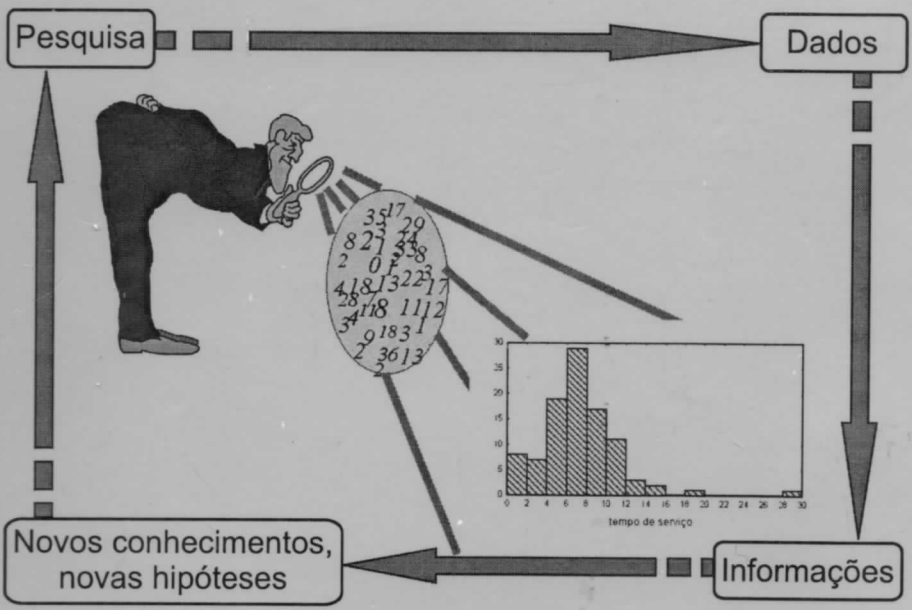
ESTATÍSTICA

APLICADA ÀS CIÊNCIAS SOCIAIS

5ª edição revisada



PEDRO ALBERTO BARBETTA

EDITORA DA UFSC

Pedro Alberto Barbetta

Estatística Aplicada às Ciências Sociais

5ª edição revisada

Editora da UFSC
Florianópolis
2002

Editora da UFSC

Campus Universitário – Trindade
Caixa Postal 476
88010-970 – Florianópolis – SC
(048) 331-9408, 331-9605 e 331-9686
(048) 331-9680
edufsc@editora.ufsc.br
http://www.editora.ufsc.br

Capa:
Paulo Roberto da Silva

Supervisão técnico-editorial:
Aldy Vergés Maingué

Revisão:
Ana Lúcia Pereira do Amaral

Ficha Catalográfica

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

B235e Barbetta, Pedro Alberto
Estatística aplicada às Ciências Sociais / Pedro Alberto Barbetta. 5. ed. – Florianópolis : Ed. da UFSC, 2002.
340p. : il. (Série Didática)

Inclui bibliografia

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

CDU: 31:3
CDD: 300:21

Impresso no Brasil

SUMÁRIO

PREFÁCIO À 4ª EDIÇÃO

Com seis anos utilizando as edições anteriores deste livro, sugestões e contribuições de diversos professores e alunos, aos quais somos muito grato, construímos a 4ª edição com melhor apresentação, mais figuras ilustrativas, mais exemplos, vários exercícios complementares, tópicos adicionais e saídas comentadas de programas computacionais, especialmente da planilha eletrônica *Microsoft Excel*. Enfatizamos a interação entre estatística e metodologia de pesquisa. Incluímos a questão do tamanho da amostra em estudos comparativos (Capítulo 11), a análise de correlação por postos (Capítulo 13) e, principalmente, complementamos a análise de regressão, introduzindo a análise de resíduos, transformações e uma introdução à regressão múltipla (Capítulo 13). Com o grande número de programas computacionais, hoje é possível levar ao aluno as técnicas associadas à análise de regressão, sem precisar apresentar um exaustivo curso de matemática e de estatística. A análise de regressão é extremamente importante na pesquisa das ciências sociais e humanas, como poderá ser percebido no Capítulo 13.

Pedro Alberto Barbetta

PREFÁCIO

Nas reuniões sobre o ensino da estatística, muito se tem discutido sobre o problema de oferecer disciplinas introdutórias em cursos das áreas das Ciências Sociais e Humanas. A maior dificuldade está no fato de que os métodos estatísticos são embasados numa rigorosa formulação matemática e de que os alunos destas áreas, em geral, não têm grande familiaridade com a matemática. Na tentativa de tentar contornar este problema, 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 livro apresenta uma introdução à estatística, juntamente com uma orientação básica de como planejar e conduzir uma pesquisa social. Além disso, todos os capítulos iniciam com problemas práticos que motivam e justificam a introdução de técnicas estatísticas.

O texto começou a ser escrito em 1989 e suas versões preliminares já foram amplamente testadas em disciplinas de estatística ministradas na UFSC, abrangendo os cursos de Ciências Sociais, Psicologia, Administração, Biblioteconomia, Arquitetura e Urbanismo, além das pós-graduações em Administração e Enfermagem. Os alunos destes cursos merecem nossa imensa gratidão porque através de suas críticas e sugestões conseguimos aperfeiçoar nosso material e chegar à versão atual, que tem recebido muitos elogios. Agradecemos, também, as contribuições dos professores Sílvia Nassar, Edla F. Ramos, Paulo J. Ogliari, Masanao Ohira, Antonio C. Bornia, Cristiano J.C.A. Cunha e Arno Blass e dos funcionários da Editora da UFSC pelo apoio na revisão e na editoração.

O livro inicia com uma visão geral dos métodos que serão tratados e apresenta algumas idéias básicas sobre o planejamento de uma pesquisa social (Capítulos 2 e 3). Estes itens não precisam necessariamente ser desenvolvidos no início do curso. Os Capítulos 4 a 6 trazem alguns dos principais elementos da Estatística Descritiva e da Análise Exploratória de Dados, incluindo as suas aplicações em pesquisas de campo desenvolvidas na UFSC. Alguns modelos de probabilidades, que serão necessários para o entendimento de capítulos posteriores, são apresentados nos Capítulos 7 e 8. E os Capítulos 9 a 13 introduzem alguns métodos estatísticos propriamente ditos, também com aplicações em problemas reais.

Pedro Alberto Barbetta

Capítulo 1

Introdução

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

Para 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 ilustrativo para facilitar a nossa discussão.

Exemplo 1.1 Com o objetivo de levantar conhecimentos sobre o *grau 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 dos elementos de uma população, com base nos dados de uma amostra*. Chamamos de ***população*** o conjunto de elementos que formam o universo de nosso estudo e que são passíveis de serem observados. Uma parte destes elementos é dita uma ***amostra***.

Coleta de dados

Para conhecermos certas características dos elementos de uma população (ou de uma amostra), precisamos coletar dados destes elementos. É uma fase da pesquisa que precisa ser cuidadosamente planejada, para que dos dados a serem levantados forneçam informações relevantes, em termos

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

dos objetivos da pesquisa. É no planejamento da obtenção dos dados que devemos planejar, também, *o que fazer com eles*. Esta fase do trabalho será discutida 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, residentes na região em estudo. Ao observar o grau de instrução do chefe da casa, o entrevistador classificava a resposta do entrevistado numa das três seguintes categorias: (1) *nenhum grau de instrução completo*, (2) *primeiro grau completo* e (3) *segundo grau completo*. É claro que, ao coletar os dados desta forma, já se tinha em mente os procedimentos estatísticos que seriam usados na futura análise destes dados, com a finalidade de atender aos objetivos da pesquisa.

Descrição e exploração de dados

Depois de observada uma amostra de famílias (Exemplo 1.1), ficamos com um conjunto de dados relativos à variável *grau de instrução do chefe da casa*. Estes 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 freqüências*.

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

Uma distribuição de freqüências do grau de instrução, por exemplo, deve informar *quantas* pessoas (ou a *percentagem* de pessoas) que se enquadram em cada categoria preestabelecida do grau de instrução. A Figura 1.1 mostra, sob forma de um gráfico, a distribuição de freqüências do grau de instrução do chefe da casa, numa amostra de 120 famílias (Exemplo 1.1).[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 freqüências assim como suas representações em tabelas e gráficos serão vistas nos Capítulos 4 e 5.

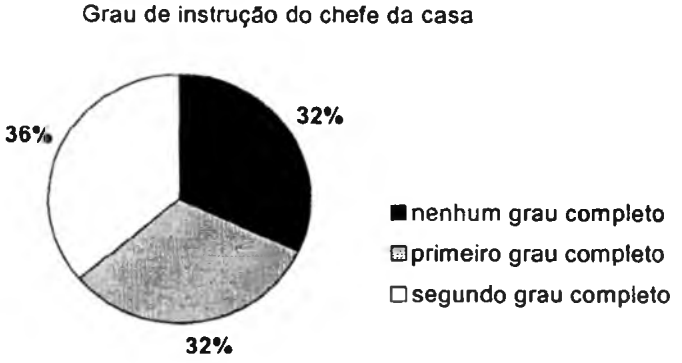


Figura 1.1 Distribuição de freqüências do grau de instrução do chefe da casa. Amostra de 120 famílias do bairro Saco Grande II, Florianópolis – SC, 1988.

A região em estudo (bairro Saco Grande II) pode ser vista como uma agregação de três localidades: Conjunto Residencial Monte Verde, Conjunto Residencial Parque da Figueira e Encosta do Morro. Considerando que haja interesse em comparar estas três localidades, construímos a Figura 1.2, que apresenta três distribuições de freqüências, sendo uma para cada localidade.

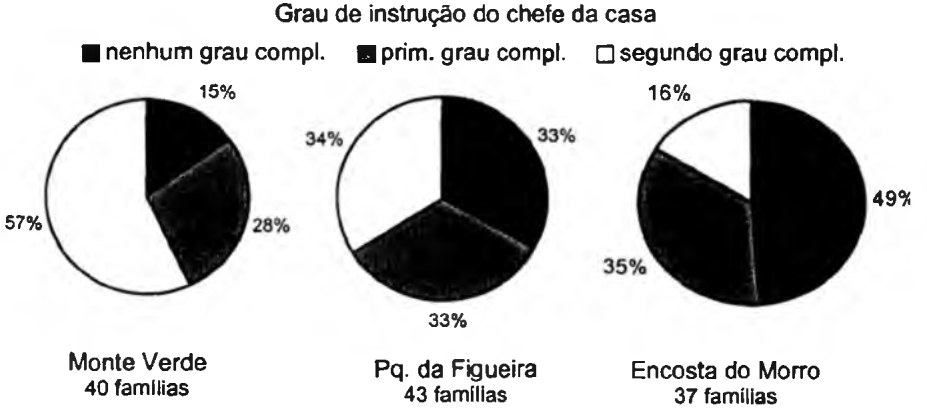


Figura 1.2 Distribuição de freqüências do grau de instrução do chefe da casa, por localidade. Amostra de 120 famílias do Bairro Saco Grande II, Florianópolis – SC, 1988.

Ao descrever os dados, começamos a *explorar* como deve ser a população de onde estes dados foram extraídos. A Figura 1.2, por exemplo, parece sugerir que, na região pesquisada, o perfil do grau 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 é chamada de *análise exploratória de dados*, que é uma tentativa de captar a essência das informações contidas nos dados, através da descrição adequada em tabelas e, principalmente, em gráficos. É a busca de um padrão 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, torna-se importante estudarmos os chamados modelos probabilísticos (Capítulos 7 e 8), que constituem uma forma de mensurar a incerteza e, em conseqüência, fornecem uma metodologia adequada para generalizar resultados da amostra para a população. Os modelos probabilísticos formam a base teórica para se completar a análise estatística de um conjunto de dados, que pode ser feita sob a forma de estimação de parâmetros ou de teste de hipóteses, como ilustraremos a seguir, após introduzir novos conceitos fundamentais.

Chamamos de ***parâmetro*** alguma 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 o segundo grau 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 segundo grau completo* 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, portanto, uma forma de inferência estatística. Uma outra forma de inferência estatística surge quando temos alguma hipótese sobre a população em estudo e queremos verificar a sua validade, a partir de uma amostra. São os chamados *testes estatísticos de hipóteses* ou *testes de significância*.

> *O cientista tem idéias sobre a natureza da realidade (idéias que ele denomina hipóteses) e freqüentemente testa suas idéias através de pesquisa sistemática* (LEVIN, 1985, p.1).

No problema do Exemplo 1.1, poderíamos ter interesse em testar a seguinte hipótese: *a distribuição do grau 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 freqüências nas três localidades. Por exemplo, enquanto no Monte Verde temos 57% de famílias com o chefe da casa possuindo o segundo grau completo, na Encosta do Morro, este percentual cai para 16%. Mas estas diferenças nos resultados da amostra são suficientes para afirmarmos que elas também existem na população?

Para inferirmos adequadamente 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 deste tipo de problema.

Em pesquisas empíricas, é fundamental se testar adequadamente 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).

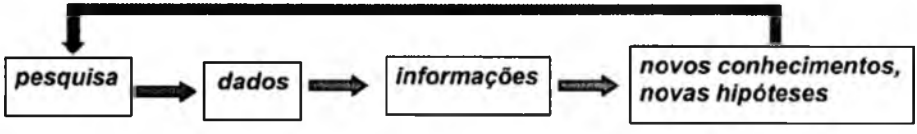


Figura 1.3 O processo interativo da evolução do conhecimento.

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

Pesquisas e dados [1]

Em nossas decisões do dia-a-dia estamos direta ou indiretamente nos baseando em dados observados. 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 deste bem com referência a outras 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.

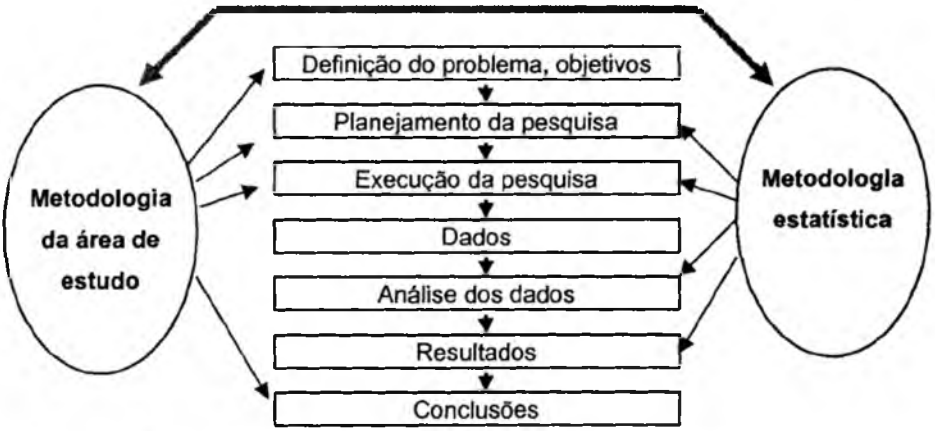


Figura 2.1 Etapas usuais de uma pesquisa quantitativa.

[1] Este capítulo teve a contribuição da Profª SÍLVIA MODESTO NASSAR (INE/ CTC/ 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 estes 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 seqüência 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 nesta empresa.
b) Conhecer a distribuição do grau de instrução dos funcionários.
c) Verificar o interesse dos funcionários em participar de programas de treinamento.
d) Avaliar o grau de satisfação dos funcionários com o trabalho que exercem na empresa.

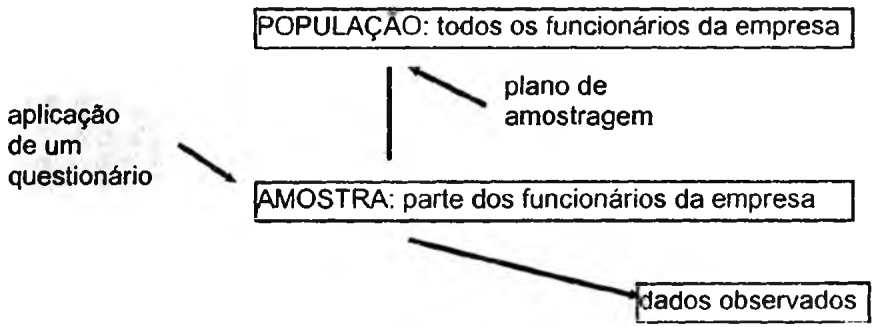
e) Verificar se existe associação entre o grau de satisfação do funcionário com a sua produtividade.[2]

A elaboração dos objetivos específicos deve ser feita de tal forma que forneça uma primeira indicação das características que precisamos observar ou medir. Por exemplo, para atingir aos objetivos do problema em questão, precisamos levantar as seguintes características de cada funcionário da empresa: *tempo de serviço, grau de instrução, interesse em participar de programas de treinamento, grau 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. Veja os seguintes exemplos.[3]

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

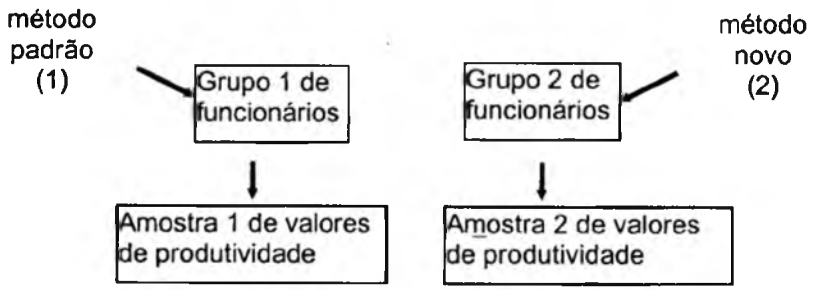


[2] Os objetivos de (a) a (d) podem ser alcançados por uma pesquisa capaz de descrever as características pertinentes da população. Por outro lado, o objetivo (e) é mais analítico, pois nele está embutida a hipótese de que exista associação entre satisfação e produtividade, que deverá ser colocada à prova.

[3] Uma descrição mais completa sobre os tipos de pesquisa pode ser encontrada em livros de metodologia de pesquisa, como em Selltiz, Wrightsman, Cook (1987) volume 1. Veja *Referências Bibliográficas* no final do livro.

O Exemplo 2.1 ilustra uma *pesquisa de levantamento* ou *survey*. Neste tipo de pesquisa observam-se diversas características dos elementos de uma certa população, utilizando-se questionários ou entrevistas. A observação é feita naturalmente e sem interferência do pesquisador. A pesquisa tipo levantamento é bastante comum nas Ciências Sociais e costuma gerar grandes conjuntos de dados. Na seqüência deste livro daremos mais destaque a este 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 em estudo. *Dados observados*: 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* em que o pesquisador exerce controle sobre o método de treinamento que vai ser aplicado a cada funcionário. Este 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, sobre a validade ou falsidade da hipótese previamente formulada.[4]

De um lado oposto, temos as situações em que conhecemos muito pouco sobre o universo a ser estudado. Nestes casos, podemos realizar uma *pesquisa qualitativa*, observando detalhadamente um pequeno número de elementos, sem uma formulação criteriosa das características a serem levantadas. Neste tipo de pesquisa 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.

Chamamos de ***população alvo*** 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. E, conseqüentemente, as conclusões baseadas nestes dados não valem, necessariamente, para todo o conjunto de funcionários.

Definimos como ***população acessível***, ou simplesmente como ***população***, o conjunto de elementos que queremos abranger em nosso estudo e que são passíveis de serem observados, com respeito às características que pretendemos levantar. Realizando adequadamente a pesquisa, podemos garantir que os seus resultados serão válidos para este conjunto de elementos.[5]

---

[4] A análise comparativa de dois conjuntos de dados será tratada no Capítulo 11.

[5] Quando houver diferença razoável entre a população alvo e a população acessível, pode haver grande viés ao generalizar os resultados da análise para toda a população alvo. Nestes casos, é recomendável citar no relatório da pesquisa a limitação de que seus resultados valem especificamente para a população definida como acessível, evitando, assim, que seus resultados 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, numa determinada época*. Mas pode também ser definida como os *indivíduos moradores do bairro* ou, ainda, como *os 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 torna-se 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 forneça 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.[6]

O planejamento da coleta de dados

Definidos os objetivos e a população a ser estudada, precisamos pensar *como* deverá ser a coleta de dados. Em muitas situações não precisamos ir até os 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 *grau de instrução dos funcionários* talvez possam ser obtidos no departamento de pessoal desta 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, torna-se necessário construir um instrumento para que sua coleta

---

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

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) Você considera a pesquisa proposta no Exercício 1 como experimental ou de levantamento? Justifique.

2.2 DADOS E VARIÁVEIS

Vamos chamar de ***variáveis*** as características que podem ser observadas (ou medidas) em cada elemento da população, sob as mesmas condições. Uma variável observada (ou medida) num elemento da população deve gerar apenas um resultado. As variáveis surgem quando perguntamos *o quê* vamos observar ou medir nos elementos de uma população.

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 pensar em observá-las com perguntas do tipo:

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

Estas perguntas, contudo, não estão identificando bem as variáveis de interesse, pois os funcionários podem interpretá-las de diferentes formas e, por exemplo, para a primeira pergunta, podem ocorrer respostas tais como: *há pouco mais de 12 anos*, *há 7 meses*, *há muito tempo*, etc., 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, como, 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*, suas 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).

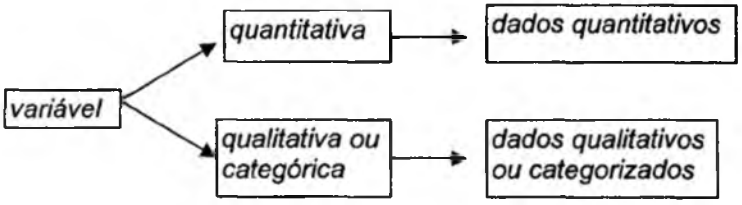


Figura 2.2 Classificação das variáveis e dos dados, em termos do nível de mensuração.

No exemplo precedente, o *tempo de serviço (em anos completos)* é uma variável quantitativa, enquanto o *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 grau 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 qualidades 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, onde o respondente vai atribuir um valor, que ele julga ser a sua 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.[7]

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.

---

[7] Uma terceira opção seria avaliar a característica *satisfação* indiretamente, considerando vários indicadores que medem esta característica, conforme alguma teoria sobre o assunto. Estes indicadores poderiam ser, por exemplo, adequação do salário, segurança no emprego, sentimento de auto-realização, sensação de autonomia, etc.

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.[8]

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

Exercícios

3) Defina variáveis 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, em 1992, 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 a serem levantadas.

---

[8] Uma outra possibilidade seria definir a variável *esportes 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.

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 grau de instrução dos funcionários e
- avaliar o grau 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*, *grau de instrução* e *grau de satisfação com o trabalho*.

b) Fazer uma revisão bibliográfica para verificar como mensurar adequadamente algumas características.

No exemplo precedente precisamos avaliar o grau de satisfação dos funcionários. Podemos procurar referências bibliográficas 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.[9]

c) Estabelecer a forma de mensuração das características (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 *grau de instrução*, em *categorias mutuamente exclusivas*, como: *nenhum grau completo*, *primeiro grau completo*, *segundo grau completo* e *superior completo*. O *grau de satisfação com o trabalho* pode ser avaliado de muitas formas diferentes. Uma destas formas poderia ser uma escala de cinco pontos, sendo 1 – *completamente insatisfeito*, 2 – *insatisfeito*, 3 – *mais ou menos satisfeito*, 4 – *satisfeito* e 5 – *completamente satisfeito*.

---

9 IBGE é a sigla da *Fundação Instituto Brasileiro de Geografia e Estatística*, órgão responsável por diversos levantamentos no Brasil, como os censos demográficos, censos agropecuários, censos industriais, anuários estatísticos, estudo nacional de despesas familiares, etc.

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

A característica *grau de satisfação com o trabalho* pode ser avaliada sob vários enfoques, como, por exemplo, 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) 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 este 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 destes objetivos, pois, *quanto mais longo o questionário, menor tende a ser a qualidade e 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 do entrevistador que as transcreve para a ficha. Nesta segunda situação, o entrevistador pode ou não interferir, sob forma de esclarecimento de algum item, anotando aspectos que julgar relevante, 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, pois o entrevistador pode esclarecer os diversos itens que estão sendo indagados.

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 este instrumento está bom. Neste 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: ambigüidade 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 91.1, com finalidades puramente acadêmicas.

O problema de pesquisa: A relação de um aluno universitário e o curso que está fazendo.

Objetivo geral: Num curso universitário, conhecer melhor a relação entre o aluno e o curso. Em particular, no curso de Ciências da Computação da UFSC.

Objetivos específicos:

1) Avaliar o grau de satisfação do aluno com o curso que está realizando.

2) Verificar se existe associação entre o grau 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 91.1.[10]

Amostra: Optamos por um processo rápido e fácil para a seleção da amostra. Tomamos três disciplinas obrigatórias das três últimas fases e aplicamos o questionário em sala de aula. A amostra foi, então, formada pelos alunos presentes nos dias de aplicação dos questionários.[11]

---

[10] Como se pretende avaliar a satisfação do aluno com o curso, a população deve ser formada por alunos que já conviveram com as diversas fases deste curso, donde a definimos como o conjunto de alunos que estavam cursando as três últimas fases.

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

Forma de mensuração das variáveis

Satisfação com o curso: é feita através da avaliação numérica, numa escala de 1 (um) a 5 (cinco), de acordo com o grau que o aluno julgar que melhor se adapte à sua satisfação com o curso, complementando com avaliações de aspectos específicos do curso, como corpo docente, recursos materiais disponíveis e é feito através do conteúdo curricular.

Desempenho do aluno: Índice de Aproveitamento Acumulado, calculado pela instituição, em função dos conceitos (ou notas) obtidos pelo aluno nas disciplinas cursadas. Então, os dados relativos a esta variável são dados secundários.

Aspectos positivos e *negativos do curso*: serão observados de duas maneiras: (1) avaliações numéricas, numa escala de 1 (um) a 5 (cinco), de acordo com o grau que o aluno julgar que melhor se adapte à sua concordância com alguns aspectos do curso e (2) deixar o aluno descrever livremente o principal aspecto positivo e negativo do curso. Nesta segunda situação, as categorias destas duas variáveis serão criadas após a realização de uma análise das respostas dos questionários, isto é, 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 o curso que você está realizando na UFSC? _____.

2) Qual a fase predominante em que você se encontra? _____.

