos da Cidade de Criciúma, Santa Catarina.* UFSC, Florianópolis, 1995. Com adaptações.

Referências

AGRESTI, A. *Analysis of ordinal categorical data.* USA: John Wiley, 1984.

BARBETTA, P. A.; REIS, M. M.; BORNIA, A. C. *Estatística para cursos de Engenharia e Informática.* São Paulo: Atlas, 2004.

BOLFARINE, H.; BUSSAB, W. O. *Elementos de amostragem.* São Paulo: Edgard Blücher, 2005.

BLALOCK, H. M. *Social statistics.* USA: Mc. Graw-Hill, 1960.

BOX, G. E. P.; HUNTER, W. G.; HUNTER, J. S. *Statistics for experimenters.* Canadá: John Wiley, 1978.

BUSSAB, W. O.; MORETTIN, P. A. *Estatística básica.* 5.ed. São Paulo: Saraiva, 2002.

CHATTERJEE, S.; HADI, A. S.; PRICE, B. *Regression analysis by examples.* 3. ed. USA: John Wiley, 2000.

COCHRAN, W. G. *Sampling techniques.* 3. ed. USA: John Wiley, 1977.

FISHER, R. A. *The design of experiments.* 6. ed. Edinburgo: Oliver and Boyd, 1951.

LEACH, C. *Introduction to statistics:* a nonparametric approach for the social sciences. USA: John Wiley, 1979.

LEVIN, J. *Estatística aplicada às ciências humanas.* 2. ed. São Paulo: Harbra, 1985.

LEVINE, D. M.; BERENSON, M. L.; STEPHAN, D. *Estatística*: teoria e aplicações usando o Excel. Rio de Janeiro: LTC, 2000.

MAGALHÃES, A. N.; LIMA, A. C. P. *Noções de probabilidade e estatística.* 4. ed. São Paulo: EDUSP, 2002.

MENDENHALL, N. *Probabilidade e estatística,* v. 1 e 2. Rio de Janeiro: Campos, 1985.

NOETHER, G. F. *Introdução à Estatística:* uma abordagem não paramétrica. 2. ed. Rio de Janeiro: Guanabara Dois, 1983.

SELLTIZ, G. I.; WRIGHTSMAN, L. S.; COOK, S. W. *Métodos de pesquisa nas relações sociais.* 4. ed. São Paulo: EPU, 1987.

SIEGEL, S. *Estatística não paramétrica aplicada às ciências do comportamento.* Rio de Janeiro: Mc. Graw Hill, 1975.

STIGLER, S. M. *The history of statistics*: the measurement of uncertainty before 1900. USA: Harward, 1986.

STEVENSON, W. J. *Estatística aplicada à administração.* São Paulo: Harbra, 1981.

TEXEIRA, E.; MEINERT, E. M.; BARBETTA, P. A. *Análise sensorial de alimentos.* Florianópolis: Editora da UFSC, 1987.

TRIOLA, M. F. *Introdução à Estatística.* 9. ed. Rio de Janeiro: LTC, 2005.

WONNACOTT, T. H.; WONNACOTT, R. J. *Estatística aplicada à Economia e à Administração.* Rio de Janeiro: Livros Técnicos e Científicos, 1981.

Apêndice

Tabela 1 Números aleatórios

59	58	48	36	47	92	85	05	08	65	47	49	10	41	05	10	75	59	75	99	17	28	97	99	75
53	26	21	50	21	37	93	85	52	86	86	22	75	34	37	69	85	25	03	78	50	26	18	25	10
07	02	16	58	67	05	32	93	87	84	31	30	62	78	60	59	90	24	22	07	74	43	43	56	91
92	87	67	56	36	58	58	16	88	16	17	83	52	09	99	86	17	20	95	93	01	46	77	18	11
90	57	05	58	96	84	33	68	15	87	28	18	08	76	89	94	60	94	48	76	92	93	49	13	91
24	26	56	02	33	33	21	75	54	04	96	28	85	78	11	54	01	92	86	36	65	19	45	97	79
20	09	49	50	27	33	86	85	59	39	02	25	60	56	26	01	11	24	44	15	58	00	54	54	09
22	74	50	39	12	83	91	03	38	78	85	56	78	41	44	26	04	12	13	50	38	15	61	02	51
10	45	36	09	86	07	68	31	98	41	98	17	56	93	84	16	01	48	99	36	44	61	71	69	67
09	82	11	18	29	96	19	12	47	26	26	01	14	78	55	33	11	13	56	95	68	66	57	90	33
04	63	02	45	50	61	91	02	14	07	57	36	29	12	74	89	47	84	89	69	13	85	22	66	83
55	93	05	63	30	40	05	51	03	31	68	15	33	85	87	94	80	24	96	62	31	38	95	35	38
66	15	07	64	38	16	44	52	26	42	34	65	99	71	63	87	22	04	62	15	76	94	00	00	77
96	31	72	41	94	47	03	44	73	77	96	17	02	97	50	26	67	60	63	57	66	81	92	03	20
07	10	58	83	63	35	47	34	05	38	92	26	05	33	40	91	23	43	68	72	29	74	60	67	01
04	47	64	02	49	10	52	21	00	80	40	56	68	97	32	43	46	70	65	08	96	52	25	29	44
56	24	53	31	96	65	42	53	27	78	23	30	61	34	18	56	59	23	69	27	83	66	60	03	12
98	15	27	91	71	24	15	28	61	91	83	49	05	82	54	53	59	30	25	19	36	31	31	56	58
36	96	23	77	26	79	74	28	12	16	08	88	07	28	71	45	43	40	07	66	11	26	38	51	87
66	01	53	03	67	92	27	27	17	54	31	23	30	42	83	85	78	21	68	34	86	33	77	84	40
48	07	09	48	65	92	33	41	97	63	48	97	19	86	81	10	85	42	84	49	03	82	01	82	88
95	44	86	84	32	09	03	56	46	96	64	51	33	75	10	29	00	99	23	82	92	31	77	08	17
91	73	15	42	46	72	21	07	34	11	92	70	89	58	54	11	30	93	38	29	00	53	93	14	09
08	35	79	86	83	06	89	37	82	12	81	14	08	82	04	91	88	04	86	36	18	10	09	78	99
37	20	97	09	96	86	34	77	09	31	04	38	18	79	61	68	66	47	40	35	40	16	50	22	54
79	14	72	97	40	90	98	64	42	25	72	95	89	98	59	03	73	02	95	47	34	85	74	60	90
58	55	07	49	26	08	02	70	20	14	57	17	20	89	16	07	86	05	38	61	69	48	78	18	62
77	93	74	07	34	23	49	25	23	87	43	93	35	93	02	80	94	57	16	22	73	67	28	75	37
91	82	56	78	91	47	22	60	09	32	67	02	21	71	61	12	83	08	40	00	52	23	47	46	58
53	66	43	91	44	19	05	53	26	31	89	52	31	98	20	03	70	03	61	07	52	79	97	75	92
91	03	23	35	58	48	22	68	98	07	12	20	88	41	89	19	00	56	88	74	96	71	20	52	46
70	35	43	62	20	81	20	95	72	99	80	91	40	17	51	26	71	79	23	17	01	25	48	07	82
93	85	01	86	56	78	48	74	55	63	62	09	64	35	47	08	70	04	66	86	08	91	83	42	94
75	40	86	33	31	96	06	26	53	07	41	58	96	29	23	17	71	66	60	72	07	18	47	73	75
37	15	68	73	37	31	76	55	39	13	49	61	13	83	90	53	47	54	53	52	80	30	40	35	21
35	88	34	83	04	71	67	75	40	83	99	97	96	83	32	16	04	27	99	31	49	80	34	34	95
73	06	78	79	97	28	86	29	45	91	76	44	64	99	81	33	95	06	94	26	85	78	57	43	12
94	70	05	36	32	38	44	59	60	01	13	74	03	30	33	24	79	77	71	87	41	57	07	96	68
09	65	41	62	93	63	28	60	59	28	29	08	69	81	67	60	57	53	64	28	12	24	35	23	49
12	39	50	50	09	22	70	54	75	38	78	56	79	26	62	79	37	83	33	92	33	30	61	41	90

Nota: Os espaços entre os números são apenas para facilitar a leitura, mas os números podem ser lidos com a quantidade de algarismos que se queira.