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

   a) Didática dos professores de seu curso ..... (1 2 3 4 5)
   b) Grau de conhecimento dos professores ..... (1 2 3 4 5)
   c) Bibliografia disponível ..... (1 2 3 4 5)
   d) Laboratórios e outros recursos materiais ..... (1 2 3 4 5)
   e) Conteúdo dos programas das disciplinas oferecidas ..... (1 2 3 4 5)
   f) Encadeamento das disciplinas ..... (1 2 3 4 5)
   g) Satisfação com o curso, num sentido geral ..... (1 2 3 4 5)

4) Apresente o principal ponto positivo e negativo de seu curso, em termos do ensino ministrado.

   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 visão 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, enquanto que as respostas do subitem (g) serão usadas com vistas aos objetivos 1 e 2.

O item 4 procura complementar as informações do item 3, através de uma *pergunta aberta*.

O item 5 é uma medida de desempenho do aluno no curso, calculado pela instituição (índice de aproveitamento acumulado), para propósitos de matrícula. Como, em geral, os alunos não sabem o valor deste 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 serão usadas para, juntamente com outras informações, atingir o objetivo 2.[12]

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.

---

[12] 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.

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 grau mínimo e 5 o grau máximo, para as seguintes características relacionadas com você e seu curso.

   a) Didática dos professores de seu curso ..... (1 X 3 4 5)
   b) Grau de conhecimento dos professores..... (1 2 3 X 5)
   c) Bibliografia disponível ..... (1 X 3 4 5)
   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, em termos do ensino ministrado.

   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).

É comum armazenar os dados numa matriz (ou quadro), onde cada coluna se refere a uma variável e cada linha a um respondente.[13] A Tabela 2.1 mostra os dados armazenados dos cinco primeiros respondentes. Os dados observados do questionário que acabamos de mostrar estão na primeira linha desta tabela.

Tabela 2.1 Armazenamento dos dados de cinco respondentes.

	Item do questionário									
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

[13] 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.

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*, 9 – *Outros*. E para o item (4.b), *ponto negativo*, foram: 1 – *Professores*, 2 – *Laboratórios e Recursos Materiais*, 3 – *Currículo e Disciplinas*, 4 – *Aplicações*, 5 – *Atualização*, 9 – *Outros*.

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

ANEXO

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

	Item do questionário									X 5
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

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
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
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 deste alimento. Estamos fazendo uma *amostragem*, ou seja, extraindo do *todo* (população) uma *parte* (amostra), com o propósito de avaliarmos (*inferirmos*) a qualidade de tempero de todo o alimento.

Nas pesquisas científicas, em que se quer conhecer algumas características de uma população, também é muito comum observar-se apenas uma amostra de seus elementos e, a partir dos resultados dessa amostra, obter valores aproximados, ou *estimativas*, para as características populacionais de interesse. Este tipo de pesquisa é usualmente chamado de *levantamento por amostragem.*

Num levantamento por amostragem, a seleção dos elementos que serão efetivamente observados deve ser feita sob uma metodologia adequada, de tal forma que os resultados da amostra sejam informativos, para avaliar características de toda a população. 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, chamamos de ***população*** um conjunto de elementos passíveis de serem mensurados, com respeito às variáveis que se pretende levantar. A população pode ser formada por pessoas, famílias, estabelecimentos industriais, ou qualquer outro tipo de elementos, dependendo basicamente dos objetivos da pesquisa.

É comum termos interesse em descrever certas características específicas dos elementos da população, que denominaremos ***parâmetros.*** Veja os exemplos seguintes.

---

[1] Este capítulo teve a contribuição da Profª SÍLVIA MODESTO NASSAR (INE / CTC / UFSC).

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 pode ser definida como todos eleitores com domicílio eleitoral no município.[2] Os principais parâmetros devem ser *as percentagens de votos de cada candidato à prefeitura, no momento da pesquisa*.

Exemplo 3.3 Para planejar políticas de recursos humanos numa empresa, com milhares de funcionários, podemos realizar uma pesquisa para avaliar alguns parâmetros da população de funcionários desta empresa, tais como: *tempo médio de serviço dos funcionários na empresa, 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 típicas em que se recomenda utilizar amostragens. A Figura 3.1 ilustra uma pesquisa eleitoral, onde se tem o interesse na percentagem de votos de cada candidato (parâmetros).

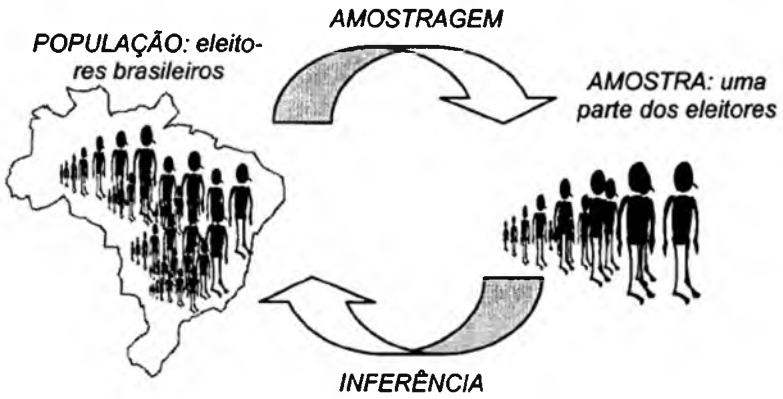


Figura 3.1 Pesquisa eleitoral: um caso típico de levantamento por amostragem.

---

[2] Na prática, a população acessível se restringe aos eleitores residentes no município.

O termo ***inferência estatística*** refere-se ao uso apropriado dos dados da amostra para se ter algum conhecimento sobre os parâmetros da população. Os valores calculados a partir dos dados da amostra, com o objetivo de avaliar parâmetros desconhecidos, são chamados de ***estimativas*** desses parâmetros. Numa pesquisa eleitoral, por exemplo, as percentagens de cada candidato, divulgadas antes da eleição, são, na verdade, *estimativas*.

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 este programa de treinamento.

Por que amostragem?

Citaremos quatro razões para o uso de amostragem em levantamentos de grandes populações.

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 é o controle dos entrevistadores.[3]

Quando o uso de amostragem não é interessante?

Citaremos três situações em que pode não valer a pena a realização de uma amostragem.

1) *População pequena.* Sob o enfoque de amostragens aleatórias que estudaremos neste capítulo, se a população for pequena (digamos, de 50 elementos) para termos uma amostra capaz de gerar resultados precisos para os parâmetros da população, necessitamos de uma amostra relativamente grande (em torno de 80% da população). Geralmente é mais relevante o tamanho absoluto da amostra do que a percentagem que

---

[3] O termo *censo* refere-se à pesquisa de toda a população.

ela representa na população. Voltemos à situação de verificar o tempero de um alimento em preparação. Desde que o alimento esteja bem mexido, uma amostra de uma colher é suficiente, independentemente de estarmos preparando uma pequena ou grande quantidade de alimento. Na Seção 3.4 voltaremos a discutir tamanho de amostra.

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 estas características tem-se o parâmetro *número de habitantes residentes no país*, que é fundamental para o planejamento do país. Desta forma, o parâmetro *número de habitantes* precisa ser avaliado com grande precisão e, por isto, se pesquisa toda a população.

Plano de amostragem

Para fazermos 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.[4] Os parágrafos seguintes tentam esclarecer melhor estes termos.

Para efetuar a seleção dos elementos que farão parte da amostra, precisamos estabelecer a ***unidade de amostragem***, ou seja, 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 e que, de alguma forma, estejam associadas aos 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

---

[4] Muitas vezes o termo *plano de amostragem* é usado para designar somente a técnica de *seleção* dos *elementos*.

(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*. Estas amostragens 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. Neste livro, daremos ênfase a estes tipos de amostragens.

Estudaremos, inicialmente, algumas formas de seleção dos elementos que irão compor a amostra. Posteriormente discutiremos a questão do tamanho da amostra.

3.1 AMOSTRAGEM ALEATÓRIA SIMPLES

Para a seleção de 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 de pertencer à amostra*.[5]

O uso de tabelas de números aleatórios

As tabelas de números aleatórios facilitam o processo de seleção de uma amostra aleatória. Estas tabelas são formadas por sucessivos sorteios

---

[5] Estas propriedades podem ser verificadas através do cálculo de probabilidades. A probabilidade de um particular elemento da população pertencer à amostra é dada por n/N.

de algarismos do conjunto {0, 1, 2,...,9}, com reposição. Uma destas tabelas encontra-se no apêndice, donde extraímos uma parte e apresentamos a seguir. Os espaços colocados a cada dois algarismos servem, apenas, para facilitar a visualização da tabela, não interferindo na sua utilização.

Números Aleatórios

98 08 62 48 26	45 24 02 84 04	44 99 90 88 96	39 09 47 34 07	35 44 13 18 80
33 18 51 62 32	41 94 15 09 49	89 43 54 85 81	88 69 54 19 94	37 54 87 30 43
80 95 10 04 06	96 38 27 07 74	20 15 12 33 87	25 01 62 52 98	94 62 46 11 71

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.[6]

POPULAÇÃO: funcionários da empresa

Aristóteles	Anastácia	Arnaldo	Bartolomeu	Bernardino
Cardoso	Carlito	Cláudio	Ermílio	Ercílio
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 Cesar			

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. Ercílio
11. Ernestino	12. Endevaldo	13. Francisco	14. Felício	15. Fabrício
16. Geraldo	17. Gabriel	18. Getúlio	19. Hiraldo	20. João da Silva
21. Joana	22. Joaquim	23. Joaquina	24. José da Silva	25.José de Souza
26. Josefa	27. Josefina	28. Maria José	29. Maria Cristina	30. Mauro
31. Paula	32. Paulo Cesar			

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. Usaremos, neste

[6] 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.

exemplo, a primeira linha, desprezando os valores que estiverem fora do conjunto {01, 02,...,32} e os valores que se repetirem.

Números aleatórios extraídos da tabela: 08 26 24 02 04.
Amostra: {Cláudio, Josefa, José da Silva, Anastácia, Bartolomeu}

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 destes valores, observado na amostra de funcionários, é chamado de *amostra da variável X*, conforme ilustrado a seguir:

Amostra de funcionários:

{Cláudio, Josefina, José da Silva, Anastácia, Bartolomeu}

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

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

Exercícios

1) Considerando a população do Exemplo 3.4, extraia uma amostra aleatória simples de $n = 10$ funcionários. Use a segunda linha da tabela de números aleatórios (Tabela I 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	Emílio	13	Ercílio	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 Cesar	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 1650 a 8840. 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 estas 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 amostra aleatória simples, por um processo bem mais rápido do que aquele que discutimos na seção anterior. Por exemplo, se queremos tirar uma amostra de 1.000 fichas, dentre uma população de 5.000 fichas, podemos tirar, sistematicamente, uma ficha a cada cinco ($^{5.000}/_{1.000} = 5$). 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, e a relação N/n é chamada de ***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. Ercílio
11. Ernestino	12. Endevaldo	13. Francisco	14. Felício	15. Fabrício
16. Geraldo	17. Gabriel	18. Getúlio	19. Hiraldo	20. João da Silva
21. Joana	22. Joaquim	23. Joaquina	24. José da Silva	25. José de Souza
26. Josefa	27. Josefina	28. Maria José	29. Maria Cristina	30. Mauro
31. Paula	32. Paulo Cesar			

Devemos sortear um elemento dentre os seis primeiros. Podemos fazer isto extraindo um número, de um algarismo, da tabela de números aleatórios. Tomaremos, para este exemplo, o primeiro número da segunda linha. O número é "3", ou seja, o primeiro funcionário da amostra é o "Arnaldo". Os demais são obtidos pelo intervalo de seleção "6", a partir do Arnaldo, resultando na seguinte amostra[7]:

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

Amostragem estratificada

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

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

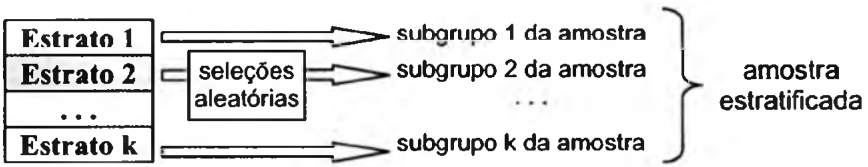


Figura 3.2 O processo de amostragem estratificada.

---

[7] 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 selação for grande (digamos, maior do que 10) a diferença será desprezível.

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.

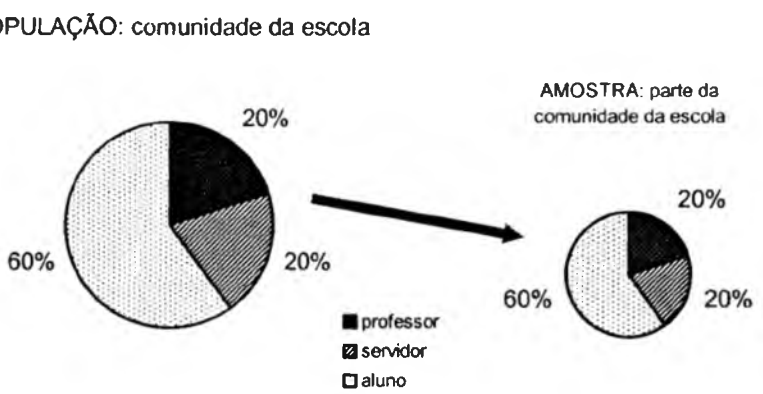


Figura 3.3 Ilustração de uma amostragem estratificada proporcional.

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

Exemplo 3.6 Com o objetivo de levantar 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%)	$n_p = (0,20).10 = 2$
Servidores	10/50 = 0,20 (ou 20%)	$n_s = (0,20).10 = 2$
Alunos	30/50 = 0,60 (ou 60%)	$n_a = (0,60).10 = 6$

Para selecionar aleatoriamente dois professores, usaremos a numeração já existente na população, substituindo o "10" por "0". Neste caso, podemos usar a tabela de números aleatórios, tomando valores com um algarismo. Usando, por exemplo, a primeira linha da tabela de números aleatórios (98 08...), temos os seguintes professores selecionados: {P9, P8}, correspondentes aos dois primeiros números desta linha.[8]

Para os servidores, usando a segunda linha da tabela de números aleatórios (33 18...), com o mesmo processo de numeração, temos: {S3, S1}.

Para os alunos, precisamos extrair números de dois algarismos. Usando a própria numeração da população e a terceira linha da tabela (80 95 10 04 06 96 38 27 07 74 20...), temos: {A10, A4, A6, A27, A7, A20}.

A amostra {P9, P8, S3, S1, A10, A4, A6, A27, A7, A20} 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 precisos, quando comparada com uma amostra aleatória simples.[9]

---

[8] Os números aleatórios foram extraídos da tabela de números aleatórios que se encontra no apêndice deste livro.

[9] No presente contexto, entende-se por resultados mais precisos aqueles que provavelmente estejam mais próximos dos parâmetros da população de onde foi extraída a amostra.

Amostragem estratificada uniforme: seleciona-se a mesma quantidade de elementos em cada estrato. No exemplo precedente, para selecionar uma amostra estratificada uniforme de, digamos, $n = 12$ indivíduos da comunidade da escola, devemos selecionar 4 indivíduos de cada categoria (Exercício 6).

A amostragem estratificada uniforme costuma ser usada em situações em que o maior interesse é obter estimativas separadas para cada estrato, ou ainda, 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 provêem 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, deve-se agregar os resultados de cada estrato por uma média aritmética ponderada, levando-se em consideração a proporcionalidade de cada estrato na população.[10]

Amostragem de conglomerados

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

Chamamos *conglomerado* a um grupamento de elementos da população. Por exemplo, numa população de domicílios de uma cidade, os quarteirões formam *conglomerados* de domicílios.

Este tipo de amostragem consiste, num primeiro estágio, em selecionar conglomerados de elementos. Num segundo estágio, ou se observam todos os elementos dos conglomerados selecionados no primeiro estágio (*amostragem de conglomerados em um estágio*), ou, como é mais comum, 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. Veja a Figura 3.4.

---

[10] Ver Cochran (1977).

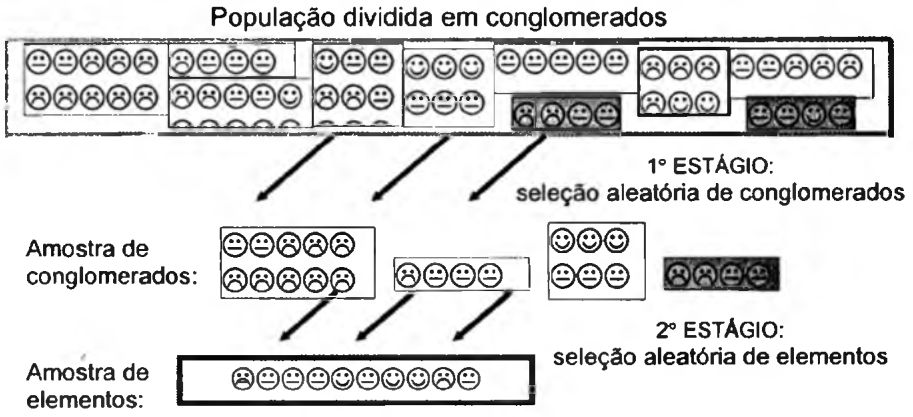


Figura 3.4 Ilustração do processo de amostragem de conglomerados em dois estágios.

Em algumas pesquisas em grande escala, a amostragem pode ser feita em mais estágios. Por exemplo, para selecionar uma amostra de domicílios do Estado de Santa Catarina, podemos, no primeiro estágio, selecionar municípios; no segundo estágio, selecionar quarteirões e, finalmente, no terceiro estágio, selecionar domicílios.

Chamamos de ***fração de amostragem*** a 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 cada elemento da população tem a mesma probabilidade de pertencer à amostra.

Exemplo 3.7 Considere o problema de selecionar uma amostra de domicílios de uma cidade. Podemos tomar as ruas como conglomerados, como indicado no quadro abaixo, 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, como exemplo, selecionar 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. Neste estágio, as unidades de amostragem são as ruas que vamos considerar numeradas, como segue: 1 = *A*, 2 = *B*, 3 = *C*, 4 = *D* e 5 = *E*. Tomemos, por exemplo, números da primeira linha da tabela de números aleatórios do apêndice (98 08 62 48 26 45). Os números grifados têm correspondência com as ruas, donde temos a amostra de conglomerados (ruas): *B*, *D* e *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 segunda linha da tabela de números aleatórios, e usando a própria numeração de identificação, chegamos nos domicílios *B9, B10, B4, B6, B7, B12* e *B1*.

Rua D. Tomando, por exemplo, a quarta linha da tabela de números aleatórios, sorteamos os domicílios *D2* e *D4*.

Rua E. Usando a quinta linha, sorteamos *E1, E8, E6* e *E3*.

Amostra selecionada: {*B9, B10, B4, B6, B7, B12, B1, D2, D4, E1, E8, E6, E3*}.

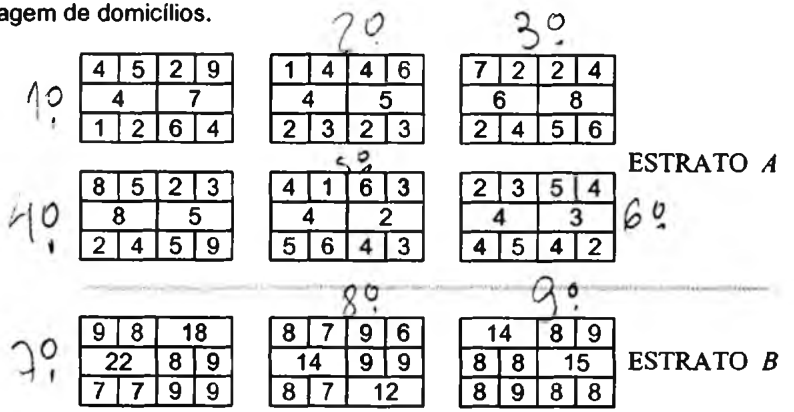
O leitor deve observar que, ao contrário dos planos discutidos anteriormente, a amostragem de conglomerados não exige uma lista de todos os elementos da população. Basta, no primeiro estágio, uma lista de conglomerados e, no segundo estágio, uma lista de elementos, mas somente para os conglomerados previamente selecionados. Por este aspecto, em pesquisas onde os elementos da população estão dispersos sobre grandes áreas territoriais, a amostragem de conglomerados torna-se muito mais econômica do que a aleatória simples.

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 observada numa amostragem de domicílios.



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

b) Extraia uma amostra aleatória de tamanho $n = 9$, através de 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 selecionados.

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 isto 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

Este tipo de amostragem assemelha-se, numa primeira fase, com a amostragem estratificada proporcional. A população é vista de forma segregada, dividida em diversos subgrupos. Seleciona-se, para fazer parte da amostra, 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, por exemplo, a população pode ser dividida por localidade, por nível de instrução, por faixas de renda, etc. Veja o Exercício 10 para saber como dividir a população com mais de uma variável estratificadora.

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.

Numa população deste tipo, a utilização de uma amostragem aleatória pode não ser recomendável, já que temos uma população pequena.[11] 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 objetivos a descrição de certas características da população. Em muitos casos, porém, o principal objetivo é comparar certas características em duas ou mais populações.

---

[11] A maioria das universidades brasileiras tem menos de 50 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.

Para se comparar, por exemplo, 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: 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, que tenha um setor especializado nesta doença, tomando-se todas as pessoas em tratamento. Obviamente esta 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 similares possível, a não ser o fato de que uma delas é formada por pessoas *com câncer no pulmão* e a outra, por pessoas que *não tenham câncer no pulmão*. Nestas duas amostras se estudaria e compararia o *hábito de fumar*.

Num estudo experimental, em que é possível controlar os elementos que vão pertencer a cada um dos grupos a serem comparados, a comparabilidade das amostras pode ser obtida, num primeiro momento, por uma *divisão aleatória* dos elementos entre os grupos. Por exemplo, 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.[12] 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.

---

[12] O sorteio pode ser feito usando uma tabela de números aleatórios. Veja o Exercício 5, Seção 3.1.

b) Com a finalidade de estudar o perfil dos consumidores de um supermercado, fez-se a coleta de dados durante um mês, tomando a cada dia, um consumidor da fila de cada caixa do supermercado, variando sistematicamente o horário da coleta dos dados.

c) Para avaliar a qualidade dos itens que saem de uma linha de produção, observaram-se todos os itens das 14 às 14 horas e 30 minutos.

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 o 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 estes exames, decidiu-se estratificar a população por nível escolar (1° grau e 2° grau) 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
1º grau	48%	14%
2º grau	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.[13] Também não abordaremos aspectos financeiros, mesmo sabendo que muitas vezes o tamanho da amostra fica restrito aos recursos disponíveis.

[13] Para outros tipos de amostragens aleatórias, o leitor pode consultar livros próprios de amostragens, como Cochran (1977). Veja *Referências Bibliográficas* no final do livro.

Outros pontos importantes na determinação do tamanho da amostra são a heterogeneidade da população em estudo e os tipos de parâmetros que se deseja estimar (proporções, médias, etc.). Estes ingredientes entrarão em fórmulas mais refinadas, as quais apresentaremos no Capítulo 9. Nesta seção, trataremos de uma formulação bastante genérica, usada em pesquisas em que se deseja estimar diversos parâmetros, especialmente proporções (ou percentagens) de ocorrência de determinados atributos.[14]

Alguns conceitos

Como já definimos, o termo ***parâmetro*** é usado para designar alguma característica descritiva dos elementos da população. De forma análoga, chamaremos de ***estatística*** alguma característica descritiva dos elementos da amostra.[15] Por exemplo, na população dos funcionários de uma empresa, a *percentagem de funcionários favoráveis a um programa de treinamento* é um parâmetro. Numa amostra a ser retirada de 200 destes funcionários, a *percentagem de favoráveis ao programa de treinamento*, nesta amostra, é uma estatística.

Ao observarmos efetivamente uma amostra de 200 funcionários, se encontrarmos 60% de favoráveis, este valor é chamado de *estimativa* do referido parâmetro. Então, uma ***estimativa*** é o valor acusado por uma certa estatística, considerando a particular amostra observada.

Chamamos de ***erro amostral*** a diferença entre o valor que a estatística pode acusar e o verdadeiro valor do parâmetro que se deseja 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 dos parâmetros de interesse. Por exemplo, na divulgação de pesquisas eleitorais, é comum encontrarmos no relatório algo como: *a presente pesquisa tolera um erro de 2%*. Isto quer dizer que, quando a pesquisa aponta determinado candidato com 20% de preferência do

---

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

[15] A estatística, quando usada para avaliar (ou estimar) o valor de um parâmetro, também é chamada de *estimador*.

eleitorado, está afirmando, na verdade, que a preferência por este candidato é um valor do intervalo de 18% a 22% (ou seja, 20% $\pm$ 2%).

A especificação do *erro amostral tolerável* deve ser feita sob um enfoque probabilístico, pois, por maior que seja a amostra, existe sempre o risco de o sorteio gerar uma amostra com características bem diferentes das da população de onde ela está sendo extraída. Contudo, este enfoque probabilístico será introduzido somente no Capítulo 9. Por ora, deixaremos num sentido coloquial certas expressões, tais como: *provavelmente, com alto nível de confiança,* etc.[16]

Uma fórmula para o cálculo do 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 conhecer o tamanho da população, através da seguinte expressão:

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

Conhecendo o tamanho N da população, podemos corrigir o cálculo anterior, por:

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

Exemplo 3.8 Planeja-se um levantamento por amostragem para avaliar diversas características da população das $N = 200$ famílias moradoras de um certo bairro. Estas características (parâmetros) são especialmente do tipo *percentagens*, tais como, a *percentagem de famílias que usam programas de alimentação popular, a percentagem de famílias que moram em casas próprias,* etc. Qual deve ser o tamanho mínimo de uma amostra aleatória

[16] Para o leitor que já tenha algum conhecimento de Estatística, observamos que a formulação ora apresentada baseia-se na estimação de uma proporção, no caso de maior heterogeneidade, sob o nível de confiança de 95% (aproximado).

simples, tal que possamos admitir, com alta confiança, que os erros amostrais não ultrapassem 4% ($E_0 = 0{,}04$) ?

Solução. Uma primeira aproximação: $n_0 = \dfrac{1}{(0{,}04)^2} = 625$ famílias [1]

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

$$n = \frac{(200)\cdot(625)}{200+625} = \frac{125000}{825} = 152 \text{ famílias}$$

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 toda o município, que contém N = 200.000 famílias residentes?