Tabela 2 Distribuição binomial: probabilidade de cada valor x em função de n e π

n	x	π									
		0,05	0,1	0,15	0,2	0,25	0,3	0,35	0,4	0,45	0,5
1	0	0,9500	0,9000	0,8500	0,8000	0,7500	0,7000	0,6500	0,6000	0,5500	0,5000
	1	0,0500	0,1000	0,1500	0,2000	0,2500	0,3000	0,3500	0,4000	0,4500	0,5000
2	0	0,9025	0,8100	0,7225	0,6400	0,5625	0,4900	0,4225	0,3600	0,3025	0,2500
	1	0,0950	0,1800	0,2550	0,3200	0,3750	0,4200	0,4550	0,4800	0,4950	0,5000
	2	0,0025	0,0100	0,0225	0,0400	0,0625	0,0900	0,1225	0,1600	0,2025	0,2500
3	0	0,8574	0,7290	0,6141	0,5120	0,4219	0,3430	0,2746	0,2160	0,1664	0,1250
	1	0,1354	0,2430	0,3251	0,3840	0,4219	0,4410	0,4436	0,4320	0,4084	0,3750
	2	0,0071	0,0270	0,0574	0,0960	0,1406	0,1890	0,2389	0,2880	0,3341	0,3750
	3	0,0001	0,0010	0,0034	0,0080	0,0156	0,0270	0,0429	0,0640	0,0911	0,1250
4	0	0,8145	0,6561	0,5220	0,4096	0,3164	0,2401	0,1785	0,1296	0,0915	0,0625
	1	0,1715	0,2916	0,3685	0,4096	0,4219	0,4116	0,3845	0,3456	0,2995	0,2500
	2	0,0135	0,0486	0,0975	0,1536	0,2109	0,2646	0,3105	0,3456	0,3675	0,3750
	3	0,0005	0,0036	0,0115	0,0256	0,0469	0,0756	0,1115	0,1536	0,2005	0,2500
	4	0,0000	0,0001	0,0005	0,0016	0,0039	0,0081	0,0150	0,0256	0,0410	0,0625
5	0	0,7738	0,5905	0,4437	0,3277	0,2373	0,1681	0,1160	0,0778	0,0503	0,0313
	1	0,2036	0,3281	0,3915	0,4096	0,3955	0,3602	0,3124	0,2592	0,2059	0,1563
	2	0,0214	0,0729	0,1382	0,2048	0,2637	0,3087	0,3364	0,3456	0,3369	0,3125
	3	0,0011	0,0081	0,0244	0,0512	0,0879	0,1323	0,1811	0,2304	0,2757	0,3125
	4	0,0000	0,0005	0,0022	0,0064	0,0146	0,0284	0,0488	0,0768	0,1128	0,1563
	5	0,0000	0,0000	0,0001	0,0003	0,0010	0,0024	0,0053	0,0102	0,0185	0,0313
6	0	0,7351	0,5314	0,3771	0,2621	0,1780	0,1176	0,0754	0,0467	0,0277	0,0156
	1	0,2321	0,3543	0,3993	0,3932	0,3560	0,3025	0,2437	0,1866	0,1359	0,0938
	2	0,0305	0,0984	0,1762	0,2458	0,2966	0,3241	0,3280	0,3110	0,2780	0,2344
	3	0,0021	0,0146	0,0415	0,0819	0,1318	0,1852	0,2355	0,2765	0,3032	0,3125
	4	0,0001	0,0012	0,0055	0,0154	0,0330	0,0595	0,0951	0,1382	0,1861	0,2344
	5	0,0000	0,0001	0,0004	0,0015	0,0044	0,0102	0,0205	0,0369	0,0609	0,0938
	6	0,0000	0,0000	0,0000	0,0001	0,0002	0,0007	0,0018	0,0041	0,0083	0,0156
7	0	0,6983	0,4783	0,3206	0,2097	0,1335	0,0824	0,0490	0,0280	0,0152	0,0078
	1	0,2573	0,3720	0,3960	0,3670	0,3115	0,2471	0,1848	0,1306	0,0872	0,0547
	2	0,0406	0,1240	0,2097	0,2753	0,3115	0,3177	0,2985	0,2613	0,2140	0,1641
	3	0,0036	0,0230	0,0617	0,1147	0,1730	0,2269	0,2679	0,2903	0,2918	0,2734
	4	0,0002	0,0026	0,0109	0,0287	0,0577	0,0972	0,1442	0,1935	0,2388	0,2734
	5	0,0000	0,0002	0,0012	0,0043	0,0115	0,0250	0,0466	0,0774	0,1172	0,1641
	6	0,0000	0,0000	0,0001	0,0004	0,0013	0,0036	0,0084	0,0172	0,0320	0,0547
	7	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0006	0,0016	0,0037	0,0078
8	0	0,6634	0,4305	0,2725	0,1678	0,1001	0,0576	0,0319	0,0168	0,0084	0,0039
	1	0,2793	0,3826	0,3847	0,3355	0,2670	0,1977	0,1373	0,0896	0,0548	0,0313
	2	0,0515	0,1488	0,2376	0,2936	0,3115	0,2965	0,2587	0,2090	0,1569	0,1094
	3	0,0054	0,0331	0,0839	0,1468	0,2076	0,2541	0,2786	0,2787	0,2568	0,2188
	4	0,0004	0,0046	0,0185	0,0459	0,0865	0,1361	0,1875	0,2322	0,2627	0,2734
	5	0,0000	0,0004	0,0026	0,0092	0,0231	0,0467	0,0808	0,1239	0,1719	0,2188
	6	0,0000	0,0000	0,0002	0,0011	0,0038	0,0100	0,0217	0,0413	0,0703	0,1094
	7	0,0000	0,0000	0,0000	0,0001	0,0004	0,0012	0,0033	0,0079	0,0164	0,0313
	8	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0007	0,0017	0,0039

Tabela 2 Distribuição binomial: probabilidade de cada valor x em função de n e π (continuação)

n	x	π								
		0,55	0,6	0,65	0,7	0,75	0,8	0,85	0,9	0,95
1	0	0,4500	0,4000	0,3500	0,3000	0,2500	0,2000	0,1500	0,1000	0,0500
	1	0,5500	0,6000	0,6500	0,7000	0,7500	0,8000	0,8500	0,9000	0,9500
2	0	0,2025	0,1600	0,1225	0,0900	0,0625	0,0400	0,0225	0,0100	0,0025
	1	0,4950	0,4800	0,4550	0,4200	0,3750	0,3200	0,2550	0,1800	0,0950
	2	0,3025	0,3600	0,4225	0,4900	0,5625	0,6400	0,7225	0,8100	0,9025
3	0	0,0911	0,0640	0,0429	0,0270	0,0156	0,0080	0,0034	0,0010	0,0001
	1	0,3341	0,2880	0,2389	0,1890	0,1406	0,0960	0,0574	0,0270	0,0071
	2	0,4084	0,4320	0,4436	0,4410	0,4219	0,3840	0,3251	0,2430	0,1354
	3	0,1664	0,2160	0,2746	0,3430	0,4219	0,5120	0,6141	0,7290	0,8574
4	0	0,0410	0,0256	0,0150	0,0081	0,0039	0,0016	0,0005	0,0001	0,0000
	1	0,2005	0,1536	0,1115	0,0756	0,0469	0,0256	0,0115	0,0036	0,0005
	2	0,3675	0,3456	0,3105	0,2646	0,2109	0,1536	0,0975	0,0486	0,0135
	3	0,2995	0,3456	0,3845	0,4116	0,4219	0,4096	0,3685	0,2916	0,1715
	4	0,0915	0,1296	0,1785	0,2401	0,3164	0,4096	0,5220	0,6561	0,8145
5	0	0,0185	0,0102	0,0053	0,0024	0,0010	0,0003	0,0001	0,0000	0,0000
	1	0,1128	0,0768	0,0488	0,0284	0,0146	0,0064	0,0022	0,0005	0,0000
	2	0,2757	0,2304	0,1811	0,1323	0,0879	0,0512	0,0244	0,0081	0,0011
	3	0,3369	0,3456	0,3364	0,3087	0,2637	0,2048	0,1382	0,0729	0,0214
	4	0,2059	0,2592	0,3124	0,3602	0,3955	0,4096	0,3915	0,3281	0,2036
	5	0,0503	0,0778	0,1160	0,1681	0,2373	0,3277	0,4437	0,5905	0,7738
6	0	0,0083	0,0041	0,0018	0,0007	0,0002	0,0001	0,0000	0,0000	0,0000
	1	0,0609	0,0369	0,0205	0,0102	0,0044	0,0015	0,0004	0,0001	0,0000
	2	0,1861	0,1382	0,0951	0,0595	0,0330	0,0154	0,0055	0,0012	0,0001
	3	0,3032	0,2765	0,2355	0,1852	0,1318	0,0819	0,0415	0,0146	0,0021
	4	0,2780	0,3110	0,3280	0,3241	0,2966	0,2458	0,1762	0,0984	0,0305
	5	0,1359	0,1866	0,2437	0,3025	0,3560	0,3932	0,3993	0,3543	0,2321
	6	0,0277	0,0467	0,0754	0,1176	0,1780	0,2621	0,3771	0,5314	0,7351
7	0	0,0037	0,0016	0,0006	0,0002	0,0001	0,0000	0,0000	0,0000	0,0000
	1	0,0320	0,0172	0,0084	0,0036	0,0013	0,0004	0,0001	0,0000	0,0000
	2	0,1172	0,0774	0,0466	0,0250	0,0115	0,0043	0,0012	0,0002	0,0000
	3	0,2388	0,1935	0,1442	0,0972	0,0577	0,0287	0,0109	0,0026	0,0002
	4	0,2918	0,2903	0,2679	0,2269	0,1730	0,1147	0,0617	0,0230	0,0036
	5	0,2140	0,2613	0,2985	0,3177	0,3115	0,2753	0,2097	0,1240	0,0406
	6	0,0872	0,1306	0,1848	0,2471	0,3115	0,3670	0,3960	0,3720	0,2573
	7	0,0152	0,0280	0,0490	0,0824	0,1335	0,2097	0,3206	0,4783	0,6983
8	0	0,0017	0,0007	0,0002	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000
	1	0,0164	0,0079	0,0033	0,0012	0,0004	0,0001	0,0000	0,0000	0,0000
	2	0,0703	0,0413	0,0217	0,0100	0,0038	0,0011	0,0002	0,0000	0,0000
	3	0,1719	0,1239	0,0808	0,0467	0,0231	0,0092	0,0026	0,0004	0,0000
	4	0,2627	0,2322	0,1875	0,1361	0,0865	0,0459	0,0185	0,0046	0,0004
	5	0,2568	0,2787	0,2786	0,2541	0,2076	0,1468	0,0839	0,0331	0,0054
	6	0,1569	0,2090	0,2587	0,2965	0,3115	0,2936	0,2376	0,1488	0,0515
	7	0,0548	0,0896	0,1373	0,1977	0,2670	0,3355	0,3847	0,3826	0,2793
	8	0,0084	0,0168	0,0319	0,0576	0,1001	0,1678	0,2725	0,4305	0,6634

Tabela 2 Distribuição binomial: probabilidade de cada valor x em função de n e π (continuação)