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

$$n = \frac{(200000).(625)}{200000+625} = 623 \text{ famílias}$$

No último exemplo, vimos que a correção com o tamanho N da população, praticamente não alterou o cálculo inicial do tamanho da amostra ($n_0 = 625$ e $n = 623$). Em geral, se a população for muito grande (digamos, dezenas de milhares de elementos), o cálculo do tamanho da amostra pode ser feito pela primeira expressão:

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

sem levar em conta o tamanho exato, N, da população.

Podemos observar, também, que, para se manter o mesmo erro amostral, no Exemplo 3.8 foi necessária uma amostra abrangendo 76% da população (152 elementos extraídos de 200); enquanto que no Exemplo 3.9 foi suficiente uma amostra de apenas 0,3% da população (623 de 200.000). É, portanto, errônea a idéia de que para uma amostra ser representativa ela deva abranger uma percentagem fixa da população (veja a Figura 3.5).

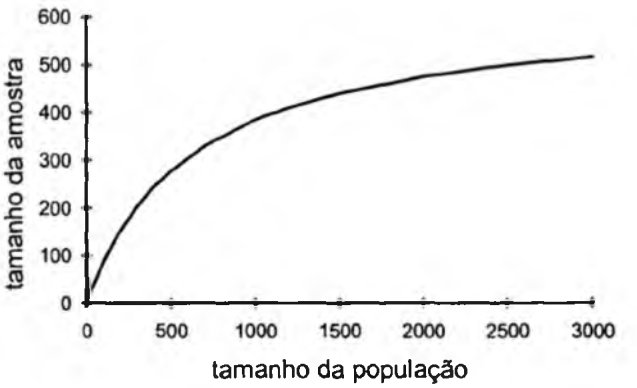


Figura 3.5 Relação entre tamanho da população e tamanho da amostra.

Tamanho da amostra em subgrupos da população

É muito 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 ter como segundo objetivo um estudo isolado de determinados bairros da cidade, e assim por diante.

Quando queremos efetuar estimativas sobre partes da população, precisamos calcular o tamanho da amostra para cada uma destas partes. O tamanho total da amostra vai corresponder à soma dos tamanhos das amostras de cada parte.

Podemos notar, pelo exposto acima, que o tamanho total da amostra deve crescer bastante quando se desejam estimativas isoladas para diversos subgrupos da população. Neste sentido, é comum o pesquisador não ser muito exigente na precisão das estimativas nos subgrupos, tolerando erros amostrais maiores.

Exemplo 3.10 Considerando o Exemplo 3.9, suponha que se deseje fazer estimativas isoladas para os seguintes estratos: (1) centro da cidade, (2) bairros e (3) periferia, mantendo-se a mesma precisão para cada estrato ($E_0 = 0{,}04$).
Neste caso, seriam necessárias:

$$n = \frac{1}{E_0^2} = \frac{1}{(0{,}04)^2} = 625 \quad \text{famílias em cada estrato}$$

e, portanto, a amostra total, deve conter: $n_{total} = 3.(625) = 1.875$ famílias.

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

Exercícios

11) Numa pesquisa, para estudar a preferência do eleitorado a uma semana da eleição presidencial, qual o tamanho de uma amostra aleatória simples de eleitores que garanta, com alta confiança, um erro amostral não superior a 2%?

12) 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 alto nível de confiança, 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 a partir de 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. Na prática, devido a uma série de razões, isto geralmente não acontece.

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, usado no cálculo do tamanho da amostra. Por isto, 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 (*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 deste município. Isto 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 isto não for possível, devemos limitar a abrangência da pesquisa à população que foi efetivamente estudada.

Falta de resposta

É comum não conseguirmos respostas de alguns elementos selecionados na amostra. Isto ocorre freqüentemente quando a população em estudo é a humana, pois, nem todos se dispõem a responder um questionário ou dar uma entrevista. O entrevistador, eticamente e respeitando o direito do entrevistado em não participar, deve ter uma 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 este problema, devemos efetuar vários retornos a estes elementos.

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 deste 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 estes erros.

Além destes 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.

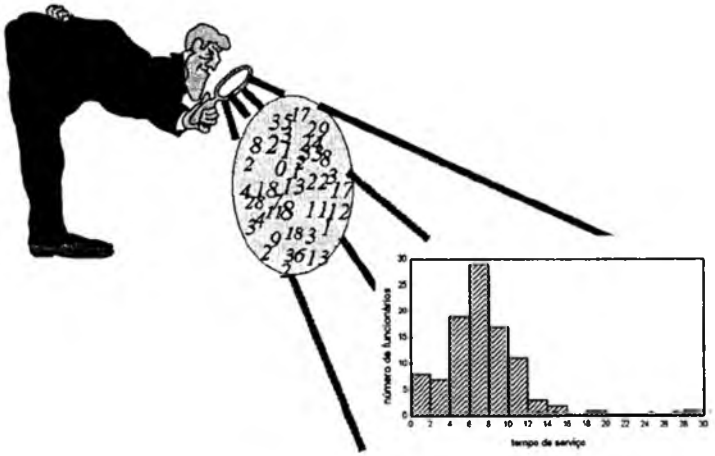
b) Retire desta população uma amostra aleatória estratificada proporcional de tamanho n = 12, usando o sexo como variável estratificadora. Use a segunda coluna da tabela de números aleatórios para o estrato dos homens e a terceira coluna para o estrato das mulheres.

c) Se o estudo tem por objetivo avaliar o tipo de brincadeira preferida por cada criança, qual 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 entre os empregados para verificar o grau de satisfação em relação à qualidade da comida no refeitório. Apresente um plano de amostragem para o presente problema.

Parte II

Descrição e exploração de dados



- Como extrair informações dos dados
- Como construir e apresentar tabelas, gráficos e medidas descritivas

Capítulo 4

Dados categorizados

Neste capítulo e nos dois seguintes, vamos considerar que os dados já foram efetivamente observados, sejam de uma amostra ou de uma população de elementos. E o objetivo básico consistirá em introduzir técnicas que permitam organizar, resumir e apresentar estes 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 começarmos a delinear hipóteses a respeito da estrutura do fenômeno em estudo. É a chamada *Análise Exploratória de Dados.*

No presente capítulo, aprenderemos a descrever e explorar dados de *variáveis qualitativas*, ou seja, aquelas cujos possíveis resultados são observados na forma de categorias. É o caso de variáveis como *grau de instrução, sexo, estado civil*, etc. Por exemplo, ao observar a variável *sexo*, num conjunto de indivíduos, estaremos classificando cada indivíduo ou 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 freqüências. A ***distribuição de freqüências*** compreende a organização dos dados de acordo com as ocorrências dos diferentes resultados observados. Ela pode ser apresentada sob forma tabular ou gráfica.

O Quadro 4.1 apresenta dados, em forma de códigos, da variável *grau de instrução do chefe da casa*, de uma amostra de 40 famílias. Estes dados fazem parte do anexo deste capítulo e serão usados para ilustrar algumas técnicas.

Quadro 4.1 Dados sobre o grau de instrução do chefe da casa, numa amostra de 40 famílias do conjunto residencial Monte Verde, Florianópolis – SC, 1988.

Códigos: 1 - nenhum grau de instrução completo; 2 - primeiro grau completo; e 3 - segundo grau completo.
Resultados observados em cada família: 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 freqüências com dados de uma variável qualitativa, basta *contar* a quantidade de resultados observados em cada categoria. A Tabela 4.1 mostra a distribuição de freqüências dos dados do Quadro 4.1.[1]

Tabela 4.1 Distribuição de freqüências do grau de instrução do chefe da casa, numa amostra de 40 famílias do conjunto residencial Monte Verde, Florianópolis – SC, 1988.

Grau de Instrução[1]	Freqüência	Percentagem
nenhum	6	15,0
primeiro grau	11	27,5
segundo grau	23	57,5
Total	40	100,0

[1] As categorias correspondem ao último grau de instrução oficialmente completado.

Fonte: Veja anexo, final deste capítulo.

[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, se vier a contribuir para uma melhor leitura dos dados em 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 texto somente deve ser feita após ela ser referenciada no texto.

A primeira coluna da Tabela 4.1 mostra todas as categorias previamente estabelecidas da variável *grau de instrução*. A segunda coluna resulta da *contagem* de quantas observações se identificam com cada categoria. São as freqüências observadas. Finalmente, a terceira coluna apresenta uma medida relativa da freqüência de cada categoria. Estas *percentagens* são obtidas dividindo-se a freqüência de cada categoria pelo número total de observações e, em seguida, multiplicando-se por 100 (cem). Estas medidas relativas são particularmente importantes para comparar distribuições de freqüências.

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

Tabela 4.2 Distribuição de freqüências do grau de instrução do chefe da casa, numa amostra de 120 famílias, dividida segundo as localidades do bairro Saco Grande II, Florianópolis – SC, 1988.

Grau de Instrução[1]	Localidade		
	Monte Verde	Pq. da Figueira	Encosta do Morro
nenhum	6 (15,0)	14 (32,6)	18 (48,7)
primeiro grau	11 (27,5)	14 (32,6)	13 (35,1)
segundo grau	23 (57,5)	15 (34,8)	6 (16,2)
Total	40 (100,0)	43 (100,0)	37 (100,0)

[1] As categorias da variável *grau de instrução* correspondem ao último grau de instrução oficialmente completado.
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, predominando o segundo grau completo. Por outro lado, temos nas famílias pesquisadas na Encosta do Morro o pior

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

perfil, em termos de grau de instrução do chefe da casa, com quase 50% deles não tendo concluído nem o primeiro grau.[3]

O leitor deve notar que, ao organizar e resumir os dados numa distribuição de freqüências, exclui-se a informação de quais elementos pertencem a cada categoria. No presente exemplo, a informação de quais famílias pertencem a cada categoria parece ser irrelevante para entender o comportamento geral da variável *grau de instrução do chefe da casa*. Em situações como esta, as distribuições de freqüências constituem um instrumento bastante útil na descrição e exploração de dados observados.

Exercícios

1) Com base nos dados do anexo deste capítulo, construa uma tabela de freqüências 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 freqüências para a variável PAP (ver anexo), para cada localidade em estudo. Apresente estas distribuições numa tabela de dupla entrada e interprete.

3) Considerando os resultados da pesquisa descrita na Seção 2.4, cujos dados estão no anexo do Capítulo 2, faça uma distribuição de freqüências 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. Elas constituem-se numa forma alternativa de apresentação de distribuições de freqüências.

Nesta seção, apresentaremos o gráfico de barras e o gráfico de setores, que são particularmente importantes na representação de distribuições de freqüências de dados categorizados.

---

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

Gráfico de barras

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

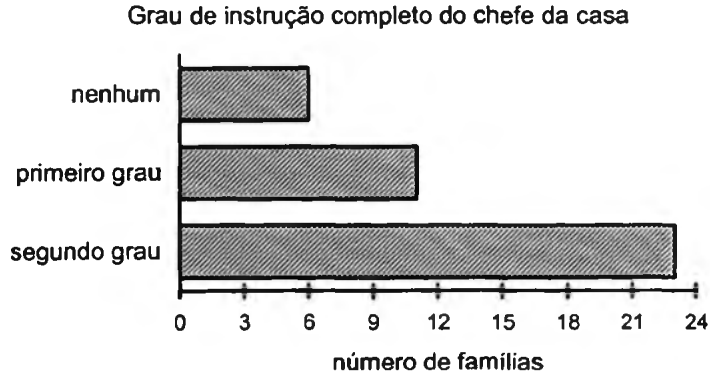


Figura 4.1 Distribuição de freqüências do grau de instrução do chefe da casa, numa amostra de 40 famílias do Conjunto Residencial Monte Verde, Florianópolis – SC, 1988.

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

Gráfico de setores

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

---

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

$$\frac{a_1}{360^0} = \frac{6}{40}$$

Relação entre o tamanho do setor (a_1) com o círculo todo (360^0).

Relação entre a freqüência da categoria (6) com o total observado (40).

Donde: $a_1 = \frac{6}{40}(360) = 54^0$

Repetindo este procedimento para as três categorias, temos:

categoria 1 (*nenhum*): setor de tamanho $a_1 = 54^0$;
categoria 2 (*primeiro grau*): setor de tamanho $a_2 = 99^0$;
categoria 3 (*segundo grau*): setor de tamanho $a_3 = 207^0$.

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

Figura 4.2 Distribuição de freqüências do grau de instrução do chefe da casa, numa amostra de 40 famílias do Conjunto Residencial Monte Verde, Florianópolis – SC, 1988.

Em se tratando da descrição de dados de variáveis ordinais, como no presente caso, deve-se dar preferência aos gráficos de barras ou de colunas, mantendo-se 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 freqüências da Tabela 4.2. No eixo horizontal, optou-se por colocar as freqüências relativas, em forma de percentagens, para facilitar a comparação.

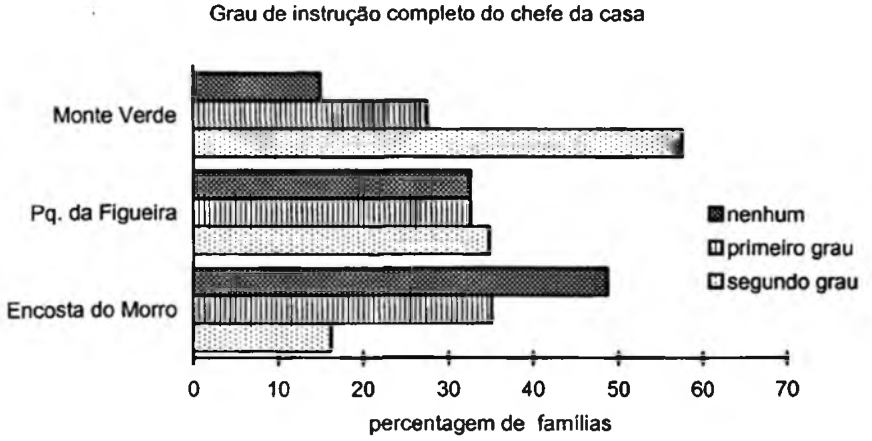


Figura 4.3 Distribuição de freqüências do grau de instrução do chefe da casa, numa amostra de 120 famílias, dividida segundo as localidades do bairro Saco Grande II, Florianópolis – SC, 1988.

Exercícios

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

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

4.3 DUPLA CLASSIFICAÇÃO

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

É muito freqüente, nas Ciências Sociais e Humanas, o interesse em verificar se duas variáveis se apresentam 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* nos indivíduos (ou famílias) pesquisados. Este tipo de análise passa pelas distribuições conjuntas de freqüências, que geralmente são apresentadas nas chamadas *tabelas de contingência* ou *tabelas de dupla entrada*, como veremos a seguir.

Para construirmos uma distribuição conjunta de freqüências, devemos observar simultaneamente as duas variáveis nos elementos em estudo. O esquema seguinte mostra a construção de uma distribuição conjunta, com as variáveis *grau de instrução do chefe da casa* e *uso de programas de alimentação popular*.

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

Dados		
família	grau 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	grau de instrução 1	2	3
1		|	|
0		|	||

Para a construção da distribuição conjunta de freqüências numa tabela de contingência, cada elemento (família) deve pertencer a uma e apenas uma casela.[5] Fazendo a classificação de todas as famílias observadas e contando as freqüências em cada casela, chegamos à Tabela 4.3. O leitor deve notar que os totais das colunas formam a distribuição de freqüências da variável *grau de instrução do chefe da casa*, quando observada

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

isoladamente, enquanto os totais das linhas constituem a distribuição da variável *uso de programas de alimentação popular.*

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

Uso de	Grau de instrução compl. do chefe da casa			
programas	nenhum	primeiro grau	segundo grau	Total
sim	31	22	25	78
não	7	16	19	42
Total	38	38	44	120

Para facilitar a análise de uma tabela de contingência, podemos incluir freqüências relativas, que podem ser calculadas em relação aos totais das linhas ou colunas, dependendo do objetivo.

A Tabela 4.4 mostra a Tabela 4.3 acrescida de percentagens em relação aos totais das colunas. Esta tabela evidencia os perfis do uso de programas de alimentação popular, considerando as famílias separadas por grau de instrução do chefe da casa (*perfis coluna*).

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

Uso de	Grau de instrução compl. do chefe da casa			
programas	nenhum	primeiro grau	segundo grau	Total
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 – Os dados da amostra parecem sugerir uma associação entre o uso de programas de alimentação popular e o grau de instrução do chefe da casa, pois, enquanto que 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 estes 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 grau de instrução do chefe da casa (*perfis linha*), considerando a amostra dividida em famílias que usam e famílias que não usam os programas. A interpretação da Tabela 4.5 é deixada para o leitor.

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

Uso de programas	Grau de instrução compl. do chefe da casa			
	nenhum	primeiro grau	segundo grau	Total
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.

Na Seção 4.1, quando discutíamos classificação simples, juntamos três distribuições de freqüências da variável *grau de instrução do chefe da casa*, correspondentes a três localidades diferentes (Tabela 4.2). Observamos, agora, que este tipo de tabela também pode ser analisada como uma tabela de contingência, como apresentado nesta seção, mesmo que na sua construção não tenhamos observado simultaneamente as duas variáveis, pois a *localidade* estava previamente estabelecida.

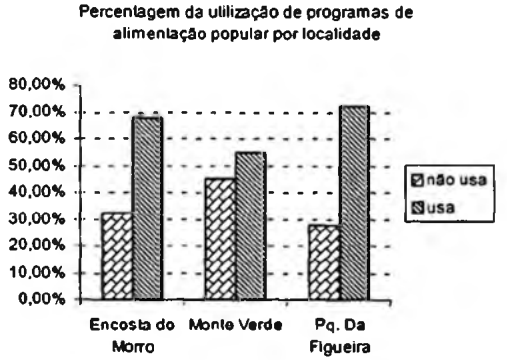
Uso do computador

Com o uso de programas computacionais de estatística, ou mesmo com planilhas eletrônicas, as tabelas e gráficos podem ser feitos com relativa facilidade. Segue uma tabela e um gráfico feitos com o auxílio do

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

Microsoft Excel, versão 1997, utilizando os dados sobre localidade e uso de programas de alimentação popular do anexo.[7]

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%



A apresentação adequada e a interpretação da tabela e do gráfico, deixamos como exercício para o leitor.

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 amostra observada 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 as tabelas.

---

[7] No *Excel*, entrar em *Dados* (menu principal) e *Relatório da Tabela Dinâmica*. O uso de percentagens é uma opção. Para fazer o gráfico de colunas, entrar na opção de gráficos.

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

Participação religiosa	Uso de bebidas alcoólicas	
	sim	não
freqüentemente	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, uma 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 esta 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	200
alto	150	300

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

Renda familiar X_1	Nível de instrução X_2	Uso de programas X_3	
		sim	não
1 baixa	1 baixo	320	80
	2 alto	80	20
2 alta	1 baixo	30	70
	2 alto	120	280

possui 3 variáveis

a) Qual a sua conclusão sobre a associação entre o *grau 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 muda em sua conclusão?

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 verificou-se 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/regular*, 2 = *bom*, 3 = *ótimo*) 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 a distribuição de freqüências de cada variável e a apresente em gráficos apropriados.

b) Construa a distribuição de freqüências conjunta. Apresente esta distribuição numa tabela de dupla entrada, calculando percentagens que enfatizam a distribuição do 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 a respeito da coleta seletiva de lixo. Uma das tabelas é apresentada a seguir:

Sistema de coleta seletiva de lixo

Grau de instrução do respondente	conhece		colabora	
	sim	não	sim	não
nenhum grau compl.	12	9	9	10
primeiro grau completo	23	3	16	15
segundo grau completo	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 suas principais informações.

ANEXO

Este anexo contém parte dos dados de entrevistas realizadas em famílias residentes na Região do Saco Grande II, Florianópolis – SC, 1988. 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.

g.i. (grau de instrução do chefe da casa): 1 = nenhum grau oficialmente completo;
2 = primeiro grau completo;
3 = segundo grau completo.

tam. (número de pessoas residentes no domicílio).

renda (renda familiar mensal, em quantidades de salários mínimos).

DADOS OBSERVADOS (120 famílias)

Nº	local	p.a.p.	g.i.	tam.	renda	Nº	local	p.a.p.	g.i	tam.	renda
1	1	0	3	4	10,3	19	1	0	3	4	5,1
2	1	0	3	4	15,4	20	1	1	3	4	12,2
3	1	1	2	4	9,6	21	1	1	3	5	5,8
4	1	0	2	5	5,5	22	1	1	3	5	12,9
5	1	1	3	4	9,0	23	1	0	3	5	7,7
6	1	1	1	1	2,4	24	1	0	2	4	1,1
7	1	0	3	2	4,1	25	1	0	2	8	7,5
8	1	1	3	3	8,4	26	1	1	3	4	5,8
9	1	1	3	6	10,3	27	1	1	1	5	7,2
10	1	1	2	4	4,6	28	1	0	3	3	8,6
11	1	0	2	6	18,6	29	1	1	2	4	5,1
12	1	1	1	4	7,1	30	1	0	3	5	2,6
13	1	0	2	4	12,9	31	1	1	3	5	7,7
14	1	0	2	6	8,4	32	1	1	2	2	2,4
15	1	0	3	3	19,3	33	1	1	3	5	4,8
16	1	0	2	5	10,4	34	1	1	1	2	2,1
17	1	1	3	3	8,9	35	1	1	1	6	4,0
18	1	0	3	4	12,9	36	1	1	1	8	12,5

continua ...

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

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

Capítulo 5

Dados quantitativos

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

5.1 VARIÁVEIS DISCRETAS

As variáveis que só assumem valores que podem ser listados são chamadas de ***variáveis discretas***. *Número de filhos de um casal* e *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 que podem assumir qualquer valor num intervalo são ditas ***variáveis contínuas***. O *peso de um indivíduo*, por exemplo, é uma variável contínua, pois o peso de um indivíduo pode ser qualquer valor no intervalo de, digamos, 0 a 300 kg.

As variáveis discretas geralmente resultam de alguma *contagem*, enquanto as contínuas costumam vir de uma *mensuração* propriamente dita.

A construção de distribuições de freqüências de dados resultantes de variáveis discretas, quando não houver grande quantidade de diferentes valores observados, pode ser feita da mesma forma que uma distribuição de freqüências de dados categorizados.[1] Como exemplo, usaremos os dados da variável *número de pessoas residentes no domicílio*, considerando uma

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

amostra de 40 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 freqüências destes dados construída através da contagem das repetições de cada resultado (ou valor) observado.

Tabela 5.1 Distribuição de freqüências do número de pessoas residentes no domicílio, numa amostra de 40 residências do Conjunto Residencial Monte Verde, Florianópolis – SC, 1988.

Número de pessoas	Freqüência 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

40 = amosha

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

A Figura 5.1 mostra duas formas alternativas de representação gráfica da distribuição de freqüências da Tabela 5.1. A primeira (Figura 5.1a) consiste em traçar riscos verticais sobre os valores efetivamente observados. A altura de cada risco deve ser proporcional à freqüência observada do correspondente valor. Na segunda representação (Figura 5.1b)

substituímos os riscos por retângulos. Estes retângulos devem ter a mesma largura e recomenda-se que sejam justapostos. O eixo vertical (das freqüências) deve sempre iniciar no zero e o eixo horizontal (dos valores da variável) pode iniciar próximo ao menor valor da variável.[2]

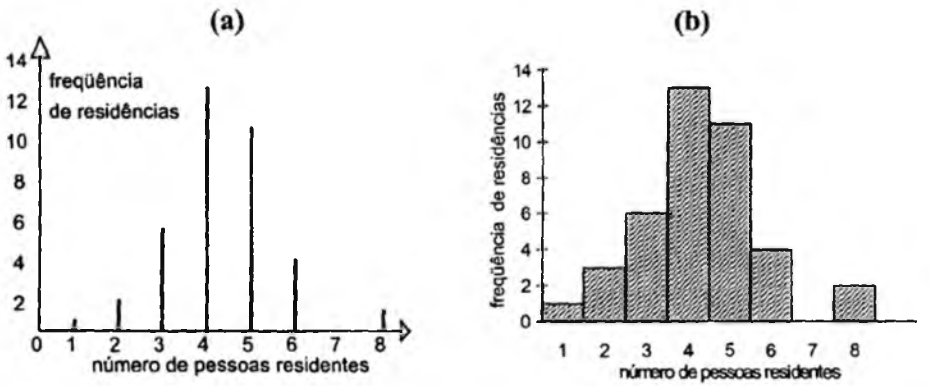


Figura 5.1 Representações gráficas da distribuição de freqüências da Tabela 5.1.

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 freqüências para o *nível de satisfação do aluno com o curso* (item 3.g do questionário);
b) apresente esta distribuição sob forma gráfica e
c) interprete.

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

---

[2] Num relatório, devemos optar em apresentar a distribuição ou numa tabela, ou num gráfico. Mas devemos lembrar que qualquer que seja a representação, esta deve vir acompanhada de um título completo, tal como na Tabela 5.1.

Distribuição do número de filhos dos pais dos respondentes

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

Apresente estas 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 dos dados observados.

Diagrama de pontos

Quando temos um conjunto com poucos dados, podemos analisá-lo através de um diagrama de pontos, isto é, fazendo com que cada resultado se identifique com um ponto na reta de números reais. A Figura 5.2 ilustra este diagrama com as taxas de crescimento demográfico dos municípios da Microrregião do Litoral do Itajaí.[3]

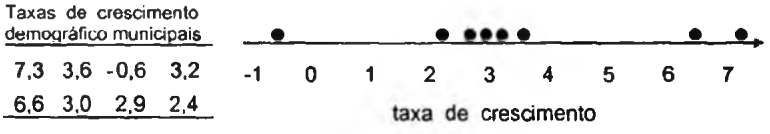


Figura 5.2 Os dados e o diagrama de pontos das taxas médias de crescimento demográfico, no período de 1970 a 1980, dos oito municípios da Microrregião do Litoral de Itajaí – SC.

[3] Os valores correspondem às taxas médias geométricas de incremento anual, 1970/80, das populações residentes dos oito municípios da Microrregião do Litoral do Itajaí. (Fonte: GAPLAN – SC e IBGE). Sobre média geométrica consultar Wonnacott, T. H. e Wonnacott, R. J. (1981).

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.

Taxas de crescimento demográfico municipais

Microrregião Serrana						Microrregião do Litoral do Itajaí			
-0,4	-1,7	-1,0	0,3	-0,3	-0,4	7,3	3,6	-0,6	3,2
-0,1	-1,2	-0,1	-0,4	0,6	0,4	6,6	3,0	2,9	2,4

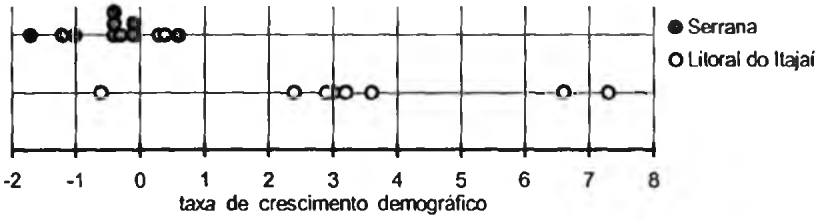


Figura 5.3 Diagrama de pontos das taxas médias de crescimento demográfico, 1970-80, dos municípios das Microrregiões Serrana e Litoral de Itajaí – SC.

Interpretação da Figura 5.3 – Os municípios do Litoral de Itajaí apresentam, em geral, taxas de crescimento demográfico maiores do que os municípios da Microrregião Serrana. Nesta segunda microrregião, a maioria dos municípios apresentam taxas negativas de crescimento populacional, enquanto que no Litoral de Itajaí, apenas um município apresenta taxa negativa. Também observamos que os dois grupamentos de municípios se diferenciam quanto à dispersão dos valores. Enquanto na Microrregião Serrana os municípios apresentam taxas de crescimento bem próximas, caracterizando uma relativa homogeneidade, no Litoral de Itajaí as taxas de crescimento populacional variam bastante de município para município.[4]

[4] A interpretação torna-se mais interessante quando se colocam algumas informações complementares, como, por exemplo, as atividades econômicas das duas microrregiões. Enquanto os municípios do Litoral do Itajaí têm no turismo e na pesca suas principais fontes de renda, nos municípios da Microrregião Serrana predominam as atividades rurais em pequenas propriedades agrícolas.

Tabela de freqüências

Nas Ciências Sociais, geralmente trabalhamos com conjuntos de centenas ou milhares de observações, onde o diagrama de pontos torna-se impraticável. Nestes casos, podemos construir distribuições de freqüências, grupando 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.

Usaremos, como exemplo ilustrativo, os dados da variável *taxa de mortalidade infantil dos 34 municípios da Microrregião Oeste Catarinense, ano de 1982.*[5]

Dados

32,3	62,2	10,3	22,0	13,1	9,9	11,9	20,0	36,4	23,5	18,0	22,6
20,3	38,3	19,6	27,2	28,9	18,4	27,3	21,7	23,7	13,9	36,3	32,9
29,7	25,4	23,8	15,7	17,0	39,2	22,7	29,9	18,3	33,0		

Considerando que todos os valores estão no intervalo de 9,9 a 62,2, devemos definir um conjunto de classes mutuamente exclusivas, tais que, quando reunidas, elas contenham este intervalo. Uma possível escolha seria construir 7 (sete) classes com amplitude aproximada de 10 (dez), como segue: de 0,0 a 9,9; de 10,0 a 19,9; ...; de 60,0 a 69,9. Para simplificar a notação, representaremos estas classes por: 0,0 |— 10,0; 10 |— 20; ...; 60 |— 70; onde o símbolo "|—" significa o intervalo entre os dois valores, incluindo o valor do lado esquerdo e excluindo o valor do lado direito.

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

---

[5] Observamos que a *taxa de mortalidade infantil* corresponde ao número médio de mortes, dentre 1000 crianças nascidas vivas, antes de completarem um ano de vida. Os dados foram extraídos da publicação *Municípios Catarinenses – Dados Básicos*, 1987, GAPLAN – SC, que utiliza-se dos dados levantados pelo IBGE.

classes	contagem	freqüência
0 ⊢ 10	|	1
10 ⊢ 20	|||||	10
20 ⊢ 30	||||| ||||| |||||	15
30 ⊢ 40	||||| ||	7
40 ⊢ 50		0
50 ⊢ 60		0
60 ⊢ 70	|	1

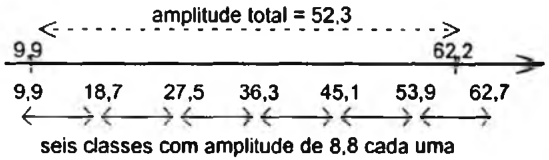
Na apresentação de uma tabela de freqüências, é comum colocar também os *pontos médios* das classes, isto é, para cada classe, calcular a média dos seus limites. Por exemplo, na classe 0 ⊢ 10, tem-se o ponto médio 5 (pois, cinco é a média entre zero e dez). O ponto médio representa o *valor típico* da classe, que, em muitas vezes, poderá ser usado para aproximar os demais valores da classe, como veremos no Capítulo 6. A Tabela 5.2 apresenta a distribuição de freqüência dos dados em discussão.

Tabela 5.2 Distribuição de freqüências das taxas de mortalidade infantil dos municípios da Microrregião Oeste Catarinense, 1982.

taxa de mortalidade infantil	ponto médio	freqüência de municípios	percentagem de municípios
0 ⊢ 10	5	1	2,9
10 ⊢ 20	15	10	29,4
20 ⊢ 30	25	15	44,2
30 ⊢ 40	35	7	20,6
40 ⊢ 50	45	0	0,0
50 ⊢ 60	55	0	0,0
60 ⊢ 70	65	1	2,9
Total	-	34	100,0

O número de classes a ser usado na tabela de freqüências é uma escolha arbitrária. Quanto maior o conjunto de dados, pode-se usar mais classes. 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 freqüências.

Em geral, usam-se de 5 (cinco) a 20 (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 observados.[6] Em nosso exemplo: n = 34, donde $\sqrt{34} \approx 6$. Como os dados estão compreendidos entre 9,9 e 62,2, ou seja, numa amplitude total de 62,2 - 9,9 = 52,3, para que todas classes tenham o mesmo tamanho, elas devem ter amplitude: $\frac{52,3}{6} \approx 8,8$ (na presente situação é conveniente arredondar *para cima*). Esquematicamente:



Resultando a seguinte tabela de freqüências:

classes	freqüências
9,9 |– 18,7	10
18,7 |– 27,5	13
27,5 |– 36,3	6
36,3 |– 45,1	4
45,1 |– 53,9	0
53,9 |– 62,7	1

A leitura de uma tabela com estas classes torna-se um pouco mais cansativa, comparada com a Tabela 5.2. Esta sugestão do número de classes precisa ser adaptada quando existem valores discrepantes no conjunto de dados. Nestes casos, normalmente isolam-se os valores discrepantes e refazem-se as classes.

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

Histograma

A Figura 5.4 mostra um histograma, 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 freqüência observada da correspondente classe.[7]

---

[6] Ressalta-se que é apenas uma sugestão!

[7] Quando as classes não têm a mesma amplitude, torna-se necessário fazer alguns ajustes. Veja, por exemplo, Bussab e Morettin (1985, p.18). O histograma também poderia ser feito usando percentagens, no eixo vertical, mas a sua forma não mudaria.

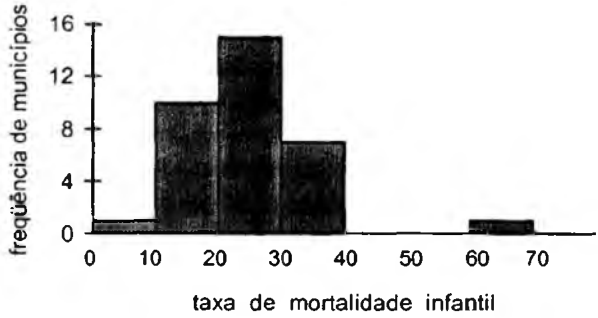
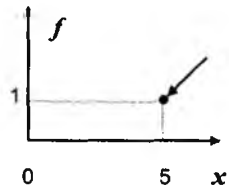


Figura 5.4 Distribuição de freqüências das taxas de mortalidade infantil dos 34 municípios da Microrregião Oeste Catarinense, 1982.

Interpretação da Figura 5.4 – Observamos uma predominância de municípios com taxas de mortalidade infantil na faixa de 10 a 30. Observamos, também, um município apontando taxa de mortalidade infantil extremamente alta, quando comparada às demais.[8]

Polígono de freqüências

O polígono de freqüências é uma representação gráfica alternativa. Para construí-lo, toma-se o ponto médio (x) e a correspondente freqüência (f) de cada classe. Colocam-se os pares (x, f) como pontos num par de eixos cartesianos. A ilustração ao lado mostra a representação do ponto (5, 1) num par de eixos cartesianos. Para completar o gráfico, devemos unir estes pontos com semi-retas, ligando os pontos extremos ao eixo horizontal.



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

---

[8] Como temos um ponto que se distancia dos demais, poderíamos considerar um maior número de classes, a fim de evidenciar melhor a distribuição dos outros valores que no presente histograma ficaram aglomerados no lado esquerdo do gráfico.

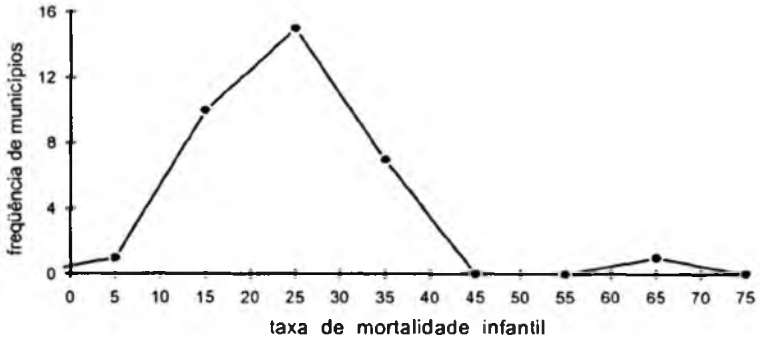


Figura 5.5 Distribuição de freqüências das taxas de mortalidade infantil dos 34 municípios da Microrregião Oeste Catarinense, 1982.

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

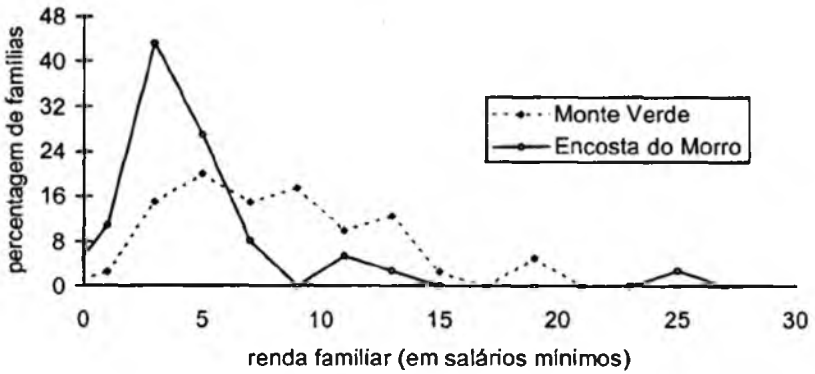


Figura 5.6 Distribuições de freqüências das rendas familiares nas localidades do Monte Verde (amostra de 40 famílias) e Encosta do Morro (amostra de 37 famílias), Bairro Saco Grande II, Florianópolis – SC, 1988.

O leitor deve observar que um gráfico deste tipo (Figura 5.6) permite explorar possíveis relações entre uma variável quantitativa (renda) e uma variável qualitativa (localidade). Ao comparar histogramas ou polígonos de freqüências, devemos observar características como a

posição no eixo horizontal, a dispersão e a assimetria. Dizemos que uma distribuição é ***simétrica*** quando um lado da distribuição é o *reflexo* do outro lado. Medidas físicas, em geral, tendem a ter distribuições razoavelmente simétricas, pois a chance de errar para mais é aproximadamente a mesma de errar para menos. Por outro lado, distribuições de renda são assimétricas, pois existe muito mais pessoas com baixa renda do que pessoas com alta renda (*principalmente no Brasil!*). Veja a Figura 5.7.

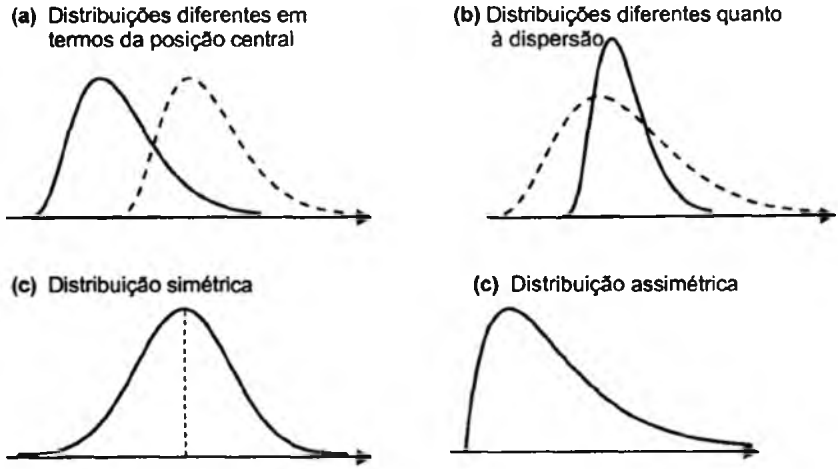


Figura 5.7 Diferentes formas de distribuições de freqüências.

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 Depto de Engenharia Mecânica e a outra no Depto 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 a que pertence.

Valores de identidade social

Depto de Eng. Mecânica	Depto 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

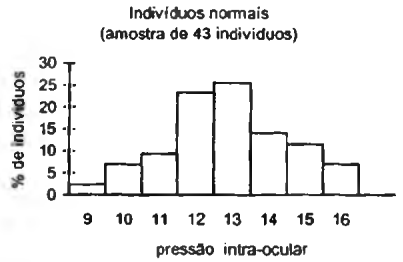
Fonte: Laboratório de Psicologia Social / UFSC, 1990.

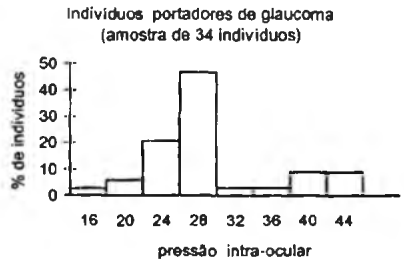
Apresente os dois conjuntos de dados num diagrama de pontos e faça uma análise comparativa.

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

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 intra-oculares para indivíduos normais e para indivíduos portadores de glaucoma. Quais as semelhanças e diferenças que podemos observar na pressão intra-ocular destes 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*, que além de fornecer a forma da distribuição de freqüências, ainda preserva, em parte, a magnitude 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 mortalidade infantil dos municípios da Microrregião Oeste Catarinense. Para facilitar a construção do *ramo-e-folhas* vamos usar, apenas, os dois algarismos mais relevantes, desprezando o algarismo decimal.[9]

---

[9] O mais correto seria arredondar ao invés de simplesmente desprezar o algarismo decimal, mas também estamos preocupados em usar um procedimento simples e rápido. A opção de se trabalhar apenas com dois algarismos baseou-se nos dados em análise. Em algumas situações pode ficar mais interessante trabalhar com números de três dígitos, deixando dois nos *ramos* e um nas *folhas*. O importante é que depois de os dados estarem expostos num *ramo-e-folhas* podemos visualizar bem a forma da distribuição.

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, por exemplo, o valor "32" fica representado por "3 | 2" (veja a quarta linha do *ramo-e-folhas*, Figura 5.8a, o "62" por "6 | 2" (última linha) e assim por diante.

Na apresentação final de um *ramo-e-folhas*, devemos também ordenar as *folhas*, como mostra a Figura 5.8b. 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, "0 | 9 " representa "9 ", "1 | 0 " representa "10 ", etc.

Dados

32,3	62,2	10,3	22,0	13,1	9,9	11,9	20,0	36,4	23,5	18,0	22,6	20,3	38,3	19,6	27,2	28,9
18,4	27,3	21,7	23,7	13,9	36,3	32,9	29,7	25,4	23,8	15,7	17,0	39,2	22,7	29,9	18,3	33,0

Dados com os dois algarismos mais relevantes:

32	62	10	22	13	9	11	20	36	23	18	22	20	38	19	27	28
18	27	21	23	13	36	32	29	25	23	15	17	39	22	29	18	33

(a)

0	9
1	0318983578
2	203207871395329
3	2686293
4	
5	
6	2

(b)

0	9
1	0133578889
2	001222333577899
3	2236689
4	
5	
6	2

unidade = 1
0 | 9 representa 9

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 freqüências, como se fosse um *histograma deitado*. Compare o *ramo-e-folhas* da Figura 5.8b com o histograma da Figura 5.4.

Na Figura 5.8b, notamos que o valor "62" está distante dos demais. É o que chamamos de *valor discrepante*. Podemos, então, estudá-lo separadamente e distribuir melhor os demais valores, duplicando o número

de *ramos* (veja a Figura 5.9).[10] É importante que se tenha a mesma quantidade de possíveis algarismos em cada *ramo* para não distorcer a forma da distribuição. No caso, os algarismos (*folhas*) de 0 a 4 pertencem ao *ramo* tipo "* " e de 5 a 9 ao *ramo* tipo "• ".

```
0•  9
1*  0133
1•  578889
2*  001222333
2•  577899          unidade = 1
3*  223             valor discrepante: 6 | 2
3•  6689
```

Figura 5.9 Apresentação, em ramo-e-folhas, das taxas de mortalidade infantil dos municípios da Microrregião Oeste Catarinense, 1982.

A Figura 5.9 mostra a distribuição com mais detalhes. Podemos observar que, excluindo o valor discrepante 62, os outros valores se distribuem de forma razoavelmente simétrica.

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.

6.512	8.453	30.592	9.279	105.083	21.083	17.968	25.089	14.867
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.

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).