n	x	π									
		0,05	0,1	0,15	0,2	0,25	0,3	0,35	0,4	0,45	0,5
9	0	0,6302	0,3874	0,2316	0,1342	0,0751	0,0404	0,0207	0,0101	0,0046	0,0020
	1	0,2985	0,3874	0,3679	0,3020	0,2253	0,1556	0,1004	0,0605	0,0339	0,0176
	2	0,0629	0,1722	0,2597	0,3020	0,3003	0,2668	0,2162	0,1612	0,1110	0,0703
	3	0,0077	0,0446	0,1069	0,1762	0,2336	0,2668	0,2716	0,2508	0,2119	0,1641
	4	0,0006	0,0074	0,0283	0,0661	0,1168	0,1715	0,2194	0,2508	0,2600	0,2461
	5	0,0000	0,0008	0,0050	0,0165	0,0389	0,0735	0,1181	0,1672	0,2128	0,2461
	6	0,0000	0,0001	0,0006	0,0028	0,0087	0,0210	0,0424	0,0743	0,1160	0,1641
	7	0,0000	0,0000	0,0000	0,0003	0,0012	0,0039	0,0098	0,0212	0,0407	0,0703
	8	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0013	0,0035	0,0083	0,0176
	9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0008	0,0020
10	0	0,5987	0,3487	0,1969	0,1074	0,0563	0,0282	0,0135	0,0060	0,0025	0,0010
	1	0,3151	0,3874	0,3474	0,2684	0,1877	0,1211	0,0725	0,0403	0,0207	0,0098
	2	0,0746	0,1937	0,2759	0,3020	0,2816	0,2335	0,1757	0,1209	0,0763	0,0439
	3	0,0105	0,0574	0,1298	0,2013	0,2503	0,2668	0,2522	0,2150	0,1665	0,1172
	4	0,0010	0,0112	0,0401	0,0881	0,1460	0,2001	0,2377	0,2508	0,2384	0,2051
	5	0,0001	0,0015	0,0085	0,0264	0,0584	0,1029	0,1536	0,2007	0,2340	0,2461
	6	0,0000	0,0001	0,0012	0,0055	0,0162	0,0368	0,0689	0,1115	0,1596	0,2051
	7	0,0000	0,0000	0,0001	0,0008	0,0031	0,0090	0,0212	0,0425	0,0746	0,1172
	8	0,0000	0,0000	0,0000	0,0001	0,0004	0,0014	0,0043	0,0106	0,0229	0,0439
	9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0016	0,0042	0,0098
	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010
11	0	0,5688	0,3138	0,1673	0,0859	0,0422	0,0198	0,0088	0,0036	0,0014	0,0005
	1	0,3293	0,3835	0,3248	0,2362	0,1549	0,0932	0,0518	0,0266	0,0125	0,0054
	2	0,0867	0,2131	0,2866	0,2953	0,2581	0,1998	0,1395	0,0887	0,0513	0,0269
	3	0,0137	0,0710	0,1517	0,2215	0,2581	0,2568	0,2254	0,1774	0,1259	0,0806
	4	0,0014	0,0158	0,0536	0,1107	0,1721	0,2201	0,2428	0,2365	0,2060	0,1611
	5	0,0001	0,0025	0,0132	0,0388	0,0803	0,1321	0,1830	0,2207	0,2360	0,2256
	6	0,0000	0,0003	0,0023	0,0097	0,0268	0,0566	0,0985	0,1471	0,1931	0,2256
	7	0,0000	0,0000	0,0003	0,0017	0,0064	0,0173	0,0379	0,0701	0,1128	0,1611
	8	0,0000	0,0000	0,0000	0,0002	0,0011	0,0037	0,0102	0,0234	0,0462	0,0806
	9	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0018	0,0052	0,0126	0,0269
	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0007	0,0021	0,0054
	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0005
12	0	0,5404	0,2824	0,1422	0,0687	0,0317	0,0138	0,0057	0,0022	0,0008	0,0002
	1	0,3413	0,3766	0,3012	0,2062	0,1267	0,0712	0,0368	0,0174	0,0075	0,0029
	2	0,0988	0,2301	0,2924	0,2835	0,2323	0,1678	0,1088	0,0639	0,0339	0,0161
	3	0,0173	0,0852	0,1720	0,2362	0,2581	0,2397	0,1954	0,1419	0,0923	0,0537
	4	0,0021	0,0213	0,0683	0,1329	0,1936	0,2311	0,2367	0,2128	0,1700	0,1208
	5	0,0002	0,0038	0,0193	0,0532	0,1032	0,1585	0,2039	0,2270	0,2225	0,1934
	6	0,0000	0,0005	0,0040	0,0155	0,0401	0,0792	0,1281	0,1766	0,2124	0,2256
	7	0,0000	0,0000	0,0006	0,0033	0,0115	0,0291	0,0591	0,1009	0,1489	0,1934
	8	0,0000	0,0000	0,0001	0,0005	0,0024	0,0078	0,0199	0,0420	0,0762	0,1208
	9	0,0000	0,0000	0,0000	0,0001	0,0004	0,0015	0,0048	0,0125	0,0277	0,0537
	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0008	0,0025	0,0068	0,0161
	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010	0,0029
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002

Tabela 2 Distribuição binomial: probabilidade de cada valor x em função de n e π (continuação)

n	x	π								
		0,55	0,6	0,65	0,7	0,75	0,8	0,85	0,9	0,95
9	0	0,0008	0,0003	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	1	0,0083	0,0035	0,0013	0,0004	0,0001	0,0000	0,0000	0,0000	0,0000
	2	0,0407	0,0212	0,0098	0,0039	0,0012	0,0003	0,0000	0,0000	0,0000
	3	0,1160	0,0743	0,0424	0,0210	0,0087	0,0028	0,0006	0,0001	0,0000
	4	0,2128	0,1672	0,1181	0,0735	0,0389	0,0165	0,0050	0,0008	0,0000
	5	0,2600	0,2508	0,2194	0,1715	0,1168	0,0661	0,0283	0,0074	0,0006
	6	0,2119	0,2508	0,2716	0,2668	0,2336	0,1762	0,1069	0,0446	0,0077
	7	0,1110	0,1612	0,2162	0,2668	0,3003	0,3020	0,2597	0,1722	0,0629
	8	0,0339	0,0605	0,1004	0,1556	0,2253	0,3020	0,3679	0,3874	0,2985
	9	0,0046	0,0101	0,0207	0,0404	0,0751	0,1342	0,2316	0,3874	0,6302
10	0	0,0003	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	1	0,0042	0,0016	0,0005	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000
	2	0,0229	0,0106	0,0043	0,0014	0,0004	0,0001	0,0000	0,0000	0,0000
	3	0,0746	0,0425	0,0212	0,0090	0,0031	0,0008	0,0001	0,0000	0,0000
	4	0,1596	0,1115	0,0689	0,0368	0,0162	0,0055	0,0012	0,0001	0,0000
	5	0,2340	0,2007	0,1536	0,1029	0,0584	0,0264	0,0085	0,0015	0,0001
	6	0,2384	0,2508	0,2377	0,2001	0,1460	0,0881	0,0401	0,0112	0,0010
	7	0,1665	0,2150	0,2522	0,2668	0,2503	0,2013	0,1298	0,0574	0,0105
	8	0,0763	0,1209	0,1757	0,2335	0,2816	0,3020	0,2759	0,1937	0,0746
	9	0,0207	0,0403	0,0725	0,1211	0,1877	0,2684	0,3474	0,3874	0,3151
	10	0,0025	0,0060	0,0135	0,0282	0,0563	0,1074	0,1969	0,3487	0,5987
11	0	0,0002	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	1	0,0021	0,0007	0,0002	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	2	0,0126	0,0052	0,0018	0,0005	0,0001	0,0000	0,0000	0,0000	0,0000
	3	0,0462	0,0234	0,0102	0,0037	0,0011	0,0002	0,0000	0,0000	0,0000
	4	0,1128	0,0701	0,0379	0,0173	0,0064	0,0017	0,0003	0,0000	0,0000
	5	0,1931	0,1471	0,0985	0,0566	0,0268	0,0097	0,0023	0,0003	0,0000
	6	0,2360	0,2207	0,1830	0,1321	0,0803	0,0388	0,0132	0,0025	0,0001
	7	0,2060	0,2365	0,2428	0,2201	0,1721	0,1107	0,0536	0,0158	0,0014
	8	0,1259	0,1774	0,2254	0,2568	0,2581	0,2215	0,1517	0,0710	0,0137
	9	0,0513	0,0887	0,1395	0,1998	0,2581	0,2953	0,2866	0,2131	0,0867
	10	0,0125	0,0266	0,0518	0,0932	0,1549	0,2362	0,3248	0,3835	0,3293
	11	0,0014	0,0036	0,0088	0,0198	0,0422	0,0859	0,1673	0,3138	0,5688
12	0	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	1	0,0010	0,0003	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	2	0,0068	0,0025	0,0008	0,0002	0,0000	0,0000	0,0000	0,0000	0,0000
	3	0,0277	0,0125	0,0048	0,0015	0,0004	0,0001	0,0000	0,0000	0,0000
	4	0,0762	0,0420	0,0199	0,0078	0,0024	0,0005	0,0001	0,0000	0,0000
	5	0,1489	0,1009	0,0591	0,0291	0,0115	0,0033	0,0006	0,0000	0,0000
	6	0,2124	0,1766	0,1281	0,0792	0,0401	0,0155	0,0040	0,0005	0,0000
	7	0,2225	0,2270	0,2039	0,1585	0,1032	0,0532	0,0193	0,0038	0,0002
	8	0,1700	0,2128	0,2367	0,2311	0,1936	0,1329	0,0683	0,0213	0,0021
	9	0,0923	0,1419	0,1954	0,2397	0,2581	0,2362	0,1720	0,0852	0,0173
	10	0,0339	0,0639	0,1088	0,1678	0,2323	0,2835	0,2924	0,2301	0,0988
	11	0,0075	0,0174	0,0368	0,0712	0,1267	0,2062	0,3012	0,3766	0,3413
	12	0,0008	0,0022	0,0057	0,0138	0,0317	0,0687	0,1422	0,2824	0,5404

Tabela 2 Distribuição binomial: probabilidade de cada valor x em função de n e π (continuação)

n	x	π									
		0,05	0,1	0,15	0,2	0,25	0,3	0,35	0,4	0,45	0,5
13	0	0,5133	0,2542	0,1209	0,0550	0,0238	0,0097	0,0037	0,0013	0,0004	0,0001
	1	0,3512	0,3672	0,2774	0,1787	0,1029	0,0540	0,0259	0,0113	0,0045	0,0016
	2	0,1109	0,2448	0,2937	0,2680	0,2059	0,1388	0,0836	0,0453	0,0220	0,0095
	3	0,0214	0,0997	0,1900	0,2457	0,2517	0,2181	0,1651	0,1107	0,0660	0,0349
	4	0,0028	0,0277	0,0838	0,1535	0,2097	0,2337	0,2222	0,1845	0,1350	0,0873
	5	0,0003	0,0055	0,0266	0,0691	0,1258	0,1803	0,2154	0,2214	0,1989	0,1571
	6	0,0000	0,0008	0,0063	0,0230	0,0559	0,1030	0,1546	0,1968	0,2169	0,2095
	7	0,0000	0,0001	0,0011	0,0058	0,0186	0,0442	0,0833	0,1312	0,1775	0,2095
	8	0,0000	0,0000	0,0001	0,0011	0,0047	0,0142	0,0336	0,0656	0,1089	0,1571
	9	0,0000	0,0000	0,0000	0,0001	0,0009	0,0034	0,0101	0,0243	0,0495	0,0873
	10	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006	0,0022	0,0065	0,0162	0,0349
	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0012	0,0036	0,0095
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0016
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
14	0	0,4877	0,2288	0,1028	0,0440	0,0178	0,0068	0,0024	0,0008	0,0002	0,0001
	1	0,3593	0,3559	0,2539	0,1539	0,0832	0,0407	0,0181	0,0073	0,0027	0,0009
	2	0,1229	0,2570	0,2912	0,2501	0,1802	0,1134	0,0634	0,0317	0,0141	0,0056
	3	0,0259	0,1142	0,2056	0,2501	0,2402	0,1943	0,1366	0,0845	0,0462	0,0222
	4	0,0037	0,0349	0,0998	0,1720	0,2202	0,2290	0,2022	0,1549	0,1040	0,0611
	5	0,0004	0,0078	0,0352	0,0860	0,1468	0,1963	0,2178	0,2066	0,1701	0,1222
	6	0,0000	0,0013	0,0093	0,0322	0,0734	0,1262	0,1759	0,2066	0,2088	0,1833
	7	0,0000	0,0002	0,0019	0,0092	0,0280	0,0618	0,1082	0,1574	0,1952	0,2095
	8	0,0000	0,0000	0,0003	0,0020	0,0082	0,0232	0,0510	0,0918	0,1398	0,1833
	9	0,0000	0,0000	0,0000	0,0003	0,0018	0,0066	0,0183	0,0408	0,0762	0,1222
	10	0,0000	0,0000	0,0000	0,0000	0,0003	0,0014	0,0049	0,0136	0,0312	0,0611
	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0010	0,0033	0,0093	0,0222
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0019	0,0056
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0009
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
15	0	0,4633	0,2059	0,0874	0,0352	0,0134	0,0047	0,0016	0,0005	0,0001	0,0000
	1	0,3658	0,3432	0,2312	0,1319	0,0668	0,0305	0,0126	0,0047	0,0016	0,0005
	2	0,1348	0,2669	0,2856	0,2309	0,1559	0,0916	0,0476	0,0219	0,0090	0,0032
	3	0,0307	0,1285	0,2184	0,2501	0,2252	0,1700	0,1110	0,0634	0,0318	0,0139
	4	0,0049	0,0428	0,1156	0,1876	0,2252	0,2186	0,1792	0,1268	0,0780	0,0417
	5	0,0006	0,0105	0,0449	0,1032	0,1651	0,2061	0,2123	0,1859	0,1404	0,0916
	6	0,0000	0,0019	0,0132	0,0430	0,0917	0,1472	0,1906	0,2066	0,1914	0,1527
	7	0,0000	0,0003	0,0030	0,0138	0,0393	0,0811	0,1319	0,1771	0,2013	0,1964
	8	0,0000	0,0000	0,0005	0,0035	0,0131	0,0348	0,0710	0,1181	0,1647	0,1964
	9	0,0000	0,0000	0,0001	0,0007	0,0034	0,0116	0,0298	0,0612	0,1048	0,1527
	10	0,0000	0,0000	0,0000	0,0001	0,0007	0,0030	0,0096	0,0245	0,0515	0,0916
	11	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006	0,0024	0,0074	0,0191	0,0417
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0016	0,0052	0,0139
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010	0,0032
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

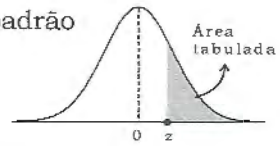
Tabela 2 Distribuição binomial: probabilidade de cada valor x em função de n e π (continuação)

n	x	π								
		0,55	0,6	0,65	0,7	0,75	0,8	0,85	0,9	0,95
13	0	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	1	0,0005	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	2	0,0036	0,0012	0,0003	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000
	3	0,0162	0,0065	0,0022	0,0006	0,0001	0,0000	0,0000	0,0000	0,0000
	4	0,0495	0,0243	0,0101	0,0034	0,0009	0,0001	0,0000	0,0000	0,0000
	5	0,1089	0,0656	0,0336	0,0142	0,0047	0,0011	0,0001	0,0000	0,0000
	6	0,1775	0,1312	0,0833	0,0442	0,0186	0,0058	0,0011	0,0001	0,0000
	7	0,2169	0,1968	0,1546	0,1030	0,0559	0,0230	0,0063	0,0008	0,0000
	8	0,1989	0,2214	0,2154	0,1803	0,1258	0,0691	0,0266	0,0055	0,0003
	9	0,1350	0,1845	0,2222	0,2337	0,2097	0,1535	0,0838	0,0277	0,0028
	10	0,0660	0,1107	0,1651	0,2181	0,2517	0,2457	0,1900	0,0997	0,0214
	11	0,0220	0,0453	0,0836	0,1388	0,2059	0,2680	0,2937	0,2448	0,1109
	12	0,0045	0,0113	0,0259	0,0540	0,1029	0,1787	0,2774	0,3672	0,3512
	13	0,0004	0,0013	0,0037	0,0097	0,0238	0,0550	0,1209	0,2542	0,5133
14	0	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	1	0,0002	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	2	0,0019	0,0005	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	3	0,0093	0,0033	0,0010	0,0002	0,0000	0,0000	0,0000	0,0000	0,0000
	4	0,0312	0,0136	0,0049	0,0014	0,0003	0,0000	0,0000	0,0000	0,0000
	5	0,0762	0,0408	0,0183	0,0066	0,0018	0,0003	0,0000	0,0000	0,0000
	6	0,1398	0,0918	0,0510	0,0232	0,0082	0,0020	0,0003	0,0000	0,0000
	7	0,1952	0,1574	0,1082	0,0618	0,0280	0,0092	0,0019	0,0002	0,0000
	8	0,2088	0,2066	0,1759	0,1262	0,0734	0,0322	0,0093	0,0013	0,0000
	9	0,1701	0,2066	0,2178	0,1963	0,1468	0,0860	0,0352	0,0078	0,0004
	10	0,1040	0,1549	0,2022	0,2290	0,2202	0,1720	0,0998	0,0349	0,0037
	11	0,0462	0,0845	0,1366	0,1943	0,2402	0,2501	0,2056	0,1142	0,0259
	12	0,0141	0,0317	0,0634	0,1134	0,1802	0,2501	0,2912	0,2570	0,1229
	13	0,0027	0,0073	0,0181	0,0407	0,0832	0,1539	0,2539	0,3559	0,3593
	14	0,0002	0,0008	0,0024	0,0068	0,0178	0,0440	0,1028	0,2288	0,4877
15	0	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	1	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	2	0,0010	0,0003	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	3	0,0052	0,0016	0,0004	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000
	4	0,0191	0,0074	0,0024	0,0006	0,0001	0,0000	0,0000	0,0000	0,0000
	5	0,0515	0,0245	0,0096	0,0030	0,0007	0,0001	0,0000	0,0000	0,0000
	6	0,1048	0,0612	0,0298	0,0116	0,0034	0,0007	0,0001	0,0000	0,0000
	7	0,1647	0,1181	0,0710	0,0348	0,0131	0,0035	0,0005	0,0000	0,0000
	8	0,2013	0,1771	0,1319	0,0811	0,0393	0,0138	0,0030	0,0003	0,0000
	9	0,1914	0,2066	0,1906	0,1472	0,0917	0,0430	0,0132	0,0019	0,0000
	10	0,1404	0,1859	0,2123	0,2061	0,1651	0,1032	0,0449	0,0105	0,0006
	11	0,0780	0,1268	0,1792	0,2186	0,2252	0,1876	0,1156	0,0428	0,0049
	12	0,0318	0,0634	0,1110	0,1700	0,2252	0,2501	0,2184	0,1285	0,0307
	13	0,0090	0,0219	0,0476	0,0916	0,1559	0,2309	0,2856	0,2669	0,1348
	14	0,0016	0,0047	0,0126	0,0305	0,0668	0,1319	0,2312	0,3432	0,3658
	15	0,0001	0,0005	0,0016	0,0047	0,0134	0,0352	0,0874	0,2059	0,4633

Tabela 3 Coeficientes binomiais

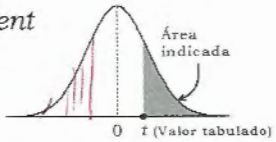
n	$\binom{n}{0}$	$\binom{n}{1}$	$\binom{n}{2}$	$\binom{n}{3}$	$\binom{n}{4}$	$\binom{n}{5}$	$\binom{n}{6}$	$\binom{n}{7}$	$\binom{n}{8}$	$\binom{n}{9}$	$\binom{n}{10}$
0	1										
1	1	1									
2	1	2	1								
3	1	3	3	1							
4	1	4	6	4	1						
5	1	5	10	10	5	1					
6	1	6	15	20	15	6	1				
7	1	7	21	35	35	21	7	1			
8	1	8	28	56	70	56	28	8	1		
9	1	9	36	84	126	126	84	36	9	1	
10	1	10	45	120	210	252	210	120	45	10	1
11	1	11	55	165	330	462	462	330	165	55	11
12	1	12	66	220	495	792	924	792	495	220	66
13	1	13	78	286	715	1287	1716	1716	1287	715	286
14	1	14	91	364	1001	2002	3003	3432	3003	2002	1001
15	1	15	105	455	1365	3003	5005	6435	6435	5005	3003
16	1	16	120	560	1820	4368	8008	11440	12870	11440	8008
17	1	17	136	680	2380	6188	12376	19448	24310	24310	19448
18	1	18	153	816	3060	8568	18564	31824	43758	48620	43758
19	1	19	171	969	3876	11628	27132	50338	75582	92378	92378
20	1	20	190	1140	4845	15504	38760	77520	125970	167960	184756

Tabela 4 Distribuição normal padrão



	segunda decimal de **z**									
z	0	1	2	3	4	5	6	7	8	9
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2842	0,2810	0,2776
0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0722	0,0708	0,0694	0,0681
1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
1,8	0,0359	0,0352	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
2,9	0,0019	0,0018	0,0017	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
3,0	0,00135									
3,5	0,000 233									
4,0	0,000 031 7									
4,5	0,000 003 40									
5,0	0,000 000 287									

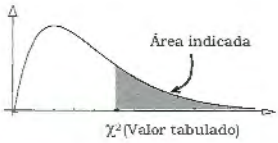
Tabela 5 Distribuição ***t*** de *Student*