[10] Este mesmo raciocínio pode ser feito com um histograma, basta construirmos classes com amplitudes menores. Se, por exemplo, com os dados em questão, construirmos classes com amplitude 5 (cinco), tais como: 5 ⊢ 10, 10 ⊢ 15, etc., teremos um gráfico equivalente à Figura 5.10.

```
0* | 33
0• | 56889999
1* | 112234
1• | 677899
2* | 114
2• | 5678
3* | 003          unidade = 1.000
3• |              0 | 3 representa 3.000
4* | 4            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. Compare a interpretação que você fez com o histograma do Exercício 5.

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

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 freqüências, onde tornou-se 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 de seus dados por algumas medidas, ou *estatísticas*. Por exemplo, para se conhecer o *peso típico* de crianças nascidas numa comunidade, podemos calcular a *média* ou a *mediana* dos pesos destas crianças ao nascerem. Para se ter idéia da magnitude de *variação do peso* destas crianças, podemos calcular o chamado *desvio padrão*. Em suma, neste capítulo vamos aprender a calcular e interpretar certas medidas, que fornecem informações específicas de um conjunto de valores de certa variável.

Primeiramente, consideraremos a média e o desvio padrão, que são as medidas mais usadas para estudar a posição central e a dispersão de um conjunto de valores. 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 observados de uma certa variável X, podemos definir a média aritmética por

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

onde ΣX indica a soma dos valores observados da variável X.

Exemplo 6.1 A Tabela 6.1 mostra as notas finais dos alunos de três turmas e a nota média de cada turma. E a Figura 6.1 mostra estes três conjuntos de valores representados em diagramas de pontos. As setas apontam para as posições das médias aritméticas.

Tabela 6.1 Notas finais de três turmas de estudantes e a média de cada turma.

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

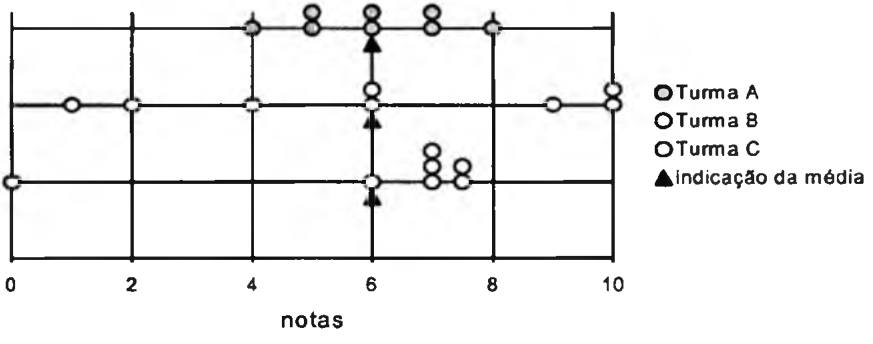


Figura 6.1 Representação das distribuições das notas de três turmas e as correspondentes posições das médias aritméticas.

Observando a Figura 6.1, percebemos que em cada diagrama de pontos, a média aritmética apresenta-se, de alguma forma, na posição central dos valores observados. Mais precisamente, podemos dizer que a média aritmética indica o *centro* de um conjunto de valores, considerando o conceito físico de *ponto de equilíbrio*. Se imaginarmos os pontos como

pesos sobre uma tábua, a *média* é a posição em que um suporte equilibraria esta tábua.

Na Figura 6.1, também observamos que os três conjuntos de valores, apesar de estarem distribuídos sob diferentes formas, apontam para uma mesma média aritmética. Isto mostra que a média aritmética *resume* o conjunto de dados, em termos de uma *posição central*, ou de um *valor típico*, mas não fornece qualquer informação sobre outros aspectos da distribuição. Comparando, por exemplo, as notas da Turma A com as notas da Turma B, verificamos que o segundo conjunto de notas é bem mais *disperso*, indicando que a Turma B é mais heterogênea em termos das notas obtidas. No conjunto de notas da Turma C, observamos um ponto discrepante dos demais, uma nota extremamente baixa, acarretando um valor para a média abaixo da maioria das notas da turma.[1]

Para melhorar o resumo dos dados, podemos apresentar, ao lado da média aritmética, uma medida da dispersão destes dados, 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 contida na 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 destes desvios. Ilustramos, a seguir, as etapas de cálculo usando o conjunto de 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 em relação à média	$X-\overline{X}$	-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, $(x-\bar{X})^2$. 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] Donde definimos a ***variância*** de um conjunto de valores, pela expressão

$$S^2 = \frac{\Sigma(X-\bar{X})^2}{n-1}$$

onde $\Sigma(x-\bar{X})^2$ 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 eqüivale à unidade de medida dos dados ao quadrado. Neste contexto, é mais comum se 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

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

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

Ao compararmos os desvios padrão de vários conjuntos de dados, podemos avaliar quais se distribuem de forma mais (ou menos) dispersa. O desvio padrão será sempre *não negativo* e será tão maior quanto mais

---

[2] Muitos autores costumam diferenciar a fórmula da variância quando os dados se referem a uma população ou a uma amostra. Neste enfoque, 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, é recomendável usar o denominador $n - 1$. Por simplicidade, vamos considerar sempre o segundo caso.

dispersos forem os valores observados. 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

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 padrão, concluímos que os alunos da Turma A obtiveram notas relativamente próximas uma das outras, quando comparados aos alunos das outras turmas. Por outro lado, as notas dos alunos da Turma B foram as que se apresentaram de forma mais heterogênea. Estas conclusões podem ser obtidas tanto pela análise das medidas descritivas (Tabela 6.2) quanto pela análise das distribuições dos valores (Figura 6.1).

Exercícios

1) Faça os cálculos dos desvios padrão 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 padrão 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

Compare sua descrição sobre a tabela com a interpretação que fizemos sobre os diagramas de pontos da Figura 5.3 (Capítulo 5).

6.2 FÓRMULAS ALTERNATIVAS 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 alternativa para o cálculo do 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 quadrática dos valores;

$\overline{X}^2$ é o valor da média elevado ao quadrado; e

n é o número de valores do conjunto de dados.

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	$(\overline{X} = 6)$
Valores ao quadrado	X^2:	16	25	25	36	36	49	49	64	$(\sum X^2 = 300)$

Donde:

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

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

Um 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 $\quad 4(1) + 5(2) + 6(2) + 7(2) + 8(1) = \sum(Xf)$

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

$$\Sigma(X^2 f) = 4^2 + 5^2(2) + 6^2(2) + 7^2(2) + 8^2$$

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

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

A Tabela 6.3 mostra a seqüência 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	Freqüência f	Xf	X^2f
4	1	4	16
5	2	10	50
6	2	12	72
7	2	14	98
8	1	8	64
Total	8	48	300

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

Em situações em que existam muitas repetições de valores, o procedimento previamente exposto facilita o cálculo de $\overline{X}$ e S, como também reduz a possibilidade de erros computacionais.

Dados em tabelas de freqüências

Como vimos na Tabela 6.3, quando os dados estão dispostos em tabelas de freqüências, podemos usar a própria tabela para facilitar a seqüência de cálculos. Porém, se a variável for contínua, com os dados grupados em classes, os cálculos de $\overline{X}$ e S somente poderão ser feitos de forma aproximada, usando os *pontos médios* das classes como se fossem os

próprios valores da variável.[3] O Exemplo 6.2 ilustra uma destas situações, usando uma distribuição de freqüências construída no capítulo anterior.

Exemplo 6.2 Cálculo aproximado de $\overline{X}$ e S com dados grupados em classes de freqüências. A Tabela 6.4 mostra a seqüência dos cálculos.

Tabela 6.4 Distribuição de freqüências das taxas de mortalidade infantil dos municípios da Microrregião Oeste Catarinense, 1982, e cálculos intermediários para obtenção de $\overline{X}$ e S.

Taxa de Mortalidade Infantil	Ponto médio X	Freqüência de famílias f	Xf	X^2f
0 ⊢ 10	5	1	5	25
10 ⊢ 20	15	10	150	2250
20 ⊢ 30	25	15	375	9375
30 ⊢ 40	35	7	245	8575
40 ⊢ 50	45	0	0	0
50 ⊢ 60	55	0	0	0
60 ⊢ 70	65	1	65	4225
Total	-	34	840	24450

Donde:[4]

$$\overline{X} = \frac{840}{34} = 24{,}71 \quad \text{e} \quad S = \sqrt{\frac{24450 - (34).(24{,}71)^2}{33}} = 10{,}57$$

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.

---

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

[4] Se tivéssemos feito os cálculos diretamente com os 34 valores da taxa de mortalidade infantil, encontraríamos $\overline{X}$ = 24,86 e S = 10,37.

5) Calcule a média e o desvio padrão da seguinte distribuição de freqüências.

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

Tamanho da família	Freqüência 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) Desenhe um histograma para a distribuição de freqüências da Tabela 6.4 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 freqüências construída no capítulo anterior, Exercício 5 (cálculo aproximado).

8) Sejam os dados do anexo do Capítulo 2.
- **a)** Calcule as médias e os desvios padrão das respostas dos itens 3(a) a 3(g) do questionário.
- **b)** Apresente estes resultados numa tabela.
- **c)** Interprete os resultados, 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 consideradas.
- **b)** Calcule o desvio padrão da renda familiar em cada localidade.
- **c)** Apresente estes resultados numa tabela.
- **d)** O que você pode concluir a partir destes 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, estas 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 freqüente*, 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 conseqüência, são mais recomendadas para a análise de dados que possam conter estes tipos de valores.

A mediana

A mediana procura avaliar o centro de um conjunto de valores, no sentido 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), já que a quantidade de valores com magnitude inferior a 4 é a mesma que a quantidade de valores com magnitude superior a 4.

Nem todos os conjuntos de dados têm um valor central tão nítido como o exposto acima.[5] Neste sentido, precisamos de uma definição mais precisa para a mediana.

Define-se a ***mediana*** de um conjunto de valores como o valor 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

[5] No conjunto de dados {3, 5, 6, 7, 10, 11}, qualquer valor entre 6 e 7 poderia ser usado como a mediana, enquanto no conjunto {3, 4, 5, 5, 5, 6} não teríamos qualquer valor com a propriedade de que metade dos valores tem magnitudes inferiores a ele e a outra metade tem magnitudes superiores.

dos dois valores de posições mais próximas a $\frac{n+1}{2}$. Vamos representar a mediana por M_d.

EXEMPLOS:

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

$\Rightarrow$ posição $\frac{n+1}{2} = 4 \Rightarrow M_d = 7$

b) $\{5, 3, 2, 8, 4\} \xrightarrow{\text{ordenando}} \{2, 3, \underline{4}, 5, 8\}$, posição $\frac{n+1}{2} = 3 \Rightarrow M_d = 4$

c) $\{3, 5, \underline{6, 7}, 10, 11\} \Rightarrow$ posição $\Rightarrow \frac{n+1}{2} = 3{,}5 \Rightarrow M_d = \frac{6+7}{2} = 6{,}5$

Quando os dados estão apresentados num *ramo-e-folhas* é muito fácil obter a mediana, pois, neste caso, os valores já estão ordenados (veja o exemplo seguinte).

Exemplo 6.3 Obtenção da mediana de dados apresentados em *ramo-e-folhas*, ilustrado pelas taxas de mortalidade infantil dos municípios da Microrregião Oeste de Santa Catarina.[6]

0	9
1	0133578889
2	001222333577899
3	2236689
4	
5	unidade = 1
6	2

$n = 34$

$\Rightarrow$ posição $\frac{n+1}{2} = \frac{35}{2} = 17{,}5$

$\Rightarrow M_d = \frac{22+23}{2} = 22{,}5$

Podemos considerar o valor $M_d = 22{,}5$ como o *valor típico* das taxas de mortalidade infantil dos municípios da Microrregião Oeste Catarinense, pois metade dos municípios acusam taxas de mortalidade infantil inferiores a 22,5 e a outra metade tem níveis mais elevados de mortalidade infantil.

---

6 A construção do *ramo-e-folhas* deste exemplo foi feita na Seção 5.7.

Comparação entre média e mediana

A Figura 6.3 mostra os valores da média e da mediana no diagrama de pontos dos dados do Exemplo 6.3. Note que o valor discrepante 62 *puxa* mais a média do que a mediana.

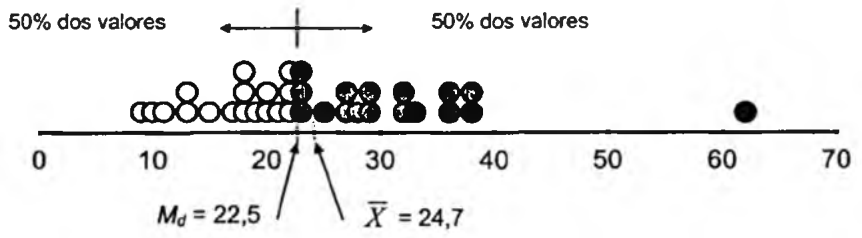


Figura 6.3 Posição da média e da mediana no diagrama de pontos das taxas de mortalidade infantil dos municípios da Microrregião Oeste de Santa Catarina.

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 coincidem numa mesma posição. Em distribuições assimétricas, a média tende a se deslocar para o lado da cauda mais longa.

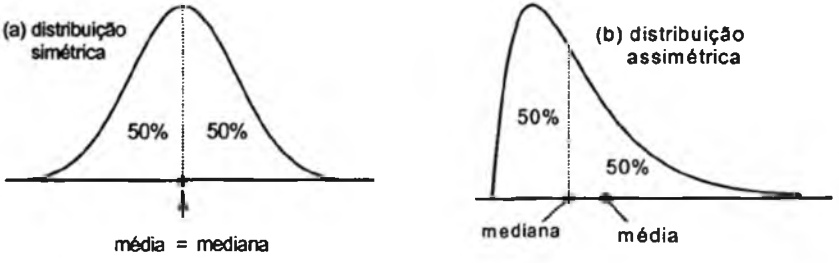


Figura 6.4 Posições da média e da mediana segundo a forma (simétrica ou assimétrica) da distribuição.

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 que a mediana surge como uma alternativa para representar a posição central em distribuições

muito assimétricas.[7] 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 do conjunto de valores. 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.

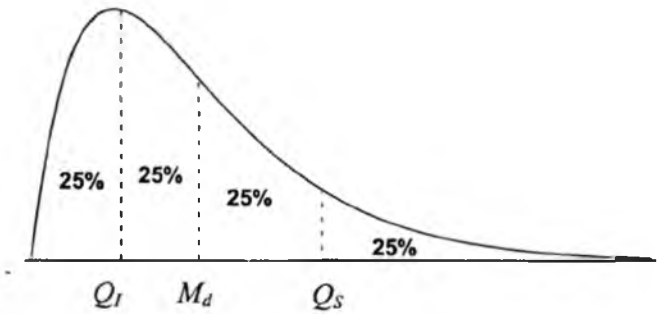


Figura 6.5 Os quartis dividem a distribuição em 4 partes iguais.

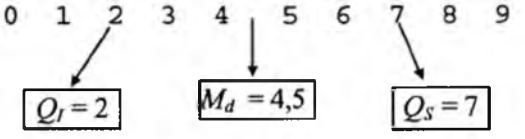
---

[7] Mesmo para variáveis que supostamente tenham distribuições razoavelmente simétricas, a média e a mediana podem não se igualarem, já que, em geral, estamos observando apenas alguns valores (amostras) destas variáveis. Para variáveis com distribuições razoavelmente simétricas, a média é a medida de posição central mais adequada, por usar o máximo de informações contidas nos dados. A média é calculada usando propriamente a magnitude dos valores, enquanto a mediana utiliza somente na ordenação dos valores.

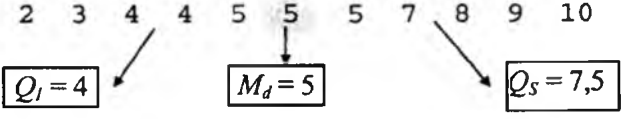
Dado um conjunto de dados ordenados, podemos obter, de forma aproximada, o quartil inferior, Q_I, como a mediana dos valores de posições menores ou iguais à posição da mediana. A mediana dos valores de posições maiores ou iguais à posição da mediana corresponde ao quartil superior, Q_S.[8]

EXEMPLOS:

a) Dados: 2, 0, 5, 7, 9, 1, 3, 4, 6, 8. Ordenando:



b) Dados:



No Exemplo (b), onde a mediana coincidiu com um valor do conjunto de dados, por convenção contamos este valor tanto para a obtenção de Q_I quanto para a obtenção de Q_S.

Exemplo 6.3 (continuação) Obtenção dos quartis de dados apresentados em *ramo-e-folhas*. Taxas de mortalidade infantil dos municípios da Microrregião Oeste de Santa Catarina.

0	9		$\Rightarrow M_d = 22{,}5$
1	0133578889		
2	001222333577899		$\Rightarrow Q_I$ = 18 (mediana dos 17 menores valores)
3	2236689		
4			
5			$\Rightarrow Q_S$ = 29 (mediana dos 17 maiores valores)
6	2	unidade = 1	

Com estas duas novas medidas, Q_I e Q_S, podemos dizer que 25% dos municípios da Microrregião Oeste Catarinense têm taxas de mortalidade infantil não superiores a 18, enquanto existem 25% de municípios nesta microrregião com taxas iguais ou superiores a 29. Podemos dizer, também,

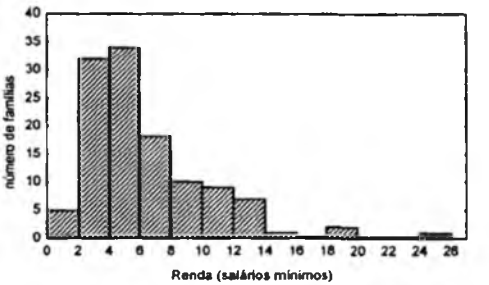
[8] Dado um conjunto de valores, nem sempre conseguimos dividi-lo exatamente em quatro partes iguais. O procedimento exposto oferece uma solução aproximada, mas bastante satisfatória quando a quantidade de valores for grande e com poucas repetições.

que os 50% dos municípios mais típicos desta microrregião, em termos de mortalidade infantil, acusam taxas variando de 18 a 29.

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 seguir, apresenta-se as medidas descritivas da renda, em salários mínimos, de uma amostra de famílias de um bairro de Florianópolis (anexo do Capítulo 4). Estas medidas foram obtidas através da planilha eletrônica *Excel*.[9] Ao lado é apresentado o histograma de freqüências para facilitar a interpretação.[10]

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



Em termos de posição central, tem-se a *média*, a *mediana* e a *moda*. Esta última medida apresenta o valor mais freqüente do conjunto de dados. O fato de a média apresentar um valor maior do 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 este coeficiente se aproxima de zero. Coeficiente de

[9] No *Microsoft Excel*, várias técnicas estatísticas podem ser feitas acionando no menu principal "ferramentas", "suplementos" e solicitando que se instale as "ferramentas de análise". Para obter as medidas descritivas, acionar "ferramentas", "análise de dados" e "estatísticas descritivas".

[10] O histograma foi construído com o apoio do *STATISTICA* 5.1. Ver www.statsoft.com.br

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 a diferença entre o máximo (E_S) e o mínimo (E_I), e a "contagem" é o número de valores usado no cálculo das medidas descritivas.

Esquema dos cinco números

O esquema dos cinco números é uma forma de apresentação dos quartis e extremos, como mostra a Figura 6.6. Através destes números podemos ter informações sobre a posição central, dispersão e assimetria da distribuição de freqüências, como ilustra a Figura 6.7.

$n = 34$

M_d	22,5	
Q	18	29
E	9	62

Figura 6.6 Esquema dos cinco números, construído a partir dos dados do Exemplo 6.3.

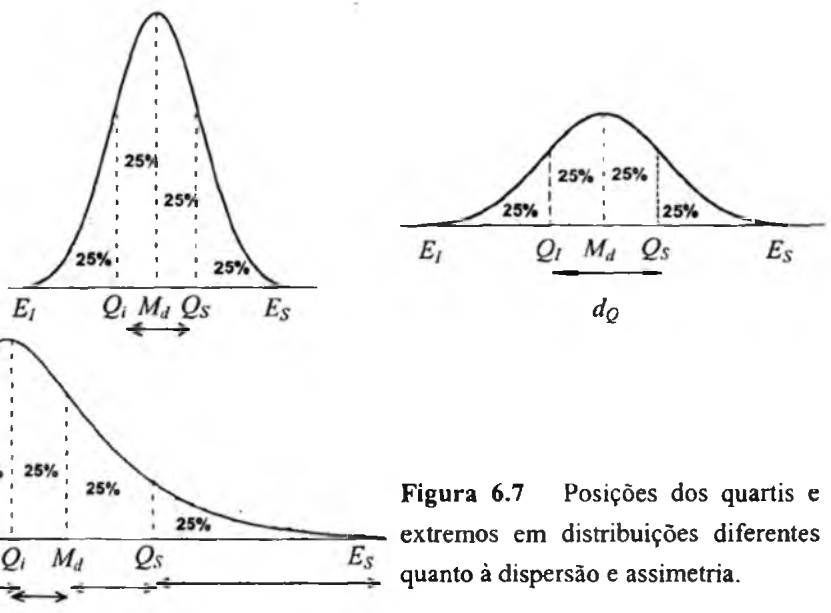
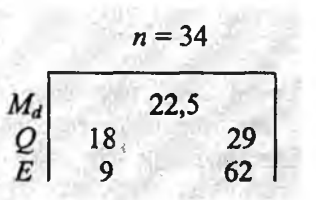


Figura 6.7 Posições dos quartis e extremos em distribuições diferentes quanto à dispersão e assimetria.

O desvio entre quartis, $d_Q = Q_S - Q_I$, é muitas vezes usado como uma medida de dispersão. Veja na Figura 6.7 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 consiste em verificar se existe algum valor do conjunto de dados que se afasta mais do que $(1,5)d_Q$ do quartil superior (ou inferior).

Exemplo 6.3 (continuação)



$$d_Q = Q_S - Q_I = 29 - 18 = 11$$

$$Q_I - (1,5)d_Q = 18 - (1,5)(11) = 1,5$$

$$Q_S + (1,5)d_Q = 29 + (1,5)(11) = 45,5$$

Pelo critério exposto, o extremo superior, 62, pode ser considerado um valor discrepante, pois está além de $(1,5)d_Q$ do quartil superior.

O Exemplo 6.4 mostra uma análise exploratória de dados, usando as medidas descritivas estudadas nesta seção.

Exemplo 6.4 Com o objetivo de comparar a distribuição da renda familiar em duas localidades, construímos, para cada localidade, um *ramo-e-folhas*, acompanhado de um esquema de cinco números, 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			Encosta do Morro		
1	1		0	19	
2	1446	unidade = 0,1	1	38	unidade = 0,1
3	9	1 | 1 representa 1,1	2	123367889	0 | 1 representa 0,1
4	0168		3	599999	
5	11588		4	0224569	
6	8	valores discrepantes:	5	0188	valores discrepantes:
7	12577	18 | 6 e 19 | 3	6	4	11 | 1, 11 | 4,
8	4469		7	19	13 | 9 e 25 | 7
9	006				
10	3349	$n = 40$			$n = 37$
11					
12	25999	M_d 7,7			M_d 3,9
13		Q 4,95 10,35			Q 2,7 5,1
14		E 1,1 19,3			E 0,1 25,7
15	4				

Notamos, inicialmente, que o nível de renda no Conjunto Residencial Monte Verde (mediana de 7,7 salários mínimos) tende a ser 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 tende a ser 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, para ambas as distribuições, 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 do conceito que apresentamos, 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 freqüências é através do chamado *diagrama em caixas* ou *desenho esquemático*. Traça-se dois retângulos: um representando o espaço entre o quartil inferior e a mediana e o outro entre a mediana e o quartil superior. Estes dois retângulos, em conjunto, representam a faixa dos 50% dos valores mais típicos da distribuição. Entre os quartis e os extremos traça-se uma linha. Caso existam valores discrepantes – além de $1{,}5(d_Q)$ –, a linha é traçada até o último valor não discrepante; e os valores discrepantes são indicados por pontos (veja a Figura 6.8).

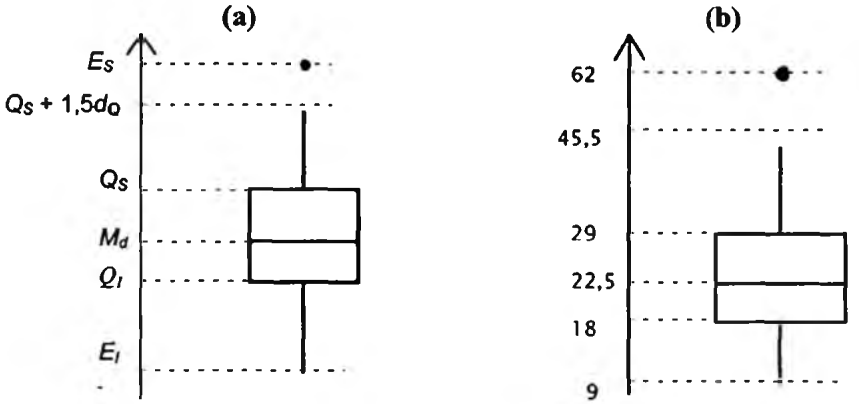


Figura 6.8 (a) Construção de um diagrama em caixas e (b) o diagrama em caixas dos dados do Exemplo 6.3.

A Figura 6.9 mostra a forma do *diagrama em caixas* para uma distribuição simétrica e para uma distribuição assimétrica. Note as diferenças e imagine como ficaria um *diagrama em caixas* se tivéssemos uma distribuição mais dispersa.

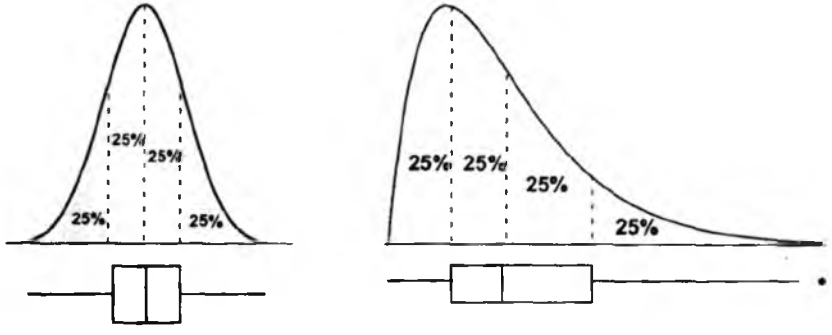


Figura 6.9 Diagrama em caixas e a forma da distribuição.

A Figura 6.10 apresenta os *diagramas em caixas* das duas distribuições de renda do Exemplo 6.4. Compare esta representação com os *ramo-e-folhas* vistos anteriormente.

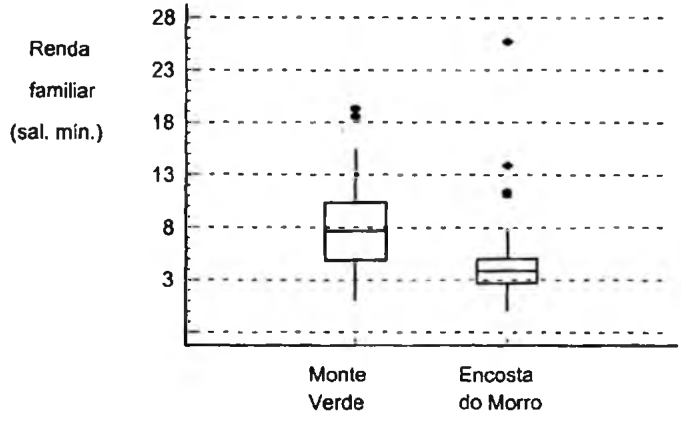


Figura 6.10 Representação das distribuições de renda do Exemplo 6.4 em diagramas em caixas.

Exercícios

10) Verifique os cálculos para a obtenção da mediana e dos quartis dos dois conjuntos de dados do Exemplo 6.4.

11) Obtenha a mediana e os quartis da distribuição de freqüências 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, o que você diria sobre a simetria da distribuição destes valores?