	Área na cauda superior								
gl	0,25	0,10	0,05	0,025	0,01	0,005	0,0025	0,001	0,0005
1	1,000	3,078	6,314	12,71	31,82	63,66	127,3	318,3	636,6
2	0,816	1,886	2,920	4,303	6,965	9,925	14,09	22,33	31,60
3	0,765	1,638	2,353	3,182	4,541	5,841	7,453	10,21	12,92
4	0,741	1,533	2,132	2,776	3,747	4,604	5,598	7,173	8,610
5	0,727	1,476	2,015	2,571	3,365	4,032	4,773	5,894	6,869
6	0,718	1,440	1,943	2,447	3,143	3,707	4,317	5,208	5,959
7	0,711	1,415	1,895	2,365	2,998	3,499	4,029	4,785	5,408
8	0,706	1,397	1,860	2,306	2,896	3,355	3,833	4,501	5,041
9	0,703	1,383	1,833	2,262	2,821	3,250	3,690	4,297	4,781
10	0,700	1,372	1,812	2,228	2,764	3,169	3,581	4,144	4,587
11	0,697	1,363	1,796	2,201	2,718	3,106	3,497	4,025	4,437
12	0,695	1,356	1,782	2,179	2,681	3,055	3,428	3,930	4,318
13	0,694	1,350	1,771	2,160	2,650	3,012	3,372	3,852	4,221
14	0,692	1,345	1,761	2,145	2,624	2,977	3,326	3,787	4,140
15	0,691	1,341	1,753	2,131	2,602	2,947	3,286	3,733	4,073
16	0,690	1,337	1,746	2,120	2,583	2,921	3,252	3,686	4,015
17	0,689	1,333	1,740	2,110	2,567	2,898	3,222	3,646	3,965
18	0,688	1,330	1,734	2,101	2,552	2,878	3,197	3,610	3,922
19	0,688	1,328	1,729	2,093	2,539	2,861	3,174	3,579	3,883
20	0,687	1,325	1,725	2,086	2,528	2,845	3,153	3,552	3,850
21	0,686	1,323	1,721	2,080	2,518	2,831	3,135	3,527	3,819
22	0,686	1,321	1,717	2,074	2,508	2,819	3,119	3,505	3,792
23	0,685	1,319	1,714	2,069	2,500	2,807	3,104	3,485	3,768
24	0,685	1,318	1,711	2,064	2,492	2,797	3,091	3,467	3,745
25	0,684	1,316	1,708	2,060	2,485	2,787	3,078	3,450	3,725
26	0,684	1,315	1,706	2,056	2,479	2,779	3,067	3,435	3,707
27	0,684	1,314	1,703	2,052	2,473	2,771	3,057	3,421	3,689
28	0,683	1,313	1,701	2,048	2,467	2,763	3,047	3,408	3,674
29	0,683	1,311	1,699	2,045	2,462	2,756	3,038	3,396	3,660
30	0,683	1,310	1,697	2,042	2,457	2,750	3,030	3,385	3,646
35	0,682	1,306	1,690	2,030	2,438	2,724	2,996	3,340	3,591
40	0,681	1,303	1,684	2,021	2,423	2,704	2,971	3,307	3,551
45	0,680	1,301	1,679	2,014	2,412	2,690	2,952	3,281	3,520
50	0,679	1,299	1,676	2,009	2,403	2,678	2,937	3,261	3,496
z	0,674	1,282	1,645	1,960	2,326	2,576	2,807	3,090	3,291

Nota: A coluna em destaque é a mais usada.

Tabela 6 Distribuição qui-quadrado



	Área na cauda superior								
gl	0,25	0,10	0,05	0,025	0,01	0,005	0,0025	0,001	0,0005
1	1,32	2,71	3,84	5,02	6,63	7,88	9,14	10,83	12,12
2	2,77	4,61	5,99	7,38	9,21	10,60	11,98	13,82	15,20
3	4,11	6,25	7,81	9,35	11,34	12,84	14,32	16,27	17,73
4	5,39	7,78	9,49	11,14	13,28	14,86	16,42	18,47	20,00
5	6,63	9,24	11,07	12,83	15,09	16,75	18,39	20,51	22,11
6	7,84	10,64	12,59	14,45	16,81	18,55	20,25	22,46	24,10
7	9,04	12,02	14,07	16,01	18,48	20,28	22,04	24,32	26,02
8	10,22	13,36	15,51	17,53	20,09	21,95	23,77	26,12	27,87
9	11,39	14,68	16,92	19,02	21,67	23,59	25,46	27,88	29,67
10	12,55	15,99	18,31	20,48	23,21	25,19	27,11	29,59	31,42
11	13,70	17,28	19,68	21,92	24,73	26,76	28,73	31,26	33,14
12	14,85	18,55	21,03	23,34	26,22	28,30	30,32	32,91	34,82
13	15,98	19,81	22,36	24,74	27,69	29,82	31,88	34,53	36,48
14	17,12	21,06	23,68	26,12	29,14	31,32	33,43	36,12	38,11
15	18,25	22,31	25,00	27,49	30,58	32,80	34,95	37,70	39,72
16	19,37	23,54	26,30	28,85	32,00	34,27	36,46	39,25	41,31
17	20,49	24,77	27,59	30,19	33,41	35,72	37,95	40,79	42,88
18	21,60	25,99	28,87	31,53	34,81	37,16	39,42	42,31	44,43
19	22,72	27,20	30,14	32,85	36,19	38,58	40,88	43,82	45,97
20	23,83	28,41	31,41	34,17	37,57	40,00	42,34	45,31	47,50
21	24,93	29,62	32,67	35,48	38,93	41,40	43,77	46,80	49,01
22	26,04	30,81	33,92	36,78	40,29	42,80	45,20	48,27	50,51
23	27,14	32,01	35,17	38,08	41,64	44,18	46,62	49,73	52,00
24	28,24	33,20	36,42	39,36	42,98	45,56	48,03	51,18	53,48
25	29,34	34,38	37,65	40,65	44,31	46,93	49,44	52,62	54,95
26	30,43	35,56	38,89	41,92	45,64	48,29	50,83	54,05	56,41
27	31,53	36,74	40,11	43,19	46,96	49,65	52,22	55,48	57,86
28	32,62	37,92	41,34	44,46	48,28	50,99	53,59	56,89	59,30
29	33,71	39,09	42,56	45,72	49,59	52,34	54,97	58,30	60,73
30	34,80	40,26	43,77	46,98	50,89	53,67	56,33	59,70	62,16
35	40,22	46,06	49,80	53,20	57,34	60,27	63,08	66,62	69,20
40	45,62	51,81	55,76	59,34	63,69	66,77	69,70	73,40	76,10
45	50,98	57,51	61,66	65,41	69,96	73,17	76,22	80,08	82,87
50	56,33	63,17	67,50	71,42	76,15	79,49	82,66	86,66	89,56
100	109,1	118,5	124,3	129,6	135,8	140,2	144,3	149,4	153,2

Nota: A coluna em destaque é a mais usada.

Tabela 7 Valor absoluto mínimo para o coeficiente de correlação ***r*** de Pearson ser significativo

	Nível de significância, α, num teste unilateral					
	0,100	**0,050**	**0,025**	**0,010**	**0,005**	**0,001**
	Nível de significância, α, num teste bilateral					
n	**0,200**	**0,100**	**0,050**	**0,020**	**0,010**	**0,002**
5	0,687	0,805	0,878	0,934	0,959	0,986
6	0,608	0,729	0,811	0,882	0,917	0,963
7	0,551	0,669	0,754	0,833	0,875	0,935
8	0,507	0,621	0,707	0,789	0,834	0,905
9	0,472	0,582	0,666	0,750	0,798	0,875
10	0,443	0,549	0,632	0,715	0,765	0,847
11	0,419	0,521	0,602	0,685	0,735	0,820
12	0,398	0,497	0,576	0,658	0,708	0,795
13	0,380	0,476	0,553	0,634	0,684	0,772
14	0,365	0,458	0,532	0,612	0,661	0,750
15	0,351	0,441	0,514	0,592	0,641	0,730
16	0,338	0,426	0,497	0,574	0,623	0,711
17	0,327	0,412	0,482	0,558	0,606	0,694
18	0,317	0,400	0,468	0,543	0,590	0,678
19	0,308	0,389	0,456	0,529	0,575	0,662
20	0,299	0,378	0,444	0,516	0,561	0,648
21	0,291	0,369	0,433	0,503	0,549	0,635
22	0,284	0,360	0,423	0,492	0,537	0,622
23	0,277	0,352	0,413	0,482	0,526	0,610
24	0,271	0,344	0,404	0,472	0,515	0,599
25	0,265	0,337	0,396	0,462	0,505	0,588
26	0,260	0,330	0,388	0,453	0,496	0,578
27	0,255	0,323	0,381	0,445	0,487	0,568
28	0,250	0,317	0,374	0,437	0,479	0,559
29	0,245	0,311	0,367	0,430	0,471	0,550
30	0,241	0,306	0,361	0,423	0,463	0,541
35	0,222	0,283	0,334	0,392	0,430	0,504
40	0,207	0,264	0,312	0,367	0,403	0,474
45	0,195	0,248	0,294	0,346	0,380	0,449
50	0,184	0,235	0,279	0,328	0,361	0,427
60	0,168	0,214	0,254	0,300	0,330	0,391
70	0,155	0,198	0,235	0,278	0,306	0,363
80	0,145	0,185	0,220	0,260	0,286	0,340
90	0,136	0,174	0,207	0,245	0,270	0,322
100	0,129	0,165	0,197	0,232	0,256	0,305

Notas: (1) Tabela construída com base na estatística

$$t = \frac{r.\sqrt{n-2}}{\sqrt{1-r^2}}$$

que tem distribuição *t de Student* com $gl = n - 2$, sob as suposições de os dados terem distribuição normal e a correlação ser linear.

(2) A coluna em destaque é a mais usada.

Tabela 8 Valor absoluto mínimo para o coeficiente de correlação por postos, r_s de Spearman, ser significativo

	Nível de significância, α, num teste unilateral					
	0,100	**0,050**	**0,025**	**0,010**	**0,005**	**0,001**
	Nível de significância, α, num teste bilateral					
n	**0,200**	**0,100**	**0,050**	**0,020**	**0,010**	**0,002**
5	0,800	0,900	1,000	1,000	-	-
6	0,657	0,829	0,886	0,943	1,000	-
7	0,571	0,714	0,786	0,893	0,929	1,000
8	0,524	0,643	0,738	0,833	0,881	0,952
9	0,483	0,600	0,700	0,783	0,833	0,917
10	0,455	0,564	0,648	0,745	0,794	0,879
11	0,427	0,536	0,618	0,709	0,755	0,845
12	0,406	0,503	0,587	0,678	0,727	0,818
13	0,385	0,484	0,560	0,648	0,703	0,791
14	0,367	0,464	0,538	0,626	0,679	0,771
15	0,354	0,446	0,521	0,604	0,657	0,750
16	0,341	0,429	0,503	0,585	0,635	0,729
17	0,328	0,414	0,488	0,566	0,618	0,711
18	0,317	0,401	0,474	0,550	0,600	0,692
19	0,309	0,391	0,460	0,535	0,584	0,675
20	0,299	0,380	0,447	0,522	0,570	0,660
21	0,292	0,370	0,436	0,509	0,556	0,647
22	0,284	0,361	0,425	0,497	0,544	0,633
23	0,278	0,353	0,416	0,486	0,532	0,620
24	0,271	0,344	0,407	0,476	0,521	0,608
25	0,265	0,337	0,398	0,466	0,511	0,597
26	0,259	0,331	0,390	0,457	0,501	0,586
27	0,255	0,324	0,383	0,449	0,492	0,576
28	0,250	0,318	0,375	0,441	0,483	0,567
29	0,245	0,312	0,369	0,433	0,475	0,557
30	0,240	0,306	0,362	0,426	0,467	0,548
35	0,220	0,282	0,336	0,399	0,442	0,530
40	0,205	0,263	0,314	0,373	0,412	0,495
45	0,193	0,248	0,295	0,351	0,388	0,466
50	0,183	0,235	0,280	0,332	0,368	0,441
60	0,167	0,214	0,255	0,303	0,335	0,402
70	0,154	0,198	0,236	0,280	0,310	0,372
80	0,144	0,185	0,221	0,262	0,290	0,348
90	0,136	0,174	0,208	0,247	0,273	0,328
100	0,129	0,165	0,197	0,234	0,259	0,311

Notas: (1) Os valores para $n \leq 30$ foram extraídos de Leach (1979) e baseiam-se na distribuição exata. Para $n > 30$, a tabela foi construída com base na estatística $z = r_s \cdot \sqrt{n-1}$, que, sob a suposição de correlação linear, tem distribuição aproximadamente normal padrão.
(2) A coluna em destaque é a mais usada.

Respostas de alguns exercícios

Capítulo 2

2) Pesquisa de levantamento, pois numa pesquisa eleitoral procura-se obter as preferências dos eleitores quanto aos candidatos, sem que o entrevistador interfira no processo, ou seja, procura-se levantar os dados naturalmente, como eles se apresentam no momento da pesquisa.

4) a) altura em centímetros (quantitativa); d) sexo, possíveis respostas: masculino e feminino (qualitativa).

6) Quando um respondente depara com um questionário muito longo, este se cansa de responder e pode deixar parte do questionário em branco, ou responder apressadamente, comprometendo as respostas.

Capítulo 3

1) {Josefa, Joana, Joaquim, José de Souza, Arnaldo, Getúlio, Hercílio, Carlito Anastácia, Cardoso}

2) {1, 2, 22, 2, 2, 2 10, 3, 5, 16}

3) {G, U, X, J} (alfabeto conforme acordo ortográfico de 1990; 26 letras)

4) Não, basta extrair 100 números da tabela, com quatro algarismos, pertencentes ao conjunto {1650, 1651, ..., 8840}, sem repetição.

11) $n = 2.500$

12) $n = 286$

CAPÍTULO 4

2) Tabela de frequências múltipla: Distribuição de uma amostra de famílias quanto ao uso de programas de alimentação popular, por localidade da residência. Bairro Saco Grande II, Florianópolis – SC, 1988.

Uso de programas de alimentação popular	Localidade		
	Monte Verde	Pq. da Figueira	Encosta do Morro
não	18 (45,0%)	12 (27,9%)	12 (32,4%)
sim	22 (55,0%)	31 (72,1%)	25 (67,6%)
Total	40 (100,0%)	43 (100,0%)	37 (100,0%)

3) Tabela de frequências: O principal ponto positivo do Curso de Ciências da Computação – UFSC, na opinião dos alunos das três últimas fases, semestre 91.1.

Ponto positivo	professores	atualização	abrangência	prática	currículo	outros
frequência	13 (26%)	6 (12%)	7 (14%)	4 (8%)	5 (10%)	15 (30%)

NOTA: Dez alunos não responderam este item. As percentagens foram calculadas sobre os 50 respondentes.

6) Tabela de frequências: Distribuição de uma amostra de famílias quanto ao uso de programas de alimentação popular, por faixa de renda. Bairro Saco Grande II, Florianópolis, 1988.

Uso de programas de alimentação popular	Renda familiar	
	até 5 sal. mín.	mais de 5 sal. mín.
não	15 (27,3%)	27 (42,2%)
sim	40 (72,7%)	37 (57,8%)
Total	55 (100,0%)	64 (100,0%)

NOTA. Houve uma não resposta na amostra de 120 famílias.

8) a) Analisando a Tabela 1, observamos haver associação entre grau de instrução e uso de programas de alimentação popular, pois, enquanto no estrato das famílias de nível de instrução baixo 70% delas usam os programas, nas famílias de nível de instrução alto este percentual cai para 40%.

b) Se separarmos a nossa população por nível de renda familiar (Tabela 2), observamos uma completa independência entre grau de instrução e uso de programas de alimentação popular. As grandes diferenças quanto ao uso ou não dos programas fica entre os dois níveis de renda familiar considerados. Isto nos leva a crer que a associação observada na Tabela 1 é, na verdade, induzida pela variável renda familiar.

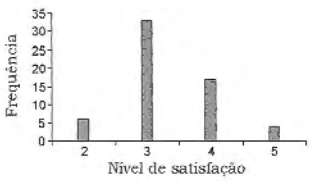
CAPÍTULO 5

1) Podemos dizer que o mais típico são residências com quatro ou cinco moradores. Não parece haver residência com número de moradores muito diferente das demais (casos discrepantes).

2) a)

Nível de satisfação	Frequência	%
2	6	10,00
3	33	55,00
4	17	28,33
5	4	6,67
Total	60	100,00

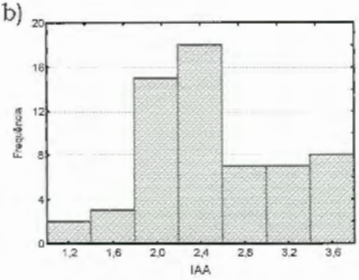
b)



5) a)

Classes	Freq.	%
1,0 |— 1,4	2	3,3
1,4 |— 1,8	3	5,0
1,8 |— 2,2	15	25,0
2,2 |— 2,6	18	30,0
2,6 |— 3,0	6	11,7
3,0 |— 3,4	8	11,7
3,4 |— 3,8	8	13,3
Total	60	100,0

b)



8)

1	3
1	5678899
2	00000111111122233333444
2	55555556667999
3	00111224
3	5556666

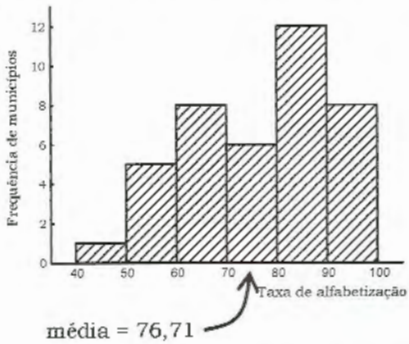
Capítulo 6

2) Média = 7 e desvio padrão = 0

4) Média = 7,6 e desvio padrão = 2,37

5) Média = 4,3 e desvio padrão = 1,45

6)



7) a) Média = 2,311 e desvio padrão = 1,206

8) Tabela: Medidas descritivas de algumas características do Curso Ciências da Computação – UFSC, na visão dos alunos das três últimas fases.

	Características do Curso						
	professores (didática)	professores (conhec.)	bibliografia disponível	recursos materiais	conteúdo das disc.	currículo	satisfação em geral
Média	2,77	3,23	2,20	2,30	3,40	3,35	3,32
DP	0,62	0,67	0,94	1,05	0,69	0,90	0,75

10) a) $M_d = 15$; $Q_I = 10,5$ e $Q_S = 19,5$

b) $M_d = 15,5$; $Q_I = 10,5$ e $Q_S = 19,5$

11) $M_d = 4$; $Q_i = 3,5$ e $Q_s = 5$

12) $M_d = 2,45$; $Q_I = 2,10$ e $Q_S = 2,97$

13) $E_i = 1$; $Q_i = 2$; $M_d = 4$; $Q_s = 5$ e $E_s = 12$

16) Não, para se ter a taxa de alfabetização da Unidade da Federação, precisa-se calcular a média ponderada pela população adulta de cada município.

CAPÍTULO 7

1) a)

Resultados	1	2	3	4	5	6	7	8	9	10
Probabilidades	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1

b) $A = \{2, 4, 6, 8, 10\}$; $B = \{1, 3, 5, 7, 9\}$ e $C = \{1, 2\}$.

c) $P(A) = 1/2$; $P(B) = 1/2$ e $P(C) = 1/5$.

2)

Resultados	homem	mulher
Probabilidades	1/3	2/3

3) a)

Resultados	*A*	*B*	*branco ou nulo*
Probabilidades	0,30	0,50	0,20

b) 0,80

4) a) 78/120 b) 44/120 c) 76/120 d) 25/120 e) 53/120
f) 25/44 g) 25/78

5) 0,4225

6) a) É binomial com $n = 3$ e $\pi = 5/8$.

b) Não é binomial. Os ensaios não são independentes.

c) É binomial com $n = 20$ e π = proporção de mulheres na população, na época da pesquisa.

d) É binomial com $n = 500$ e π = proporção de pessoas favoráveis em SC, na época da pesquisa.

e) Não é binomial. O parâmetro π não é constante ao longo dos ensaios;

f) É binomial com $n = 100$ e π = proporção de recém-nascidos em SC com menos de 2 kg, na época da pesquisa.

g) Não é binomial. A característica em estudo não pode ser identificada em apenas dois resultados, em cada ensaio.

7) a) 0,3125 b) 0,500

8) 0,3770

9) Binomial com $n = 5$ e $\pi = 0,40$; ou seja:

x	0	1	2	3	4	5
p(x)	0,0778	0,2592	0,3456	0,2304	0,0768	0,0102

11) a) 0,663 b) 0,337 c) 0,3174

12)

Resultado	0,0	0,2	0,4	0,6	0,8	1,0
Probabilidade	0,0778	0,2592	0,3456	0,2304	0,0768	0,0102

13) a) 0,2753 b) 0,0334

14) 0,0702

16) a) 0,1646 b) 0,1317 c) 0,7901
17) a) 0,7082 b) 0,0027 c) 0,2918
18) 8/15
19) a) 0,6553 b) 0,2458 c) 0,7379
20) a) 0,3284 b) 0,6219
21) a) 0,3874 b) 0,0702
22) a) 0,3125 b) 0,3437
23) 0,0781

Capítulo 8

1) a) 2 b) 1,5 c) 0 d) -0,5
2) 0,50
3) a) 1,33 b) 75
4) a) 0,0495 b) 0,9505 c) 0,6826 d) 0,9544 e) 0,9974
   f) 0 g) 1,65 h) 2,58
5) a) 0,0228 b) 0,9544 c) 0,1587 d) 95,44%
6) a) 0,0228 b) 68,26%
7) Ambos os eventos têm a mesma probabilidade (igual a 0,1056).
8) a) 0,1719 b) 0,1711
9) 0,6255
10) 0,0968
11) 0,985
12) 6,68%
13) a) 0,1056 b) 0,3085
14) a) 0,6826 (usando a distribuição binomial) b) 0,9032 (usando a distribuição normal)
15) a) 0,0781 b) ≈ 0
16) 85,36 minutos (ou 85 minutos e 22 segundos)

Capítulo 9

1) a) 43/90 b) 5,99
4) a) 60,0% ± 4,0% b) 60,0% ± 2,5% c) 20,0% ± 3,9%
   d) 80,0% ± 3,9% e) 50,0% ± 4,9%
   Obs.: Nível de confiança de 95% usando o valor aproximado z = 2.
5) 30,0% ± 6,4%
6) a) Na amostra: 30,0%. Na população: com 95% de confiança o intervalo 30,0% ± 4,5% contém a referida proporção.
   b) Nada. A amostragem não foi aleatória.