13) A tabela abaixo mostra a distribuição de freqüências 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.

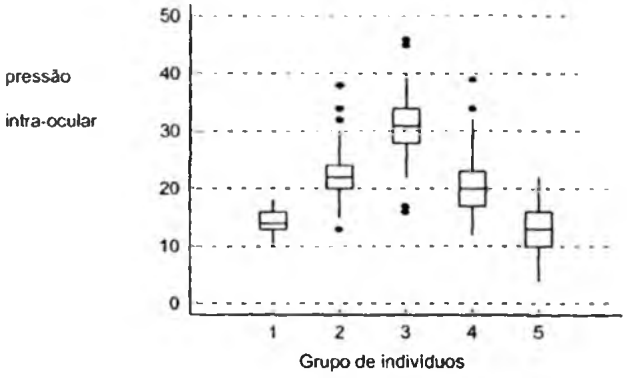
Nº de filhos	1	2	3	4	5	6	7	8	9	10	11	12
freqüência	10	45	32	50	23	23	9	7	6	2	3	2

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.

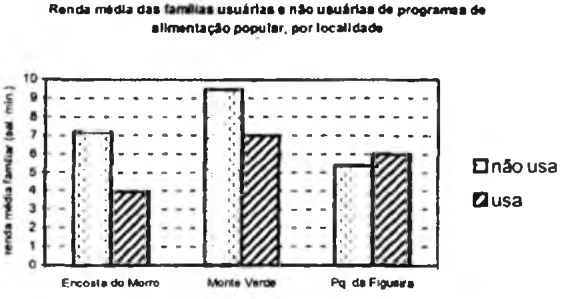
15) A figura seguinte apresenta cinco distribuições de freqüências representadas em *diagramas em caixas*. São dados de pressão intra-ocular de uma amostra de 243 indivíduos, divididos em cinco grupos, segundo a condição clínica da doença *glaucoma*. Descreva as principais informações oriundas desta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



Exercícios complementares

16) No Exemplo 6.2, calculou-se a média da taxa de mortalidade infantil dos municípios da Microrregião Oeste Catarinense. Este valor pode ser interpretado como a taxa de mortalidade infantil da referida microrregião? 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 estes 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) Apresentam-se, abaixo, algumas medidas descritivas da distribuição de salários, em R$, de três empresas do mesmo ramo.

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

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

21) Dada a tabela seguinte, compare os quatro departamentos da UFSC quanto aos escores de Identidade Social com o Departamento. Quanto maior o escore, indica identidade social mais elevada.

Medidas descritivas da 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: 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 uma amostra de observações. Desta forma, estudamos, por exemplo, a distribuição de freqüências do uso (*sim* ou *não*) de programas de alimentação popular, a partir de uma amostra de famílias de um certo bairro (Capítulo 4). Nesta abordagem, predomina o raciocínio indutivo, em que a partir da organização e descrição de dados observados, procuramos fazer conjeturas sobre o problema 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). Um exemplo deste tipo de raciocínio é apresentado a seguir.

Um problema de probabilidade: Supondo que 60% das famílias do bairro usam programas de alimentação popular, o que se pode deduzir sobre a percentagem de famílias que usam estes programas, numa amostra aleatória simples de 10 famílias?[1]

A resposta a esta indagação não é um simples número, pois, dependendo das 10 famílias selecionadas na amostra, teremos resultados diferentes. Para responder adequadamente a esta pergunta, precisamos apresentar quais são os possíveis resultados e como eles poderão ocorrer. Esta descrição é feita em termos dos chamados *modelos probabilísticos*, cuja definição formal veremos na próxima seção.

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

---

[1] Lembramos ao leitor que o termo *amostra aleatória simples* foi discutido no Capítulo 3 e significa que os elementos da amostra são extraídos da população por *sorteio*.

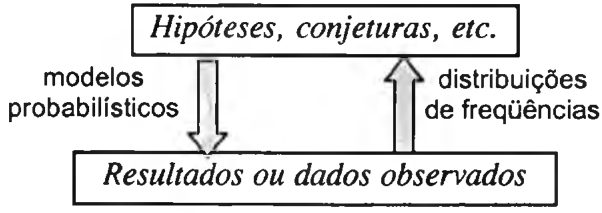


Figura 7.1 Relação entre distribuições de freqüências e modelos probabilísticos.

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, por exemplo, o seguinte experimento: *Lançar uma moeda e observar a face voltada para cima*. Os possíveis resultados são *cara* e *coroa*. Se admitirmos que a moeda é perfeitamente equilibrada e o lançamento for imparcial, podemos também dizer que a *probabilidade* de ocorrer *cara* é a mesma de ocorrer *coroa*.[2]

Espaço amostral e eventos

Dado um experimento aleatório, isto é, *alguma situação em que deve ocorrer um, dentre vários resultados possíveis*, chamamos de ***espaço amostral*** o conjunto de *todos* os resultados possíveis deste experimento. Denotaremos o espaço amostral pela letra grega 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 $\Omega = \{cara, coroa\}$.

---

[2] O leitor deve notar que estas deduções a respeito dos resultados do experimento foram feitas a partir das características físicas da moeda e do lançamento, sem observar efetivamente qualquer lançamento da moeda (ou amostra do fenômeno em estudo).

b) Lançar um dado com os lados numeradas de um a seis 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 à indagação, podemos ser levados a tomar um espaço amostral mais amplo: $\Omega^* = \{sim, não, não\ resposta\}$.

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

f) Numa certa escola de primeiro grau, 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, tal como $\Omega = \{x, \text{ tal que } x \in \Re \text{ e } 0 < x < 2{,}00 \text{ m}\}$.

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

No Exemplo 7.1, os itens de (a) a (e) são *casos discretos*, já que podemos *listar* os possíveis resultados; já no item (f) temos um exemplo do *caso contínuo*, ou seja, dentro de um intervalo de números reais, temos uma infinidade de resultados possíveis. Os *casos contínuos* serão estudados no próximo capítulo.

Chamamos de ***evento*** a qualquer conjunto de resultados possíveis.[3]

Exemplo 7.1b (continuação) Considerando o lançamento de um dado, podemos ter interesse, por exemplo, nos seguintes eventos:
A = ocorrer um número par;
B = ocorrer um número menor que 3;

---

[3] 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 todos os resultados possíveis.

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

As probabilidades são valores entre 0 (zero) e 1 (um). E a soma das probabilidades de todos os resultados possíveis do experimento deve ser igual a 1 (um).

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 destas probabilidades será apresentado posteriormente.

a) No lançamento de uma moeda, se considerarmos a moeda perfeitamente equilibrada e o lançamento imparcial, os resultados tornam-se eqüiprováveis, donde podemos alocar probabilidade 0,5 (um meio) tanto para *cara* como para *coroa*, resultando no modelo probabilístico mostrado ao lado.

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	$^1/_6$	$^1/_6$	$^1/_6$	$^1/_6$	$^1/_6$	$^1/_6$

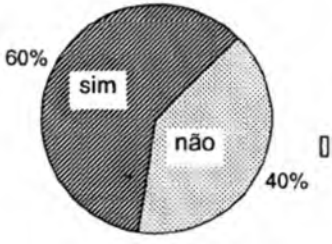
c) Na seleção de uma bola de uma urna, para construirmos um modelo para a cor da bola a ser extraída, precisamos conhecer a quantidade (ou a percentagem) de bolas de cada cor, existentes na urna. Se existirem, por

exemplo, 7 bolas azuis e 3 vermelhas e admitindo que a bola seja extraída *aleatoriamente*, temos o seguinte modelo:[4]

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 Ω = {sim, não}. Como no caso anterior, torna-se necessário o conhecimento da distribuição desta característica na população. Por exemplo, se admitirmos que em todo o bairro 60% das famílias utilizam e 40% não utilizam programas de alimentação popular e admitindo, também, que a família seja selecionada aleatoriamente, podemos explicitar o modelo probabilístico, como mostra o esquema seguinte.

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 o uso de programas de alimentação popular.

Resultado	Probabilidade
sim	0,6
não	0,4

Para a alocação das probabilidades nos diversos itens do Exemplo 7.1, usamos o chamado *princípio da eqüiprobabilidade*. Por exemplo, no problema da urna (item c), fizemos o seguinte raciocínio: "Como a seleção é

[4] Usaremos freqüentemente 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.

aleatória, toda bola da urna tem a mesma probabilidade de ser selecionada. Como existem 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 eqüiprobabilidade é usualmente enunciado em termos da probabilidade de algum evento, como apresentamos a seguir.

PRINCÍPIO DA EQÜIPROBABILIDADE. Quando as características do experimento sugerem N resultados possíveis, todos com igual probabilidade de ocorrência, a probabilidade de um certo evento A, contendo n resultados, pode ser definida por

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

ou seja,

$$P(A) = \frac{\textit{número de resultados de } A}{\textit{número total de resultados}}$$

Usando este princípio, vamos alocar probabilidades aos seguintes eventos, baseados num lançamento imparcial de um dado perfeitamente equilibrado (Exemplo 7.1b).

Eventos	Probabilidades
A = ocorrer um número par	P(A) = $^{3}/_{6}$ = ½ ou 0,5
B = ocorrer um número menor que 3	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, a partir do conhecimento das probabilidades de resultados individuais, é *somando as probabilidades dos resultados que integram o evento*. Por exemplo, no exemplo do dado, P(*ocorrer um número par*) = $^{1}/_{6}$ + $^{1}/_{6}$ + $^{1}/_{6}$ = ½. Este procedimento pode ser usado mesmo quando os resultados não são eqüiprováveis.[5]

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

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?

Solução: 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 - $^2/_{10} = ^8/_{10}$.

Dizemos que 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 destes eventos independentes.

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

Solução: P(*número par em ambos os lançamentos*) =
= P(*nº par no 1º lançamento*) . P(*nº par no 2º lançamento*) =
= (½)(½) = ¼ .

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 considerado.

Exemplo 7.4 São exemplos de ensaios de Bernoulli:

a) Seja uma 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.[6]

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

d) Lançar um dado e observar se ocorreu o ponto *seis*.[7]

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 estes casos existem apenas dois resultados a serem observados. Ou seja, o espaço amostral pode ser Ω = {*sim*, *não*} para qualquer item de (a) a (f). Sob certas suposições a respeito do experimento e admitindo o conhecimento da distribuição de *sim* e *não* na população, podemos especificar o modelo probabilístico, como ilustraremos para os itens (b) e (c).

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 o seguinte modelo probabilístico:

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

[6] 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.

[7] 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*}.

As especificações dos modelos para os outros itens ficam como exercício para o leitor.

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. Nestes casos, podemos apresentar apenas o *jeitão* do modelo, como mostra o quadro seguinte:

Resultado	Probabilidade
sim	π
não	$1 - \pi$

onde π é um valor (desconhecido) entre 0 e 1. O intervalo de 0 a 1 deve-se à própria definição de probabilidade. A probabilidade de *não*, igual a $1-\pi$, é devida ao fato de que a soma das probabilidades de todos os resultados possíveis deve ser igual a 1 (um).[8]

O número π, do modelo anterior, corresponde ao parâmetro *proporção de favoráveis ao projeto na população*. Usaremos o termo ***parâmetro*** num modelo probabilístico, para designar alguma quantidade desconhecida, mas que se tornaria conhecida se tivéssemos informações adicionais sobre a população de onde está sendo tirada a amostra, ou de características físicas do experimento em questão.

Variável aleatória

Chamamos de ***variável aleatória*** a uma característica numérica associada aos resultados de um experimento.[9] Exemplos: 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.*

---

[8] A quantidade π está sendo apresentada, no presente contexto, para designar uma probabilidade desconhecida, nada tendo a ver com o número π usado em trigonometria.

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

Podemos caracterizar um ensaio de Bernoulli por uma variável aleatória X, definida da seguinte forma: $X = 0$, se *não* e $X = 1$, se *sim*. E a formulação geral seria:

x	1	0
p(x)	π	$1 - \pi$

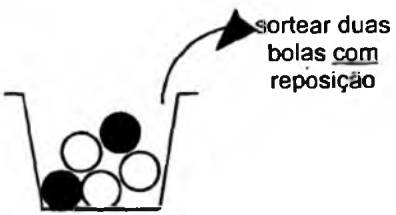
onde: π é uma quantidade entre 0 e 1;
x é um possível valor de X (no caso, 0 ou 1); e
p(x) é a probabilidade de ocorrer o valor x (isto é, p(0) é a probabilidade de X assumir o valor 0 e p(1) é a probabilidade de X assumir o valor 1).

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

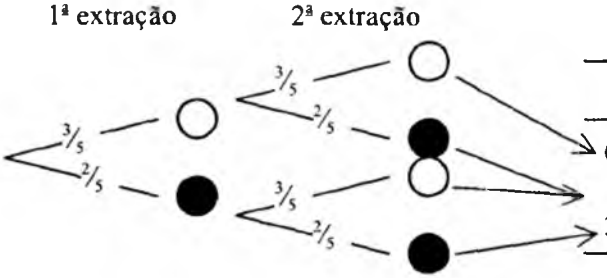
Dois ensaios de Bernoulli

Quando temos dois ensaios de Bernoulli, geralmente o interesse está na variável aleatória X = *número de ocorrências de sim nos dois ensaios*, como ilustram os exemplos seguintes.

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*. Queremos a distribuição de probabilidades da variável X = *número de bolas pretas extraídas na amostra.*



Solução: Os possíveis resultados de X = *número de bolas pretas numa amostra de duas bolas* são {0, 1, 2}. Contudo, a alocação de probabilidades para estes resultados não é uma tarefa muito fácil. Por isto, decompomos o experimento em duas partes: 1ª *extração* e 2ª *extração*, como mostra o esquema a seguir.

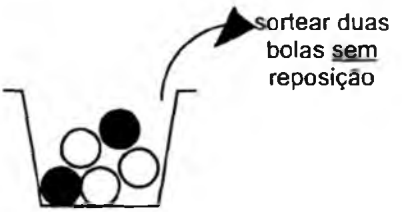


Distribuição de X

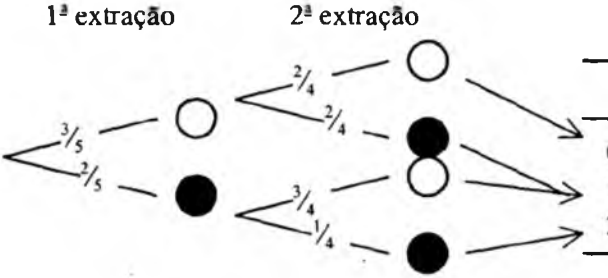
x	p(x)
0	$9/25$ (ou 0,36)
1	$12/25$ (ou 0,48)
2	$4/25$ (ou 0,16)

Para se obter a probabilidade de $X = 0$, calcula-se a probabilidade de *ocorrer bola branca na* 1ª *e bola branca na* 2ª *extração*, ou seja $(3/5)(3/5) = 9/25$ (ou 0,36). Analogamente, a probabilidade de $X = 2$ é dada por $(2/5)(2/5) = 4/25$ (ou 0,16). Um cuidado adicional deve-se ter ao calcular a probabilidade de $X = 1$, que ocorre quando acontecer *bola branca na* 1ª *e bola preta na* 2ª (com probabilidade de $(3/5)(2/5) = 6/25$), ou, *bola preta na* 1ª *e bola branca na* 2ª (com probabilidade de $(2/5)(3/5) = 6/25$). Logo, a probabilidade de $X = 1$ é $6/25 + 6/25 = 12/25$ (ou 0,48).

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 da variável aleatória X. Contudo, se o leitor refizer estes cálculos, considerando um grande número de bolas (digamos, 2000 bolas brancas e 3000 bolas pretas), a distribuição de probabilidades da variável X será praticamente a mesma, ao realizar amostragens *com* ou *sem* reposição. Neste contexto, ao tratarmos de *grandes populações*, podemos supor independência entre os ensaios, mesmo que a amostragem seja feita *sem* reposição.

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 freqüências do grau de instrução do chefe da casa e uso de programas de alimentação popular, num conjunto de 120 famílias.

	Grau de Instrução do Chefe da Casa			
Uso de programas	nenhum	primeiro grau	segundo grau	Total
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 segundo grau;

c) tal que o chefe da casa não tenha o segundo grau;

d) usuária de programas de alimentação popular e o chefe da casa ter o segundo grau;

e) usuária de programas de alimentação popular e o chefe da casa não ter o segundo grau;

f) usuária de programas de alimentação popular, considerando que o sorteio tenha sido restrito às famílias cujo chefe da casa tenha o segundo grau;

g) tal que o chefe da casa tenha o segundo grau, considerando que o sorteio tenha sido restrito às famílias usuárias de programas de alimentação popular.

5) Seja a população descrita no Exercício 4. Seleciona-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 onde observamos a presença (ou ausência) de algum atributo. Em geral, temos interesse no número (ou percentagem) de elementos que têm o atributo em estudo, numa amostra de n elementos observados.

Caracterização de um experimento binomial

Um experimento é dito binomial, quando:

> **(1)** consiste de n ensaios;
> **(2)** cada ensaio tem apenas dois resultados: *sim* ou *não*; e
> **(3)** os ensaios são *independentes* entre si, com probabilidade π de ocorrer *sim*, sendo π uma constante entre 0 e 1 ($0 < \pi < 1$).

O interesse está na distribuição de probabilidades da variável aleatória X = *número de ocorrência de sim nos n ensaios*. A distribuição de probabilidades de uma variável aleatória desse tipo é conhecida como ***distribuição binomial.*** E as quantidades n e π são os *parâmetros* da

distribuição, cuja especificação depende 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) Observar o número Y de caras, em três lançamentos imparciais de uma moeda perfeitamente equilibrada. (Neste exemplo, temos: $n = 3$ e $\pi = 0{,}5$.)

b) Observar o número X de respostas afirmativas, numa amostra aleatória de dez pessoas, indagadas a respeito de um projeto municipal, dentre uma grande população de pessoas, onde 70% delas são favoráveis. Admita que todas as pessoas dessa população responderiam *sim* ou *não* à indagação. (Neste exemplo, temos: $n = 10$ e $\pi = 0{,}7$.)

c) Observar o número F de eleitores, que se declaram a favor de um certo candidato, numa amostra de 3000 eleitores, extraída aleatoriamente de uma população de 100.000 eleitores. (Neste exemplo, temos: $n = 3000$ e π = *proporção de eleitores favoráveis ao candidato na referida população.*)

A tabela da distribuição binomial

Para conhecermos as probabilidades de uma variável com distribuição binomial, podemos fazer uso da Tabela II do apêndice (*tabela da distribuição binomial*).[10]

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.

---

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

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.

Parte da Tabela II

		π		
n	x	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, onde 70% são favoráveis a um certo projeto municipal. Qual é a probabilidade de que, numa amostra aleatória simples de 10 pessoas desta população, a maioria seja favorável ao projeto?

Solução: 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 II

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

Uma distribuição de probabilidades também pode ser apresentada sob forma gráfica, de maneira análoga às distribuições de freqüências, substituindo o eixo das freqüências por probabilidades. Veja a Figura 7.2.[11]

[11] O leitor deve notar que a variável em questão é discreta, pois só pode assumir determinados valores. Assim, estamos usando as mesmas formas gráficas descritas na Seção 5.1, que tratava de distribuições de freqüências de variáveis discretas.

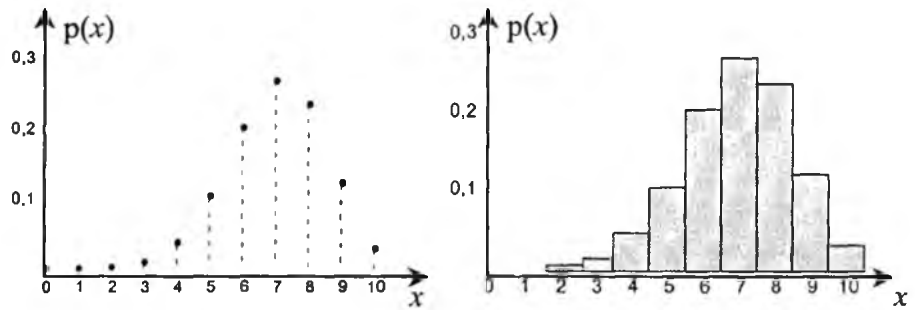


Figura 7.2 Representações gráficas da distribuição binomial com $n = 10$ e $\pi = 0{,}7$ (Exemplo 7.7b).

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 pessoas, 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 a probabilidade de ocorrer 60% ou mais de caras, ou seja, $P(X \geq 3)$, onde X é o número de vezes em que aparece *cara*.

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) Considerando o Exemplo 7.7b, mas admitindo que a distribuição da população seja 40% favorável e 60% contrária ao 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 dois ou mais favoráveis, ou seja, $P(X \geq 2)$;
- **b)** probabilidade de a amostra acusar menos de dois favoráveis, ou seja, $P(X < 2)$;
- **c)** probabilidade de a amostra acusar mais de 50% de favoráveis.

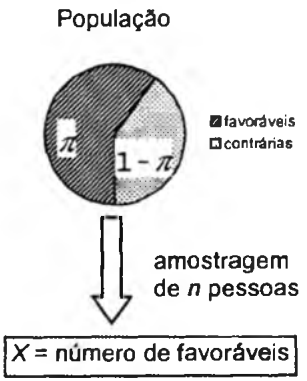
12) Considerando o Exercício 9, construa a distribuição de probabilidades da variável P = *proporção de indivíduos favoráveis na amostra de tamanho cinco.*

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 este programa de treinamento, 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, quando o processo estiver 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

Considere o seguinte experimento: seja X o número de pessoas favoráveis a um certo projeto municipal, numa amostra aleatória simples de n pessoas, extraída de uma população, onde a proporção de favoráveis é igual a π, como ilustra o esquema ao lado.



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.

Para cada uma das pessoas indagados 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.

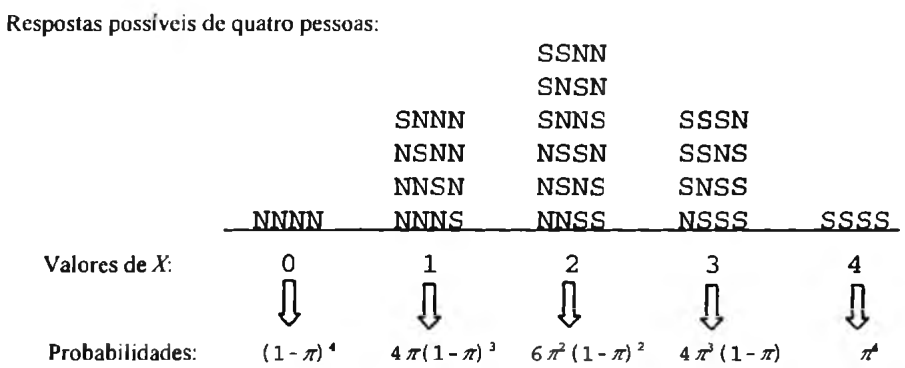


Figura 7.3 Possíveis seqüências de respostas e construção de uma distribuição binomial de probabilidades com $n = 4$ e π genérico.

Explicando as probabilidades: O evento $X = 0$ ocorre quando são sorteadas, para fazer parte da amostra, quatro pessoas contrárias ao projeto (NNNN), cuja probabilidade é $(1-\pi)(1-\pi)(1-\pi)(1-\pi)$, ou, $(1-\pi)^4$. 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 $\pi(1-\pi)^3$, a probabilidade do evento $X = 1$ é $4\pi(1-\pi)^3$. As outras probabilidades podem ser obtidas de forma análoga.

Coeficientes binomiais

Na Figura 7.3, podemos observar que, no cálculo da probabilidade do evento $X = 1$, contamos de quantas maneiras poderia aparecer uma resposta afirmativa, na amostra de quatro pessoas, e encontramos a quantidade 4 (quatro), correspondente às seguintes seqüências de respostas: SNNN, NSNN, NNSN e NNNS.

De um modo geral, na distribuição binomial, para calcular a probabilidade do evento $X = x$, onde x é um valor possível da variável aleatória X, precisamos conhecer o número de maneiras que podemos combinar as x respostas afirmativas, dentre as n respostas. Este valor, conhecido como