7) 35,0% ± 12,4%

8) 65,0% ± 8,6%

9) a) 55,0% ± 15,7% b) 72,1% ± 13,7% c) 67,6% ± 15,2%

10) a) 16,00 minutos b) 3,11 minutos c) 0,83 minutos
d) 16,00 ± 1,80 minutos

11) Nos cálculos abaixo, usamos o valor aproximado $t = 2$ (pois as amostras eram razoavelmente grandes).

Localidade	Renda média familiar mensal (em salários mínimos)
Monte Verde	8,1 ± 1,4
Pq. da Figueira	5,8 ± 0,8
Encosta do Morro	5,0 ± 1,5

Interpretação: A renda média familiar dos moradores do Monte Verde é de 8,1 salários mínimos mensais, com um erro amostral máximo (95% de confiança) de 1,4 salários mínimos. Interpretações análogas para Parque da Figueira e Encosta do Morro.
Note que com estes resultados, podemos afirmar (com pelo menos 95% de confiança), que a renda média familiar dos moradores do Monte Verde é maior do que nas duas outras localidades em estudo. Mas a diferença da renda média do Parque da Figueira e Encosta do Morro pode ser meramente casual, resultante da sorte (ou azar) das amostras extraídas, pois os intervalos de confiança têm uma área de sobreposição.

12) a) R$255,00 ± R$135,00
b) Valor, em real, que o fiscal deixa de cobrar, em média, por empresa que ele possa fazer a auditoria.
c) Com 95% de confiança, o intervalo R$255,00 ± R$135,00 contém o desconhecido valor μ.

15) 33,3% ± 7,3%

17) $n = 64$ (usando $z = 2$)

18) $n = 306$ (usando $z = 2$)

19) a) população: conjunto de todos os alunos do curso;
amostra: os 80 alunos selecionados;
parâmetro: proporção de alunos do Curso favoráveis à eliminação da disciplina de estatística;
estatística: proporção de alunos favoráveis à eliminação da disciplina de estatística dentre os 80 da amostra.
b) população: pessoas obesas da cidade;
amostra: as 20 pessoas obesas selecionadas para o estudo;
parâmetro: perda esperada de peso de uma pessoa que faça o curso;
estatística: perda média de peso das 20 pessoas selecionadas para o estudo.
c) população: pessoas fumantes da cidade;
amostra: as 100 pessoas selecionadas para o estudo;
parâmetro: proporção de fumantes da cidade que largaram o vício após a campanha.
estatística: proporção de fumantes que largaram o vício após a campanha dentre as 100 pessoas selecionadas para o estudo.

20) a) 40%
b) Com 95% de confiança, o intervalo 40,0% ± 3,4% contém a percentagem dos habitantes da cidade que apoiam a administração da prefeitura.

21) a) $n = 664$ b) 30,1% ± 4,6%. Com 99% de confiança, o intervalo 30,1% ± 4,6% contém a percentagem de pessoas que passariam a usar o produto.

22) 13,6% ± 2,6%

23) a) 3,50 ± 0,64 b) n = 98 (foram usados $t = 2{,}201$ e $N = 500$)

24) a) média = – 3,900 kg, d.p. = 8,373 kg e mediana = – 3,5 kg

b) -3,900 kg ± 5,989 kg

c) Não, pois o intervalo de confiança apresenta, também, valores positivos, ou seja, o valor esperado da variação de peso pode ser positivo (ganho de peso).

24) a) $n = 192$ b) 5,30 ± 0,46

c) Não, pois o intervalo onde deve estar a verdadeira média abrange, também, valores menores que cinco.

d) 62,5% ± 5,5%

26) 6,0%, 5,6% e 5,8%, respectivamente.

Capítulo 10

1) a) 0,0062 b) 0,3874 c) 0,0062

2) a) Rejeita H_0 b) Aceita H_0 c) Rejeita H_0

3) É possível. Por exemplo, se no teste para verificar se uma moeda é honesta ocorrer $Y = 2$ caras em $n = 12$ lançamentos, temos $p = 0{,}0384$, que rejeita ao nível de 5%, mas aceita ao nível de 1%. O inverso nunca acontece.

4) a) bilateral b) unilateral c) bilateral

5) a) 0,0031 b) 0,1937 c) 0,6127

6) a) 0,0094 b) 0,3844 c) 0,0094

8) Sim (rejeita H_0 ao nível de 5%), pois $p = 0{,}0222$ (teste unilateral).

9) Sim (rejeita H_0 ao nível de 5%), pois $p = 0{,}0014$ (teste unilateral).

10) a) H_0: Em média, a produtividade com treinamento é igual à produtividade sem treinamento. H_1: Em média, a produtividade com treinamento é maior do que a produtividade sem treinamento (teste unilateral).

b) H_0: Em média, a velocidade é igual ao valor anunciado. H_1: Em média, a velocidade é menor do que o valor anunciado (teste unilateral).

c) H_0: As produtividades médias são iguais para os dois métodos de treinamento. H_1: As produtividades médias são diferentes para os dois métodos de treinamento (teste bilateral).

11) a) Decide-se por H_1, pois o valor p é menor do que o nível de significância adotado. O risco de ele estar tomando a decisão errada é de 0,0001. (É claro que estamos considerando apenas os aspectos estatísticos).

b) Decide-se por H_0, pois o valor p é maior do que os níveis de significância normalmente adotados. Quando se aceita H_0, o valor p não oferece qualquer informação sobre o risco de se estar tomando a decisão errada.

c) Quanto menor o valor p, existe maior evidência para a rejeição de H_0 (e consequente aceitação de H_1).

12) a) Aceita H_0: a moeda é honesta ($p = 0,2892$).
b) Rejeita H_0, isto é, decide-se que a moeda é viciada ($p \approx 0,0000068$, uso da aproximação normal).

13) Hipóteses: H_0: $\pi = 0,5$ e H_1: $\pi > 0,5$ (π = probabilidade da criança acertar uma dada questão). Decisão: rejeita H_0, isto é, há evidência de que a criança tem algum conhecimento sobre o assunto ($p = 0,0031$).

14) a) H_0: $\pi = 0,25$ e H_1: $\pi > 0,25$; b) $\mu = 3$ c) $p = 0,1576$
d) Aceita H_0. Não há evidência de que a criança tem algum conhecimento sobre o assunto.

15) Decisão: rejeita H_0, isto é, há evidência de que o sistema "inteligente" adquiriu algum conhecimento sobre o assunto ($p = 0,0071$, uso da aproximação normal).

CAPÍTULO 11

1) a) H_0: não há diferença entre a percentagem de ouvintes que avaliam positivamente e a percentagem de ouvintes que avaliam negativamente a apresentação do candidato; H_1: a maior parte dos ouvintes avalia positivamente a apresentação do candidato.
b) $p = 0,1134$. Portanto, ao nível de significância de 5%, não há evidência de que houve melhora (Aceita H_0).
c) $p \approx 0$. Portanto, ao nível de significância de 5%, há evidência de melhora (Rejeita H_0).
d) $p \approx 0,00135$. Portanto, ao nível de significância de 5%, há evidência de melhora (Rejeita H_0).

3) a) H_0: em média, o curso não produz efeito no peso; H_1: em média, as pessoas que fazem o curso reduzem mais o peso do que as que não fazem o curso.
b) Ao nível de significância de 5%, rejeita H_0, isto é, podemos afirmar que o curso produz efeito no sentido desejado ($0,01 < p < 0,025$).

4) b) Rejeita H_0 ao nível de 5%, pois $t = 2,70 \rightarrow 0,01 < p < 0,025$ (teste unilateral).

5) a) Rejeita H_0 ao nível de 5%, pois, $t = 3,04 \rightarrow 0,005 < p < 0,010$ (teste unilateral).

6) Sim, rejeita H_0 ao nível de 1%, pois, $t = 6,4 \rightarrow p < 0,0005$ (teste unilateral).

7) Sim, rejeita H_0 ao nível de 1%, pois, $t = 3,09 \rightarrow 0,001 < p < 0,005$ (teste bilateral).

8) a) Não (aceita H_0 ao nível de 5%), pois $t = 1,33 \rightarrow 0,05 < p < 0,10$ (teste unilateral).
b) Mesmo que o teste rejeitasse H_0, apontando diferença significativa entre os dois grupos, não poderíamos garantir que esta diferença seja devida ao nível nutricional da mãe, pois nada garante que os dois grupos se difiram somente com respeito a este fator, já que não é uma pesquisa experimental.

9) Não (aceita H_0 ao nível de significância de 5%), pois $t = 1,018 \rightarrow 0,20 < p < 0,50$ (teste bilateral).

10) Sim (rejeita H_0 ao nível de significância de 5%), pois $t = -2,16 \rightarrow 0,02 < p < 0,05$ (teste bilateral).

12) Três testes bilaterais, admitindo $\alpha = 0,01$ para cada teste:
Monte Verde x Pq. da Figueira: existe diferença significativa, pois $t = 2,92 \rightarrow p \approx 0,005$.

Monte Verde x Encosta do Morro: existe diferença significativa, pois $t = 3,07$ → $0,002 < p < 0,005$.
Pq. da Figueira x encosta do Morro: não existe diferença significativa, pois, $t = 0,99$ → $0,20 < p < 0,50$.

13) 23 (usando o gráfico da Figura 11.11).

14) Não. Usando teste t unilateral para amostras independentes: $t = 1,51$ $(0,05 < p < 0,10)$

15) Sim. Usando teste t unilateral para dados pareados: $t = 3,10$ $(0,01 < p < 0,025)$

16) Não. Usando o teste unilateral dos sinais, $p = 0,1094$.

17) Sim. Teste t unilateral para dados pareados: $t = 1,62$ $(0,05 < p < 0,10)$.