coeficiente binomial, entra no cálculo da probabilidade como um coeficiente das potências de π e $1-\pi$, como verificamos na Figura 7.3.

Vamos representar o número de combinações que podemos fazer com x elementos, numa seqüência de n elementos (sendo $x < n$), por $\binom{n}{x}$. Este número de combinações pode ser obtido na *Tabela dos Coeficientes Binomiais* (Tabela III 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.

$x = 0$: $\binom{4}{0} = \frac{4!}{4!0!} = \frac{4!}{4!} = 1$

$x = 1$: $\binom{4}{1} = \frac{4!}{3!1!} = \frac{4.3.2.1}{3.2.1.1} = 4$

$x = 2$: $\binom{4}{2} = \frac{4!}{2!2!} = \frac{4.3.2.1}{2.1.2.1} = 6$

$x = 3$: $\binom{4}{3} = \frac{4!}{1!3!} = \frac{4.3.2.1}{1.3.2.1} = 4$

$x = 4$: $\binom{4}{4} = \frac{4!}{0!4} = \frac{4!}{4!} = 1$

Expressão geral da distribuição binomial

O raciocínio que fizemos para obter as probabilidades na Figura 7.3, pode ser generalizado para qualquer experimento binomial. E este raciocínio pode ser sintetizado pela expressão matemática que apresentamos a seguir.

Seja 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 $\{0, 1, 2,...,n\}$, 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, onde 70% são favoráveis a um certo projeto municipal (Exemplo 7.7b). Qual a

probabilidade de, numa amostra aleatória simples de quatro pessoas desta população, encontrarmos exatamente três pessoas favoráveis ao projeto?

Solução: Neste caso, X tem distribuição binomial com parâmetros $n = 4$ e $\pi = 0,7$. Então, a probabilidade pedida é dada por

$$p(3) = \binom{4}{3}.(0,7)^3.(0,3)^1 = 4.(0,7)^3.(0,3) = 0,4116$$

Se o leitor procurar na tabela da distribuição binomial (Tabela II do apêndice), deve encontrar o mesmo resultado.

Exercícios

15) Refazer o Exercício 9, sem usar a tabela da distribuição binomial.

16) (Bussab e Morettin, 1985, p.92.) 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 $\frac{2}{3}$. 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 selecionarmos, aleatoriamente e com reposição, cinco alunos desta população, quais são as probabilidades destas respostas:

a) nenhuma das cinco acusa "plenamente satisfeito"?

b) a maioria acusa "plenamente satisfeito"?

c) pelo menos uma indica "plenamente satisfeito"?

Exercícios complementares

18) De uma sala com 4 homens e 2 mulheres. Selecionar, ao acaso e sem reposição, 2 pessoas. Qual é a probabilidade de se obter exatamente 1 mulher?

19) Uma sala contém 20 mulheres e 80 homens. Se forem escolhidas, aleatoriamente e com reposição, 6 pessoas, qual é a probabilidade de que:

a) cinco ou mais sejam homens?

b) haja exatamente 2 mulheres?

c) haja pelo menos uma mulher?

20) Numa população onde 32% dos indivíduos têm alguma descendência indígena, retira-se uma amostra aleatória de 6 pessoas. Qual é a probabilidade de se encontrar

a) exatamente 2 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 deixam de pagar regularmente as suas contas (prestações). Se num particular dia, a loja vende a crédito para 10 pessoas, qual a probabilidade de que mais de 20% delas deixam de pagar regularmente as contas? Admita que as 10 pessoas que fizeram crediário nesse dia, corresponda a uma amostra aleatória de clientes potenciais desta loja.

22) Admitamos igualdade de probabilidade para o nascimento de menino e menina. De todas as famílias com 6 filhos:

a) que proporção tem 3 meninos e 3 meninas?

b) que proporção tem 4 ou mais meninas?

23) Um exame de múltipla escolha consiste em 10 questões, cada uma com 4 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"?

Capítulo 8

Distribuições contínuas e o modelo normal

Neste capítulo, estudaremos o modelo de probabilidades mais conhecido da Estatística: a chamada *distribuição normal de probabilidades*. Diversas aplicações deste modelo estarão presentes ao longo dos demais capítulos. Para podermos estudar esta distribuição, vamos, inicialmente, estender o conceito de eqüiprobabilidade para variáveis aleatórias contínuas.

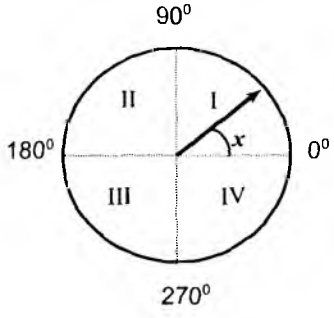
Dizemos que uma variável aleatória é ***contínua*** quando não conseguimos enumerar seus possíveis resultados, por estes formarem um conjunto infinito, num dado intervalo de números reais. Por exemplo, a altura de um indivíduo, tomado ao acaso, é uma variável aleatória contínua, pois não é possível enumerar 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

Para variáveis aleatórias contínuas, não existe interesse em atribuir probabilidades a cada particular valor, mas sim, para eventos formados por intervalos de valores. Por exemplo, ao observar a altura de um indivíduo, tomado ao acaso, não importa a probabilidade de ele medir 1,682333... metros; mas o interesse pode estar, por exemplo, 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 pára de girar. Como existem infinitos pontos no intervalo de 0 a 360^0, 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 90^0.

Admitindo que não exista alguma região de preferência para o ponteiro parar, podemos deduzir, pelo *princípio da eqüiprobabilidade*, 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 $\frac{1}{4}$.

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 "=" são equivalentes, pois, considerando a eqüiprobabilidade de todos os pontos e, considerando 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 certa função não negativa, com a área formada entre o eixo das abscissas e a curva desta função igual a 1 (um). Os eventos podem ser representados por intervalos no eixo das abscissas (eixo X), enquanto as correspondentes probabilidades, por áreas sob a curva. Apresentamos, na Figura 8.1, uma distribuição de probabilidades para o experimento do Exemplo 8.1, sob forma gráfica.

A função descrita pela Figura 8.1a se identifica com uma constante no intervalo de 0 a 360^0, 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 esta distribuição, qualquer probabilidade associada à variável X, pode ser obtida pelo cálculo de uma certa área. Neste contexto, a Figura 8.1b ilustra a probabilidade do ponteiro parar no primeiro quadrante.

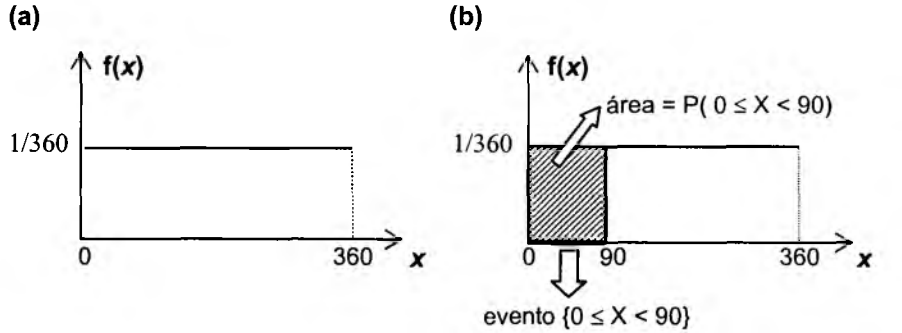


Figura 8.1 Ilustração de: (a) uma distribuição de probabilidades para a variável aleatória do Exemplo 8.1; e (b) a probabilidade do evento $\{0 \leq X < 90\}$.

Exemplo 8.2 Selecionar, aleatoriamente, de uma certa universidade, uma 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 acusar altura no intervalo de 165 a 175 cm é bem maior do que no intervalo de 190 a 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.*

---

[1] A área de um retângulo é dada por (*base*)(*altura*). Como a base é 360 e a área 1, acarreta uma altura de $1/360$.

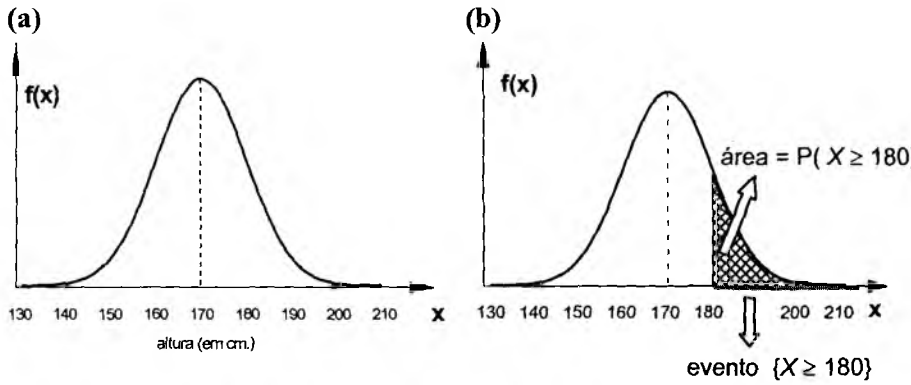


Figura 8.2 Um modelo para a altura de um aluno universitário.

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)* – este parâmetro especifica a posição central da distribuição de probabilidades.

σ *(desvio padrão)* – este parâmetro 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 das abscissas é igual a 1 (um), permitindo identificar probabilidades de eventos como áreas sob a curva, como já ilustramos na Figura 8.2b.

---

[2] Os parâmetros μ e σ do modelo normal têm analogia com as estatísticas $\overline{X}$ e S (Capítulo 6), usadas para medir, respectivamente, a posição central e a dispersão de uma distribuição de freqüências.

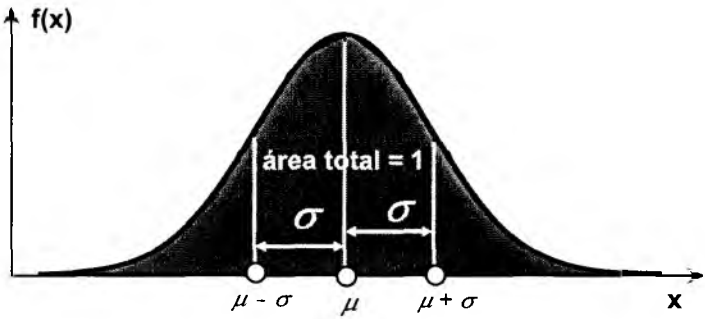


Figura 8.3 Gráfico da distribuição normal com parâmetros μ e σ.

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.

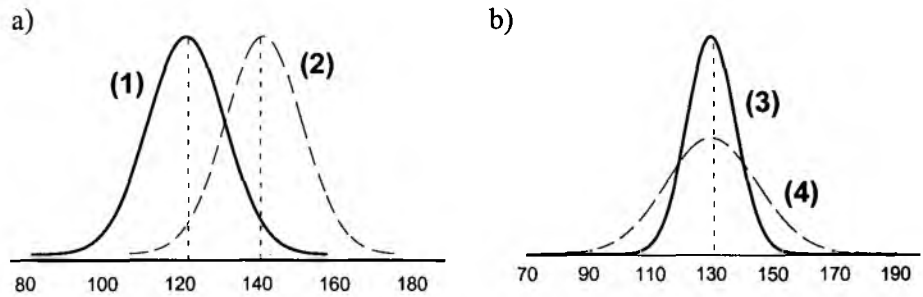


Figura 8.4 Distribuições normais em função dos parâmetros μ e σ.

As duas distribuições da Figura 8.4a podem representar, por exemplo, (1) *alturas de estudantes da primeira série do primeiro grau* 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 devem ser aproximadamente iguais ($\mu_3 \approx \mu_4$), 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 uma distribuição normal com média 0 (zero) e desvio padrão 1 (um), também conhecida como *distribuição normal padrão*.

Para que um dado valor x, de uma distribuição normal com média μ e desvio padrão σ, se transforme num valor z da distribuição normal padrão, basta fazer a seguinte operação:

$$z = \frac{x - \mu}{\sigma}$$

O valor z é conhecido como *valor padronizado*. Ele fornece uma medida relativa do valor x, em termos da distribuição da variável aleatória em estudo, como ilustramos no seguinte exemplo.

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 $z = {}^{(180 - 170)}/_{10} = 1$, ou seja, este estudante encontra-se a 1 (um) desvio padrão acima da altura média dos estudantes do sexo masculino da universidade.

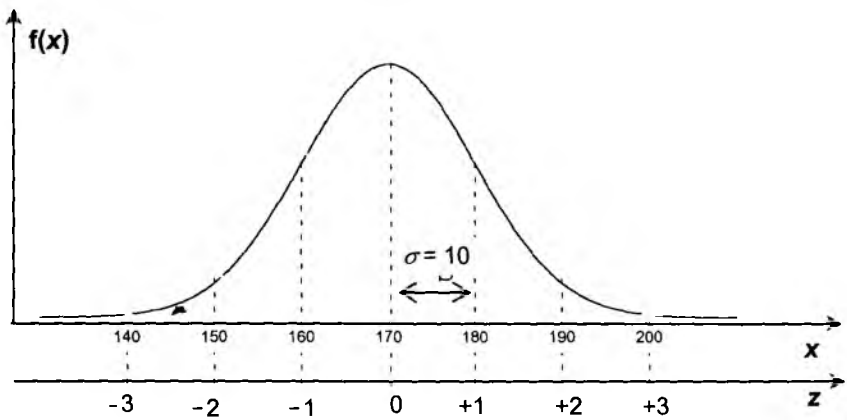


Figura 8.5 Transformação de valores de alturas de universitários (x) em valores padronizados (z).

Seja X a altura, em centímetro, de um estudante do sexo masculino, selecionado ao acaso, desta universidade. Considere que temos interesse no evento $\{X > 180\}$. A Figura 8.6 mostra a equivalência da probabilidade deste evento, $P(X > 180)$, com uma certa área na distribuição normal padrão. Para facilitar a notação, identificaremos por Z uma variável aleatória com distribuição normal padrão.

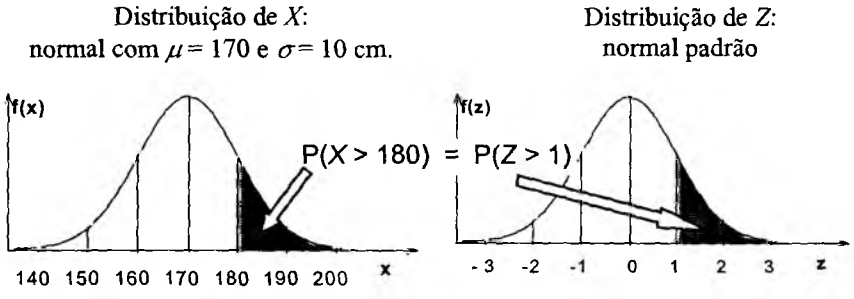


Figura 8.6 Transformação de um evento da distribuição normal de parâmetros $\mu = 170$ cm e $\sigma = 10$ cm, num evento da distribuição normal padrão.

Exercícios

1) Considerando a distribuição do Exemplo 8.3, encontre os valores padronizados para os seguintes valores de X: **a)** $x = 190$ cm; **b)** $x = 185$ cm; **c)** $x = 170$ cm; e **d)** $x = 165$ cm.

2) Ainda, considerando o Exemplo 8.3 e lembrando que a distribuição normal é perfeitamente simétrica em torno da média μ, qual é a probabilidade do estudante sorteado apresentar altura acima de 170 cm?

3) Suponha que as notas X de um vestibular tenham distribuição normal com média 60 pontos e desvio padrão 15 pontos.

a) Se você prestou este vestibular e obteve nota $x = 80$ pontos, qual é a sua posição relativa, em unidades de desvios padrão, com relação à média das notas?

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 variável com distribuição normal podem ser representadas por áreas sob a curva da distribuição normal padrão. No apêndice, apresentamos a Tabela IV, que relaciona valores positivos de z, com áreas sob a cauda superior da curva. Os valores de z são apresentados com duas decimais. A primeira decimal fica na coluna da esquerda e a segunda decimal na linha do topo da tabela. A Figura 8.7 mostra como podemos usar a Tabela IV do apêndice para encontrar, por exemplo, a área sob a cauda superior da curva, além de $z = 0{,}21$.

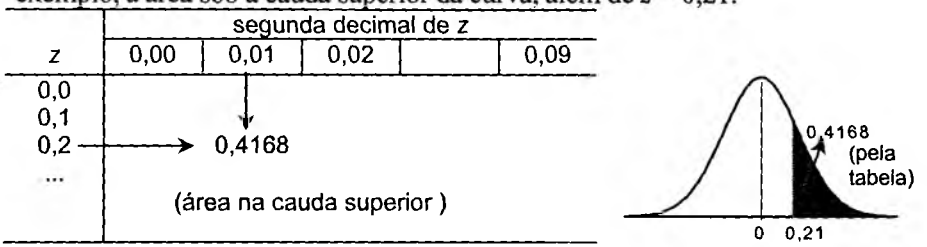


Figura 8.7 Ilustração do uso da tabela da distribuição normal padrão (Tabela IV do apêndice) para encontrar a área na cauda superior relativa ao valor de $z = 0{,}21$.

Exemplo 8.3 (continuação) Admitimos que a altura X de um estudante do sexo masculino, tomado ao acaso de uma universidade, tinha distribuição normal com média 170 cm e desvio padrão 10 cm. Vimos, também, que a probabilidade de ele acusar altura superior a 180 cm correspondia à área acima de $z = 1$ da curva normal padrão, isto é, $P(X > 180) = P(Z > 1)$. Usando a Tabela IV do apêndice, podemos encontrar esta área (probabilidade), como ilustra o esquema seguinte.

	segunda decimal de z		
z	0,00	...	0,09
...			
1,0 →	0,1587		
...			

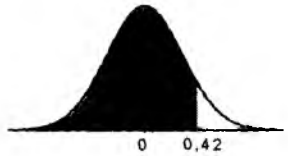
Portanto,
$P(X > 180) = 0{,}1587$.

A Tabela IV considera valores de z entre 0 (zero) e 5 (cinco). Além de $z = 5$, a área pode ser considerada 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. O

exemplo seguinte mostra como podemos operar com áreas, a fim de obter diversas probabilidades de interesse.

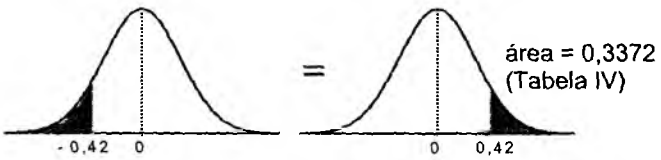
Exemplo 8.4 Seja Z uma variável aleatória com distribuição normal padrão. Vamos usar a Tabela IV para encontrar as seguintes probabilidades:

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:



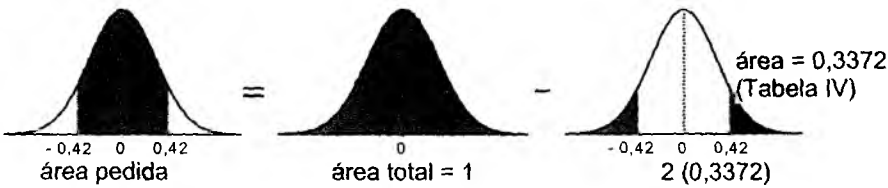
Portanto, $P(Z < 0{,}42) = 0{,}6628$.

b) $P(Z < -0{,}42)$. O esquema seguinte mostra esta probabilidade em termos de área e a correspondente operação para podermos usar a Tabela IV.



Portanto,
$P(Z < -0{,}42) =$
$= 0{,}3372$

c) $P(-0{,}42 < Z < 0{,}42)$.

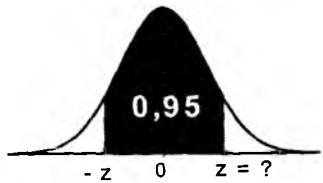


Então, $P(-0{,}42 < Z < 0{,}42) = 1 - 2\,(0{,}3372) = 0{,}3256$.

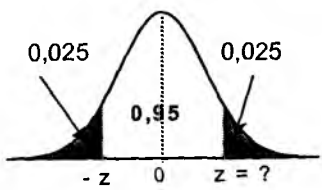
Como vimos nos exemplos precedentes, podemos obter a probabilidade de qualquer evento relativo a uma variável normal padrão,

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 o valor de z, tal que de $-z$ até z produza uma área sob a curva de 0,95? A figura ao lado ilustra esta pergunta.



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 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 IV 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		
...						

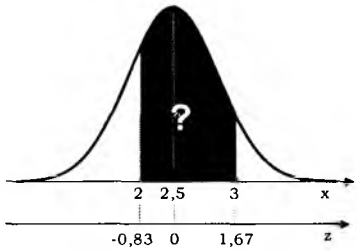
Exemplo 8.6 Suponha que o desempenho dos alunos das três últimas fases do Curso de Ciências da Computação da UFSC tenha distribuição normal de média 2,5 e desvio padrão de 0,6.[3] Selecionando aleatoriamente um aluno desta população, qual a probabilidade de ele acusar desempenho entre 2 e 3,5?

---

[3] Foram usados como estimativas de μ e σ, os valores das estatísticas $\bar{X}$ e S, calculadas a partir dos dados observados nesta população (anexo do Capítulo 2).

Solução: Primeiramente precisamos transformar os valores de desempenho, x, em valores padronizados:

$$z = \frac{x - \mu}{\sigma} = \frac{x - 2{,}5}{0{,}6}$$

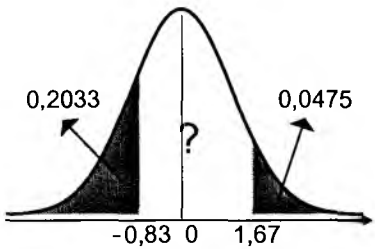


Para $x = 2$, temos: $z = {}^{(2 - 2{,}5)}/_{0{,}6} = -0{,}83$

e para $x = 3{,}5$, temos: $z = {}^{(3{,}5 - 2{,}5)}/_{0{,}6} = 1{,}67$.

A figura ao lado ilustra estas transformações.

Usando a Tabela IV do apêndice, encontramos para $z = -0{,}83$ e $z = 1{,}67$ as respectivas áreas nas extremidades da curva: 0,2033 e 0,0475 (lembrando que para valores negativos de z, como $-0{,}83$, procuramos na Tabela IV 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: $P(2 < X < 3{,}5) =$
$= 1 - (0{,}2033 + 0{,}0475) = 0{,}7492$.



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) Considerando a variável aleatória X do Exemplo 8.3, calcule:

a) $P(X > 190)$; **b)** $P(150 < X < 190)$; **c)** $P(X \leq 160)$;

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*?

8.3 DADOS OBSERVADOS E O MODELO NORMAL

A Figura 8.8 mostra um histograma de freqüências das médias diárias de pressão intra-ocular, 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.

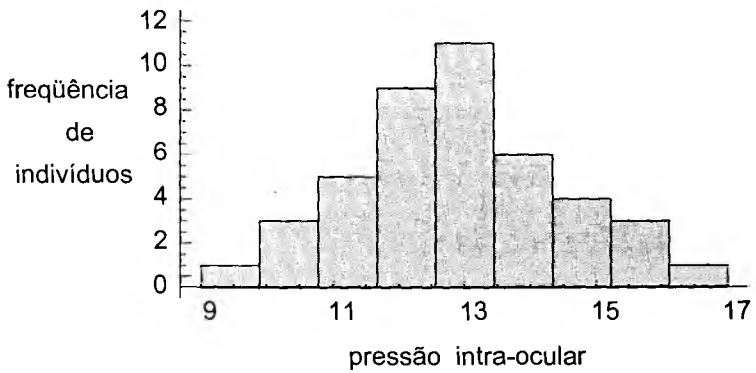


Figura 8.8 Histograma de freqüências das médias diárias de pressão intra-ocular, numa amostra de 43 indivíduos sadios.

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, onde cada valor corresponde à média aritmética de sete medidas de pressão intra-ocular, 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, pois elas podem ser vistas como *somas* de uma infinidade de componentes inerentes ao indivíduo e ao seu meio.

Quando temos dados observados de uma certa variável, que acreditamos ter distribuição aproximadamente normal, podemos usar algumas propriedades desta distribuição na análise dos dados. 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, a área sob a curva atinge, aproximadamente, 0,683; ao afastar dois desvios padrão, a área cresce para 0,955 e o afastamento de três desvios padrão gera uma área de 0,997 (veja a Figura 8.9).

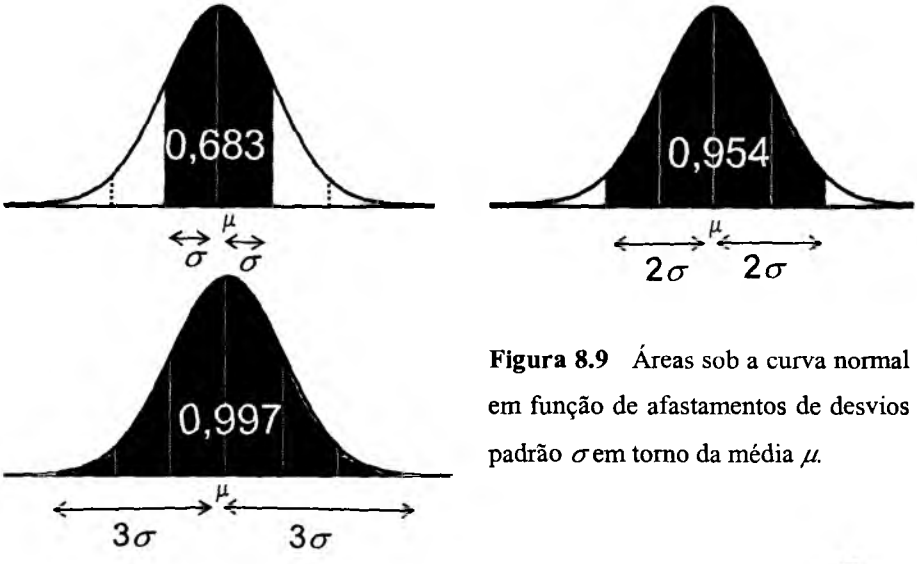


Figura 8.9 Áreas sob a curva normal em função de afastamentos de desvios padrão σ em torno da média μ.

Dado um conjunto de valores, podemos calcular a média $\overline{X}$ e o desvio padrão S, como vimos no Capítulo 6. Se estes valores se distribuem de forma parecida com um modelo normal, devemos esperar, pela propriedade que enunciamos, que ocorram:

em torno de 95% dos valores no intervalo $\overline{X} \pm 2S$ (isto é, no intervalo de $\overline{X} - 2S$ até $\overline{X} + 2S$) e

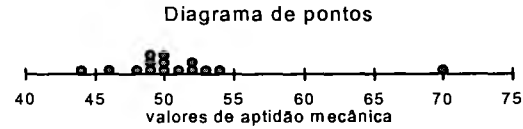
mais de 99% dos valores no intervalo $\overline{X} \pm 3S$ (isto é, no intervalo de $\overline{X} - 3S$ até $\overline{X} + 3S$).

Assim, algum valor que esteja fora do intervalo $\overline{X} \pm 3S$ pode ser considerado como um valor discrepante dos demais. E valores fora do intervalo $\overline{X} \pm 2S$ podem ser vistos como *valores suspeitos*.

Exemplo 8.7 Considere os seguintes valores, obtidos pela aplicação de um teste de aptidão mecânica, numa turma de estudantes de primeiro grau.

Valores:

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 valores comportam-se de maneira compatível com um modelo normal. Calculando a média aritmética e o desvio padrão destes dados, temos:

$\overline{X} = 51{,}1$ pontos e $S = 5{,}8$ pontos (veja as fórmulas de $\overline{X}$ e S no Capítulo 6).

Donde:

$\overline{X} \pm 2S = 51{,}1 \pm 2(5{,}8) = 51{,}1 \pm 11{,}6 \longrightarrow$ intervalo de 39,5 a 62,7 pontos;

$\overline{X} \pm 3S = 51{,}1 \pm 3(5{,}8) = 51{,}1 \pm 17{,}4 \longrightarrow$ intervalo de 33,7 a 68,5 pontos.

Verificamos que todos os valores estão no intervalo $\overline{X} \pm 2S$, com exceção do valor 70. Aliás, o 70 também não pertence ao intervalo $\overline{X} \pm 3S$, caracterizando um *ponto discrepante*. A criança que obteve o valor 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 distribuição binomial, a distribuição normal serve como uma boa aproximação. Seja, por exemplo, o problema de amostragem e as variáveis aleatórias binomiais X e Y definidas na Figura 8.10.

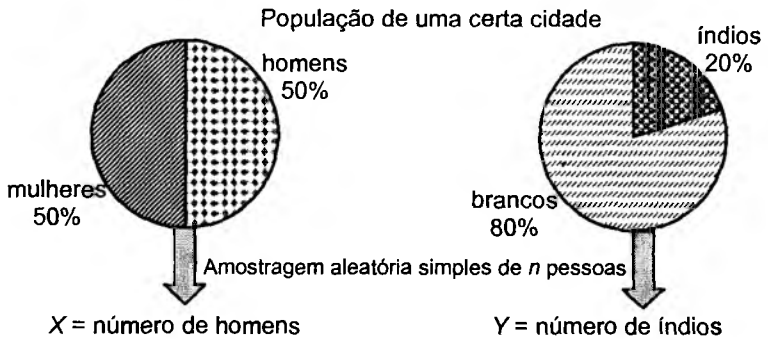


Figura 8.10 Ilustração de duas variáveis aleatórias binomiais.

A variável aleatória X tem distribuição binomial com $\pi = 0{,}5$ e Y tem distribuição binomial com $\pi = 0{,}2$. A Figura 8.11 apresenta as distribuições de probabilidades de X e Y considerando $n = 1$, 10 e 50.

Observando a Figura 8.10, verificamos 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, muitas vezes usada, considera a aproximação razoável se as duas seguintes inequações forem satisfeitas:

$$\text{(a) } n\pi \geq 5 \qquad \text{e} \qquad \text{(b) } n(1-\pi) \geq 5.$$

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:

$$\mu = n\pi \qquad \text{e} \qquad \sigma = \sqrt{n\pi(1-\pi)}$$

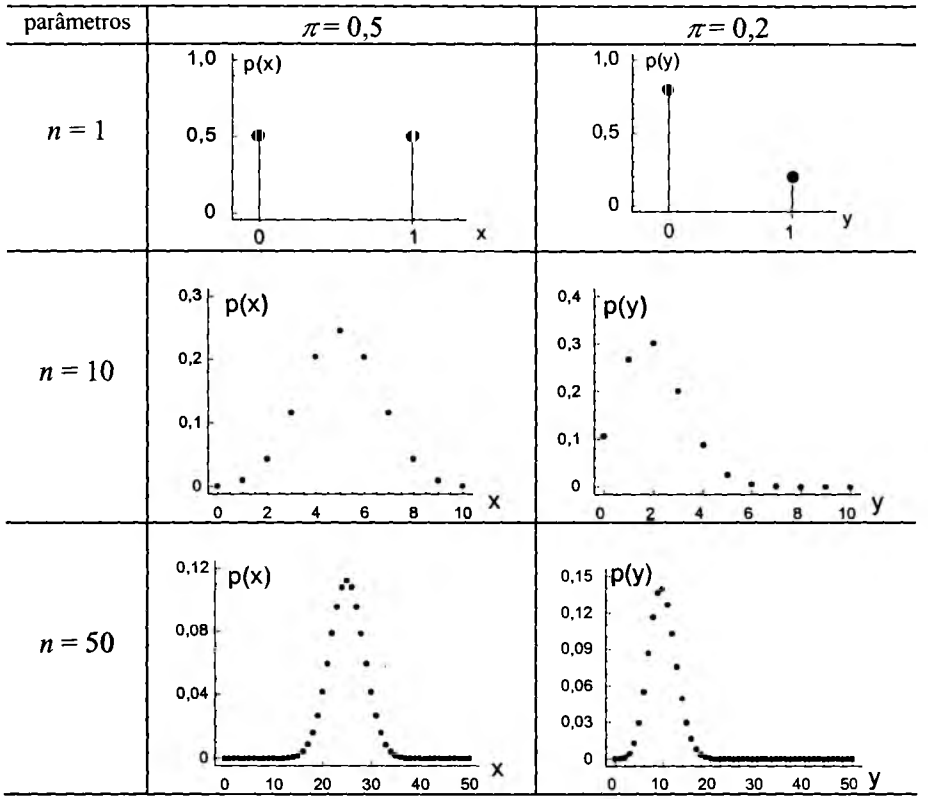


Figura 8.11 Distribuições binomiais para diferentes valores de n e π.

Exemplo 8.8 Observar o número, Y, de respostas favoráveis, numa amostra aleatória de $n = 50$ pessoas, indagadas a respeito da opinião (favorável ou contrária) sobre um projeto municipal. Admita que na população existam 40% de favoráveis.

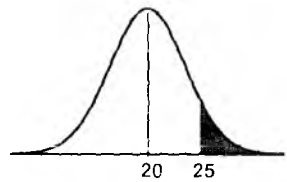
Pelas características do experimento, a variável aleatória Y tem distribuição binomial com parâmetros n = 50 e $\pi = 0,4$. Como n é grande e π não é um valor muito próximo de *zero* ou de *um*, podemos usar a aproxima-

ção normal.[4] Esta distribuição normal deve ter média μ e desvio padrão σ dados, respectivamente, por

$$\mu = n\pi = 50(0,4) = 20 \quad \text{e}$$

$$\sigma = \sqrt{n\pi(1-\pi)} = \sqrt{50(0,4)(1-0,4)} = 3,464$$

Calculemos, como exemplo, a probabilidade de ocorrer o evento *25 ou mais favoráveis na amostra*. Esta probabilidade pode ser aproximada por uma área sob a curva da distribuição normal de média μ = 20 e desvio padrão σ = 3,464, como ilustra a figura ao lado.



O valor x = 25, da distribuição normal de μ = 20 e σ = 3,464 corresponde ao seguinte valor padronizado:

$$z = \frac{x-\mu}{\sigma} = \frac{25-20}{3,464} = 1,44$$

Usando a Tabela IV (apêndice), encontramos a probabilidade 0,0749.

Correção de continuidade

Ao calcular probabilidades de eventos oriundos de experimentos binomiais como áreas sob uma curva normal, estamos procedendo uma aproximação de uma variável aleatória discreta, que só assume valores inteiros, para uma variável contínua, cujos eventos constituem intervalos de números reais. Neste contexto, devemos fazer alguns ajustes, como ilustra o exemplo seguinte.

Exemplo 8.9 Seja Y o número de caras obtidas em 10 lançamentos de uma moeda perfeitamente equilibrada.

---

[4] Poderíamos usar a regra prática: (a) $n\pi$ = (50)(0,4) = 20 e (b) $n(1-\pi)$ = (50)(1-04) = 30. Como as expressões (a) e (b) levam a valores não inferiores a 5, podemos usar a aproximação normal.

Pelas características do experimento, podemos deduzir que Y tem distribuição binomial com $n = 10$ e $\pi = 0,5$, que pode ser aproximada pela distribuição normal de média e desvio padrão dados por