18) Não. Teste t bilateral para amostras independentes: $t = 0,97$ $(0,20 < p < 0,50)$. Portanto, a diferença entre as médias amostrais pode ser explicada meramente pelo acaso.

19) Sim. Teste t unilateral para amostras independentes: $t = 3,92$ $(p < 0,0005)$.

Capítulo 12

1) a) 3,53 b) 2,40 c) Não $(0,10 < p < 0,25)$

2) Sim, pois $\chi^2 = 6,84$ é $0,005 < p < 0,01$

3) a) Sim (rejeita H_0), pois $\chi^2 = 16,25$ → $0,0025 < p < 0,005$.
b) Sim (rejeita H_0), pois $\chi^2 = 11,18$ → $0,0025 < p < 0,005$.
c) Não (aceita H_0), pois $\chi^2 = 5,14$ → $0,05 < p < 0,10$.

4) Adotando $\alpha = 0,05$. a) Não (aceita H_0), pois $\chi^2 = 2,82$ → $0,10 < p < 0,25$.
b) Sim (rejeita H_0), pois $\chi^2 = 6,72$ → $0,0025 < p < 0,05$.

5) a) $C^* = 0,107$.
b) $\phi = 0,076$. Os dados observados apresentam uma fraca associação entre *sexo* e *tabagismo*.

6) a) $C^* = 0,423$ b) $V = 0,260$

7) $\gamma = 0,3356$.

8) a) 0,214 b) -0,185

9) 0,665

10) Não. $(\chi^2 = 1,77, p > 0,25)$

11) Sim, conforme o teste qui-quadrado com correção de Yates $(\chi^2 = 2,99, 0,05 < p < 0,10)$, existe associação significativa entre o tipo de escola (pública ou particular) e o resultado no vestibular (aprovação ou reprovação), ao nível de significância de 10%.

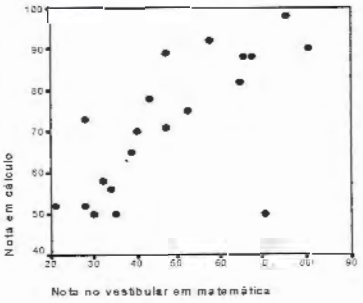
Tipo de escola	Aprovação no vestibular	
	não	sim
pública	13 (72%)	4 (33%)
particular	5 (28%)	8 (67%)
Total	18 (100%)	12 (100%)

12) Não. $(c^2 = 2,25, p > 0,25)$

13) a) Teste qui-quadrado com correção de Yates.
b) Teste t para amostras independentes.
c) Teste t para amostras independentes.

Capítulo 13

2) a) Sugere correlação positiva b) Ponto discrepante: nona observação (70, 50)



c) 0,69

d) 0,86. Correlação positiva e significativa (teste bilateral, $\alpha = 0,05$).

e) 0,66. É significativa (teste bilateral, $\alpha = 0,05$)

6) a) $r = -0,684$. Em termos dos doze municípios pesquisados, e na época de observação dos dados, verificou-se uma correlação negativa moderada entre *taxa de alfabetização* e *taxa de mortalidade infantil*. Então, para níveis maiores de alfabetização, temos uma leve tendência de redução na taxa de mortalidade infantil.

b) $r_s = -0,678$. Significativo ao nível de significância de 5% (teste bilateral); assim, podemos dizer que existe correlação (e negativa) entre essas duas variáveis, nos municípios brasileiros.

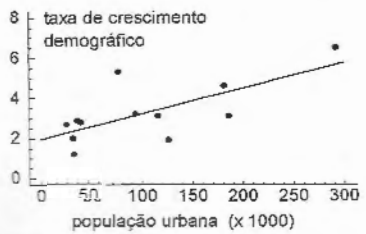
9) a) Variável dependente: nota; variável independente: número de faltas;

b) $\hat{y} = 9,51 - 0,63x$ d) $R^2 = 0,68$ e) $S_e = 1,64$

10) a) Variável dependente: taxa de crescimento demográfico; e variável independente: população urbana

b) (taxa de cresc. dem.) = 1,758 + (0,01253).(pop. urbana). Obs.: População urbana está em unidades de 1.000 habitantes.

c)



d) Predição: taxa de crescimento de 5,52.

e) $R^2 = 48\%$

12) Não. Pela tabela 7 o valor absoluto de r deveria ser no mínimo igual a 0,444 para ser significativo.

13) a) $r = -0,85$. Para as seis famílias pesquisadas, tem-se uma correlação negativa forte entre *renda familiar* e *número de filhos*.

b) $\gamma = 0{,}33$. Em relação aos dez indivíduos pesquisados, verifica-se uma correlação positiva fraca.

c) $C^* = 0{,}09$. Em relação aos cem indivíduos pesquisados, praticamente não existe associação entre altura e sexo.

14) a) $r = 0{,}925$

b) Correlação positiva forte. É também significativamente diferente de zero (Tabela 7)

c) $y = 1{,}19 + 1{,}70\ x$

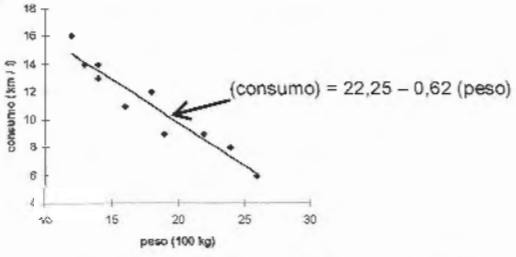
15) a) 49,1 kg b) 1,8 kg

16) a) $r = -0{,}96$ b) Correlação negativa forte

c) Variável dependente: consumo; e variável independente: peso

d) (consumo) = 22,25 – 0,62 (peso)

e)



f) Sim, verifica-se pelo gráfico do item (e) que uma relação linear parece adequar-se bem ao presente problema. Além disso, tem-se um coeficiente de determinação próximo de 1 ($R^2 = 0{,}92$).

g) 9,85 km / l.

h) Não, pois os veículos estudados estavam na faixa de 1.200 a 2.600 kg e, portanto, a equação de regressão deve ser usada apenas nesta faixa.

Coleção Didática

A interpretação de imagens aéreas
Algoritmos numéricos – sequenciais e paralelos
Análise sensorial de alimentos
Anatomia sistêmica – uma abordagem direta para o estudante
Anomalias laríngeas congênitas
Assistência social: do discurso do Estado à prática do Serviço Social
AutoCAD 2000 – guia prático para desenhos em 2D
AutoCAD 2004 – guia prático para desenhos em 2D
AutoCAD R14 – guia prático para desenhos em 2D
AutoCAD R14 – guia prático para desenhos em 3D
Avaliação nutricional de coletividades
Cálculo 1
Cálculo A
Cálculo C
Cálculo de indutância e de força em circuitos elétricos
Cálculo e Álgebra Linear com Derive
Câncer – o que você precisa saber
Cartografia – representação, comunicação e visualização de dados espaciais
Centro cirúrgico: aspectos fundamentais para Enfermagem
Classificação Decimal Universal – CDU
Construindo em alvenaria estrutural
Desenho geométrico
Desenho técnico mecânico
Diagnóstico do meio físico de bacias hidrográficas
Elementos básicos de fotogrametria e sua utilização prática
Eletromagnetismo e cálculo de campos
Eletromagnetismo para Engenharia: estática e quase estática
Eletrônica básica: um enfoque voltado à Informática
Engenharia de protocolos com LOTOS/ISO
Estatística aplicada às Ciências Sociais
Ferramentas de corte I
Ferramentas de corte II
Filtros seletores de sinais
Fundamentos de Cartografia
Fundamentos de sistemas hidráulicos
Geração de vapor
Gramática básica do Latim
Identificação de sistemas dinâmicos lineares
Influência açoriana no Português do Brasil
Inteligência Artificial
Inteligência Artificial: ferramentas e teorias
Introdução à Engenharia
Introdução à Engenharia: conceitos, ferramentas e comportamentos
Introdução à Física Nuclear e de Partículas Elementares
Introdução à Matemática
Introdução à Química Inorgânica Experimental
Introdução à Teoria dos Grafos
Introdução à Topologia Geral
Introdução ao Laboratório de Física
Latim para o português – gramática, língua e literatura
Le Français Parlé, pratique de la prononciation du Français
Macroescultura dental
Manual básico de Desenho Técnico
Maple V
Matemática – 100 exercícios de grupos
Matemática Financeira através da HP-12C
Matrizes e sistemas de equações lineares
Microbiologia – manual de aulas práticas
Monitoramento global integrado de propriedades rurais
Natação: ensine a nadar
Noções básicas de Geometria Descritiva
O papel da escola na construção de uma sociedade democrática
Óleos e gorduras vegetais – processamento e análise
Princípios de combustão aplicada
Promenades – textes et exercises pour la classe de Français
Propriedades químicas e tecnológicas do amido de mandioca e do polvilho azedo
Química Básica – teoria e experimentos
Redação
Redação oficial
Redes de Petri
Taguchi e a melhoria da qualidade: uma releitura crítica
Teaching in a clever way – tarefas comunicativas para professores de Língua Inglesa do 1o grau
Tecnologia de grupo e organização da manufatura
Teoria fundamental do motor de indução
Topografia contemporânea – Planimetria
Transmissão de energia elétrica
Unidades de informação: conceitos e competências
Ventilação industrial

Este livro foi editorado em
Korinna Bt e Arno Pro.
Miolo em papel *offset* 75g;
capa em cartão supremo 250g.
Impresso na Gráfica e Editora Copiart
em sistema de impressão *offset*.

A **Coleção Didática** da Editora da UFSC procura estabelecer uma linha objetiva de contato entre os alunos, o professor, a atividade de ensino e a sala de aula. Constitui-se de livros universitários e tem como proposta um apanhado de conteúdo programático resultante do aperfeiçoamento de textos usados em sala de aula, que incluem exercícios e demonstrações, clareza de explicação e abordagem.

9 788532 806048

Pedro Alberto Barbetta é bacharel em Estatística pela Escola Nacional de Ciências Estatísticas – ENCE/IBGE, mestre em Estatística pelo Instituto de Matemática Pura e Aplicada – IMPA/CNPq e doutor em Engenharia de Produção pela Universidade Federal de Santa Catarina – UFSC. Professor Adjunto, lotado no Departamento de Informática e Estatística da UFSC desde 1982, ministra disciplinas de Estatística nos cursos de graduação e pós-graduação em Ciências Sociais, Psicologia, Economia, Administração, Enfermagem, Neurociências e Comportamento, Engenharia de Produção e Computação. Também participa de ensino por videoconferência.
Complementos em:
<www.inf.ufsc.br/~barbetta/livro1.htm>