$$\mu = n\pi = 10(0,5) = 5 \quad \text{e} \quad \sigma = \sqrt{n\pi(1-\pi)} = \sqrt{10(0,5)(1-0,5)} = 1,58$$

Considere o evento *ocorrer quatro caras*, que pode ser escrito como $\{Y = 4\}$. Ao expressar este evento em termos de uma variável aleatória contínua X, com distribuição normal, devemos considerar um intervalo em torno do valor 4, pois, para variáveis contínuas, como já discutimos, 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 ilustra a Figura 8.12.

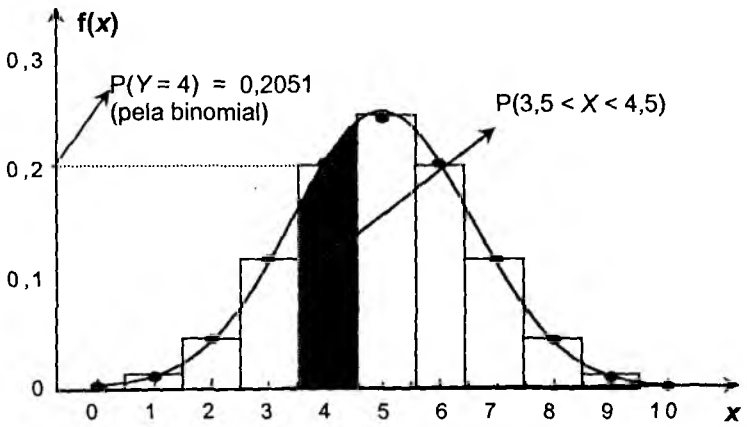


Figura 8.12 Aproximação da probabilidade do evento $\{Y = 4\}$ (da distribuição binomial) para a probabilidade do evento $\{3,5 < X < 4,5\}$ (da distribuição normal).

Usando adequadamente a distribuição normal, encontramos a probabilidade do evento $\{3,5 < X < 4,5\}$ como sendo igual a 0,2034. (Exercício: verifique o cálculo desta probabilidade).[5] Se fosse usada diretamente a distribuição binomial, chegaríamos à probabilidade igual a

[5] Neste caso, podemos usar a aproximação normal, porque $n\pi = 5$ e $n(1-\pi) = 5$, satisfazendo o critério para a aproximação.

0,2051 (Tabela II do apêndice), donde verificamos que o resultado oriundo da curva normal é bastante satisfatório.

O procedimento de subtrair e somar meia unidade, para construir um intervalo em torno de valores inteiros, é conhecido como *correção de continuidade*. Esta correção é recomendável ao aproximar uma probabilidade da distribuição binomial por uma área sob a curva normal, especialmente se o número de ensaios n não for muito grande.

Exercícios

8) Com respeito ao Exemplo 8.9, 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) Ainda com respeito ao Exemplo 8.9 calcule, pela distribuição normal, a probabilidade de ocorrer o evento *5 ou mais caras*.

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. Que proporção das notas
a) excede 85?
b) está 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 5 elementos.
b) que a amostra tenha sido de 50 elementos.

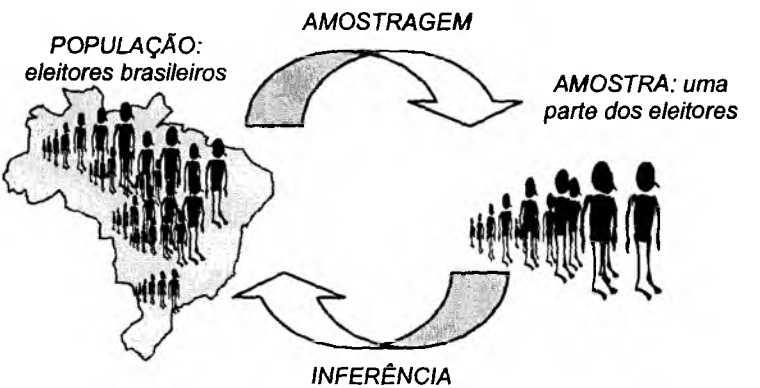
15) **a)** Um exame de múltipla escolha consiste em 10 questões, cada uma com 4 possibilidades de escolha. A aprovação exige no mínimo 50% de acertos.

Qual é a chance de aprovação se o candidato comparece ao exame sem saber absolutamente nada, apelando apenas para o "palpite"?

b) Um exame de múltipla escolha consiste em 100 questões, cada uma com 4 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"?

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 a partir de dados observados

Estimação de parâmetros

Neste capítulo, estudaremos o problema de avaliar certas características dos elementos da população, a partir de operações com os dados de uma amostra. É um raciocínio tipicamente indutivo, em que se generalizam resultados *da parte* (amostra) para *o todo* (população). Este procedimento é denominado estimação de parâmetros, e está ilustrado na Figura 9.1.

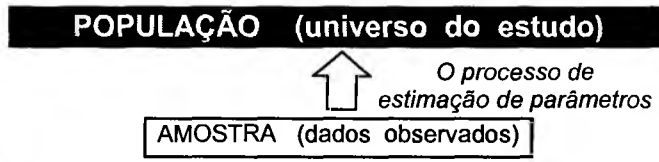


Figura 9.1 O raciocínio indutivo da estimação.

Vamos relembrar algumas definições.

Parâmetro: alguma característica descritiva dos elementos da população, como por exemplo, a média de alguma variável, a proporção de algum atributo, etc.

Estatística: alguma operação com os dados de uma amostra. Esta operação pode ser o cálculo de uma média ou de uma proporção.

A estatística, quando usada com o objetivo de avaliar, ou *estimar*, o valor de algum parâmetro, também é chamada de ***estimador***.

Exemplo 9.1 A prefeitura de uma cidade pretende avaliar a aceitação de certo projeto educacional. Depois de apresentá-lo aos moradores do município, os responsáveis por sua execução desejam avaliar o valor aproximado do parâmetro π = *proporção de favoráveis ao projeto, dentre os indivíduos residentes no município*. Para estimar este parâmetro, a prefeitura planeja observar uma amostra aleatória simples de $n = 400$ moradores e calcular o

valor da estatística P = *proporção de moradores favoráveis ao projeto na amostra* (veja a Figura 9.2).

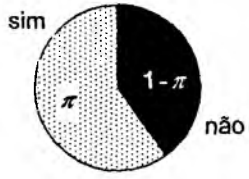
POPULAÇÃO	AMOSTRA
- todos os moradores do município; - os elementos da população estão divididos em *favoráveis* e *contrários* ao projeto; - parâmetro de interesse: π = *proporção de favoráveis.*  	- n moradores do município, selecionados aleatoriamente; - cada elemento da amostra é classificado como *favorável* ou *contrário* ao projeto; - estatística: P = *proporção de favoráveis na amostra*, isto é: $P = \dfrac{\text{nº de favoráveis na amostra}}{n}$
Qual o valor de π? ← *processo de estimação*	$\pi = P \pm$ *erro amostral*

Figura 9.2 Ilustração de um problema de estimação.

O termo *erro amostral*, que aparece na Figura 9.2, corresponde à diferença entre a estatística P e o parâmetro π.

Exemplo 9.2 Para estudar o efeito da merenda escolar, introduzida nas escolas de um grande 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, dentre todas as crianças da rede municipal de ensino, durante o primeiro ano letivo.* 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.

Quando estivermos estudando a incidência de algum atributo numa certa população, geralmente o interesse reside no parâmetro *proporção, ou percentagem, de elementos com este atributo* (é o caso do

Exemplo 9.1). Por outro lado, quando estamos pesquisando alguma característica quantitativa, como no Exemplo 9.2, torna-se mais comum o interesse em estimar o parâmetro *quantidade média da característica em questão*.

Apresentamos, a seguir, alguns parâmetros e as respectivas estatísticas, que geralmente são usadas para estimá-los. 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.

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 sobre 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.

Ao observar uma particular amostra, podemos calcular o valor da estatística que estamos usando como estimador. O valor encontrado é chamado de ***estimativa***. Por exemplo, se na amostra de $n = 400$ moradores do Exemplo 9.1 encontrarmos 240 favoráveis, temos a seguinte estimativa para o parâmetro π:

$$P = {}^{240}/_{400} = 0{,}60 \text{ (ou, 60\%)}$$

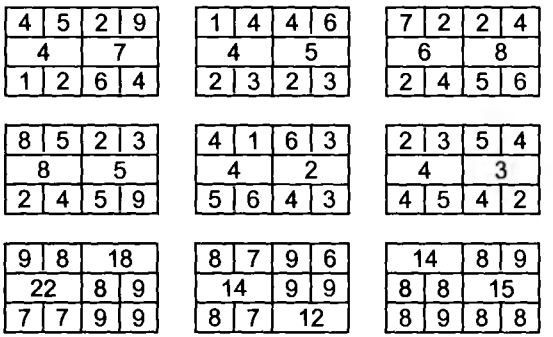
Contudo, não devemos esperar que este valor coincida com o valor do parâmetro π, pois haverá uma variação devido ao que chamamos de *erro amostral*, como foi ilustrado na Figura 9.2.

Dizemos que uma estimativa é tão mais ***precisa*** quanto menor for o seu erro amostral. Um dos principais objetivos na teoria da estimação é estimar um *limite superior provável* para o erro amostral. Este valor será a base para avaliarmos a precisão de nossa estimativa. Neste capítulo, nos preocuparemos em avaliar a precisão de estimativas de parâmetros do tipo π (*proporção de algum atributo*) e do tipo μ (*média de alguma variável quantitativa*).

Toda a formulação que apresentaremos, parte da suposição de que os dados em análise constituam uma amostra aleatória simples da população de interesse, como definido no Capítulo 3.

Exercícios

1) O esquema seguinte representa uma população de 90 domicílios, situados em quadras residenciais. Os valores dentro dos quadradinhos (domicílios) indicam o número de cômodos do respectivo domicílio.



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 20 domicílios da população do Exercício 1.[1] 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.*

9.1 DISTRIBUIÇÃO AMOSTRAL DA PROPORÇÃO

Considere a seguinte pergunta, relativa ao Exemplo 9.1: O valor de P (*proporção de favoráveis numa amostra de n = 400 moradores*) vai ser um valor próximo da verdadeira proporção π, a qual refere a todos os moradores do município?

---

[1] Se você não se lembrar de como extrair uma amostra aleatória simples, leia novamente a Seção 3.1 (Capítulo 3).

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.***

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 9.3 mostra o modelo de probabilidades, referente a cada observação, admitindo o verdadeiro valor de π conhecido e igual a 0,70.

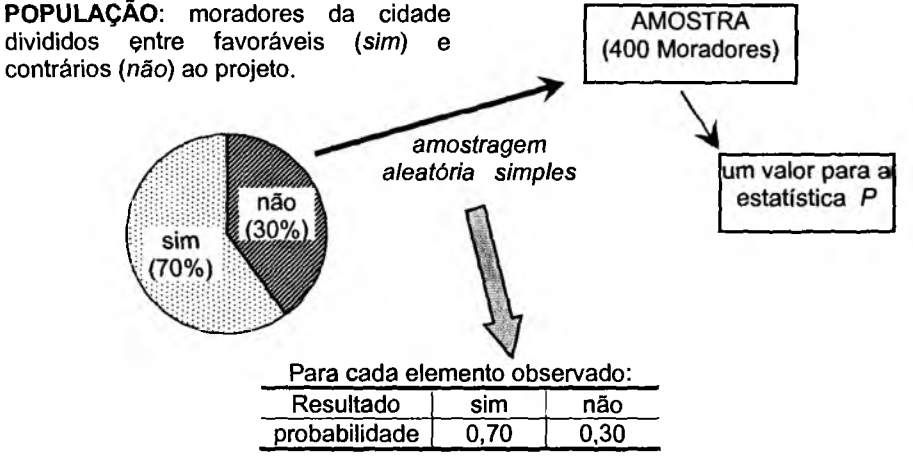


Figura 9.3 Modelo de probabilidades associado ao processo de amostragem do Exemplo 9.1, com $\pi = 0{,}70$.

Uma simulação

Para ilustrarmos a distribuição amostral de P, conforme a situação da Figura 9.3, podemos simular várias amostras de tamanho $n = 400$, segundo o modelo especificado. A simulação pode ser executada com o apoio de uma tabela de números aleatórios (Tabela I do apêndice). Cada

número de um algarismo, observado na tabela, simula a observação de um elemento da população, da seguinte forma.

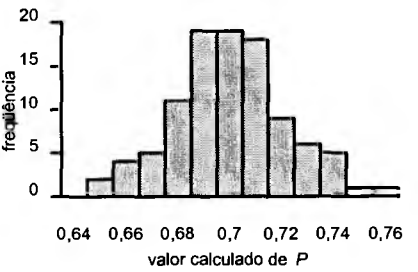
• Quando o algarismo extraído da tabela de números aleatórios for um valor do conjunto {0,1,2,3,4,5,6}, que acontece com probabilidade $^{7}/_{10}$, simula a observação de um indivíduo *favorável* ao projeto.

• Quando o algarismo extraído da tabela de números aleatórios for um valor do conjunto {7,8,9}, que acontece com probabilidade $^{3}/_{10}$, simula a observação de um indivíduo *contrário* ao projeto.

Ao observarmos 400 algarismos da tabela de números aleatórios, podemos calcular o valor de P = "proporção de números encontrados 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, através dela, termos informações sobre o erro amostral, precisamos repetir este processo várias vezes, sob as mesmas condições. Os valores da Figura 9.4 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				

Histograma de freqüências



* Valor máximo e valor mínimo.

Figura 9.4 Cem observações da distribuição amostral de P, considerando amostras de tamanho $n = 400$ e $\pi = 0,70$.

Pela Figura 9.4, verificamos que em nenhuma amostra, dentre as 100 simuladas, resultou um valor de P fora do intervalo de 0,65 a 0,76. Como, nesta situação fictícia, sabemos o valor de π (igual a 0,70), podemos afirmar que em nenhuma das amostras simuladas o erro amostral teve

magnitude superior a 0,06 (atingido por uma amostra que acusou P igual a 0,76 e, portanto, 0,76 - 0,70 = 0,06). Desta forma, 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 executada, 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 este valor, pois sempre existe o efeito do *azar* ao sortearmos os elementos que irão compor a amostra. Neste contexto, as afirmações são sempre feitas em termos de 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, acusaram erros amostrais inferiores a 0,05 (veja a Figura 9.4). Neste contexto, 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 $^{96}/_{100}$, isto é, em torno de 96%.

Teoria

Na maioria dos problemas de estimação de parâmetros não é necessário executar simulações para avaliar a precisão de uma estimativa. Por exemplo, em problemas de estimação de uma proporção, a partir de uma amostra aleatória simples, o experimento é tipicamente *binomial*, com parâmetros n (tamanho da amostra) e π (proporção do atributo em questão). Sabemos, pelo capítulo anterior, que se n for grande, a distribuição binomial se aproxima de uma *distribuição normal*, com média e desvio padrão determinados a partir de n e π, da seguinte forma:[2]

$$\mu_p = \pi \quad \text{e} \quad \sigma_p = \sqrt{\frac{\pi \cdot (1-\pi)}{n}}$$

[2] Estamos usando o sub-índice p nas notações usuais de média e desvio padrão, μ e σ, para lembrar que estes parâmetros referem-se à distribuição amostral de P.

A Figura 9.5 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, sabemos que existe 95% de probabilidade, de um valor ser observado a menos de 1,96 desvios padrão da média (Exemplo 8.5, Capítulo 8). Desta forma, se exigirmos nível de 95% de confiança, podemos explicitar um limite superior provável para o erro amostral, considerando a faixa de 1,96 desvios padrão, acima e abaixo do centro da distribuição, como mostra a Figura 9.6.

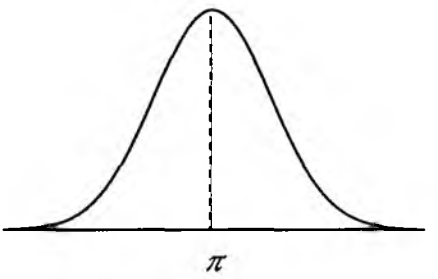


Figura 9.5 Forma aproximada da distribuição amostral de P.

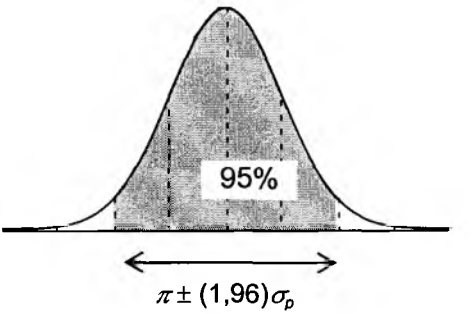


Figura 9.6 Faixa onde devem ocorrer aproximadamente 95% dos valores de P.

9.2 ESTIMAÇÃO DE UMA PROPORÇÃO

No que segue, limitou-se 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.[3]

O *desvio padrão* da distribuição amostral de P, σ_p, também conhecido como ***erro padrão*** **de** ***P***, pode ser estimado pelos dados da amostra, usando a expressão

$$S_p = \sqrt{\frac{P.(1-P)}{n}}$$

onde P é a proporção do atributo na amostra.

[3] Desde que π não seja próximo de 0 ou de 1, podemos usar a distribuição normal para $n \geq 30$. Para um maior detalhamento sobre esta aproximação, veja a Seção 8.4.

Nível de 95% de confiança

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 9.6, temos, aproximadamente, 95% de probabilidade de o valor de P cair a menos de 1,96 desvios padrão de π.

Exemplo 9.1 (continuação) Admita que na amostra de $n = 400$ elementos, encontramos 60% de favoráveis. Temos, então, $P = 0,60$ (ou 60%) e erro padrão de P dado por

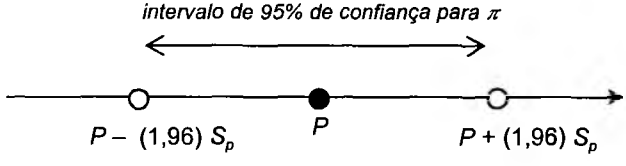
$$S_p = \sqrt{\frac{P(1-P)}{n}} = \sqrt{\frac{(0,60)(0,40)}{400}} = 0,0245$$

Usando nível de confiança de 95%, temos um erro amostral máximo provável de $(1,96)S_p = (1,96)(0,0245) = 0,048$ (ou 4,8%). Desta forma, podemos dizer que o intervalo: 60,0% ± 4,8% (isto é, o intervalo de 55,2% a 64,8%) contém, com 95% de confiança, o parâmetro π = *proporção de favoráveis em toda a população de moradores do município.*

O intervalo centrado em P e com semi-amplitude $(1,96)S_p$, ou seja:

$$P \pm (1,96)S_p$$

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

Arbitrado um nível de confiança, podemos obter o limite provável para o erro amostral, multiplicando S_p por um determinado valor z da curva normal padrão. A Figura 9.7 mostra uma tabela, construída a partir da

Tabela IV do apêndice (tabela da distribuição normal padrão), que associa os níveis de confiança mais usados, com 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 9.7 Valores de z para alguns níveis de confiança.

Fixado o nível de confiança, podemos obter o correspondente valor de z, como ilustra a Figura 9.7 e, a partir daí, calcular a estimativa do erro amostral máximo provável, $z\,S_p$, e o intervalo de confiança para π:

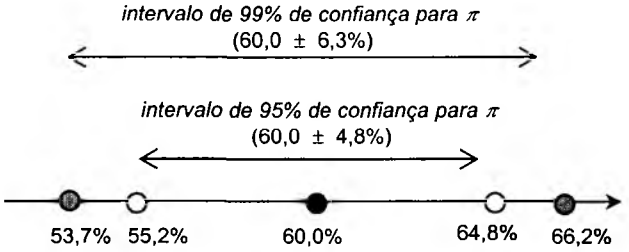
$$P \pm z\,S_p$$

Exemplo 9.1 (continuação) No exemplo em questão, poderíamos querer um nível de 99% de confiança. Então, pela tabela da Figura 9.7, temos que área = 0,99 implica z = 2,576, resultando no seguinte limite provável para o erro amostral: S_P = (2,576).(0,0245) = 0,063 (ou 6,3%). Então, com 99% de confiança, o seguinte 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%, referente 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.

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 nesta simulação, discuta sobre o erro amostral, associado a uma amostra de tamanho $n = 10$, para estimar o parâmetro π, relativo a algum atributo de uma grande população.

4) Considerando o Exemplo 9.1, faça as seguintes modificações, executando, em cada caso, um intervalo de confiança para o parâmetro π. 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 $n = 1000$ moradores, acusando 600 favoráveis.

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, dentre os consumidores em potencial de uma certa cidade, a proporção π de consumidores que passariam a usar certo produto, após experimentá-lo pela primeira vez. Para atingir este 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. Com uma leitura cuidadosa das redações, torna-se possível verificar se cada vestibulando é, ou não, favorável à mudança.

a) Foram observadas 400 redações, extraídas por sorteio, dentre todas as redações. Nesta 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 observada*? 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). Nesta 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) 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 nesta 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.*

8) A amostra descrita no Exercício 7 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 estes intervalos.
NOTA: Observe que, ao trabalhar com subgrupos de uma amostra (Exercício 8), as precisões das estimativas tendem a ser piores (intervalos de confiança mais longos), quando comparadas com à análise de toda a amostra.

9.3 ESTIMAÇÃO DE UMA MÉDIA

Para estimar o parâmetro μ (média de alguma variável quantitativa), a partir de $\bar{X}$ (média da variável observada numa amostra aleatória simples), podemos seguir os mesmos princípios da estimação de uma proporção, pois, para amostras grandes, a distribuição amostral de $\bar{X}$, também se aproxima de uma distribuição normal.

O ***erro padrão da média amostral*** pode ser estimado, a partir do desvio padrão amostral, S, segundo a expressão[4]

$$S_{\bar{X}} = \frac{S}{\sqrt{n}} \qquad \text{onde} \quad S = \sqrt{\frac{\Sigma X^2 - n\bar{X}^2}{n-1}}$$

[4] O cálculo do desvio padrão de conjunto de dados, S, foi visto no Capítulo 6.

-*Amostras grandes*

Quando temos uma amostra grande, podemos avaliar o *erro amostral máximo provável* por $z\,S_{\overline{X}}$, onde z pode ser obtido pelo esquema da Figura 9.7, em função do nível de confiança desejado.[5]

Exemplo 9.2 (continuação) Observando uma amostra aleatória simples de $n = 100$ crianças do primeiro ano letivo, nas escolas municipais, em que se estava servindo uma merenda especial, encontraram-se as seguintes estatísticas relativas à variável *ganho de peso ao longo do ano.*

Ganho médio de peso das crianças da amostra: $\overline{X} = 6{,}0$ kg;

Desvio padrão dos pesos das crianças da amostra: $S = 2{,}0$ kg.

Com o objetivo de estimar o parâmetro μ = *ganho médio de peso da população*, podemos calcular uma estimativa para o erro padrão da média amostral

$$S_{\overline{X}} = \frac{S}{\sqrt{n}} = \frac{2{,}0}{\sqrt{100}} = 0{,}2\,\text{kg}$$

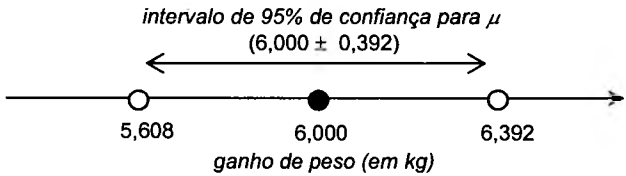
e o erro amostral máximo provável (95% de confiança)

$$(1{,}96)(0{,}2) = 0{,}392\ \text{kg}$$

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 a 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 da rede municipal de ensino.[6]



---

[5] O uso do valor z, como indicado na Figura 9.7, só é válido para amostras grandes (digamos, $n \geq 30$). Posteriormente vamos apresentar uma expressão mais geral, que vale também para amostras pequenas.

[6] Note que o intervalo de confiança de uma média é apresentado na mesma unidade de medida dos dados observados.

Amostras pequenas

Quando dispomos de uma amostra pequena (digamos, $n < 30$), não temos a garantia de que a distribuição amostral da média se aproxime de uma distribuição normal. Porém, se a variável em estudo tiver uma distribuição razoavelmente simétrica, parecida com uma normal, a teoria estatística mostra que é possível construir estimativas intervalares para a média populacional, μ, utilizando uma certa distribuição, denominada de ***t de Student***, que também é tabelada (Tabela V do apêndice).

A *distribuição t*, como mostra a Figura 9.8, 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 se aproximando da normal padrão para amostras grandes. Em geral, a distribuição é apresentada em função de um parâmetro, denominado graus de liberdade, *gl*, definido, no caso de estimação de uma média, por $gl = n - 1$.

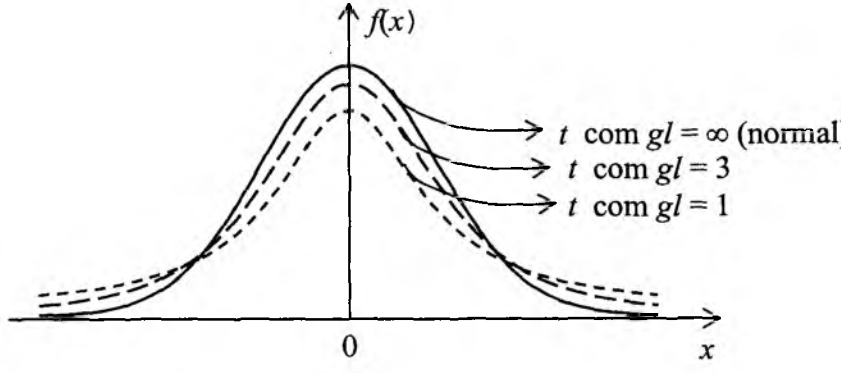


Figura 9.8 Gráficos de distribuições *t de Student* e da normal padrão.

Para obter 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 V do apêndice. Por exemplo, para $gl = 9$ e nível de confiança de 95%, devemos usar a Tabela V, como mostra a Figura 9.9.

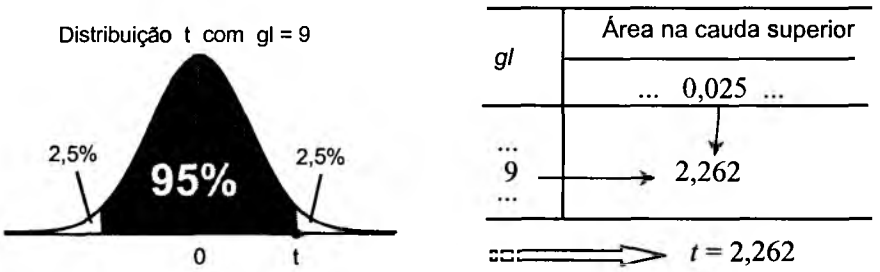


Figura 9.9 Uso da tabela da distribuição *t de Student*. Ilustração com $gl = 9$ e nível de confiança de 95%.

O intervalo de confiança para uma média μ tem a seguinte expressão geral:

$$\overline{X} \pm t\, S_{\overline{X}}$$

Exemplo 9.3 Para verificar a eficácia de um programa de prevenção de acidentes de trabalho, fez-se um estudo experimental, implementando este programa em dez empresas da construção civil, escolhidas ao acaso, numa certa região. Os dados abaixo referem-se aos *percentuais de redução de acidentes de trabalho* nas 10 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, devido ao programa preventivo, em todas as empresas da construção civil da região*. Podemos obter uma estimativa para o erro padrão da média, como segue:

$$S_{\overline{X}} = \frac{6{,}65}{\sqrt{10}} = 2{,}10$$

Usando nível de 95% de confiança, graus de liberdade $gl = 9$ (pois, $n = 10$ e $gl = n - 1$), obtemos na Tabela V (apêndice) o valor $t = 2{,}262$, donde

podemos calcular o erro máximo provável, $t.S_{\bar{X}} = (2{,}262).(2{,}10) = 4{,}75 \approx 4{,}8$. Então, temos o seguinte intervalo de 95% de confiança para o parâmetro μ:

$$18{,}0 \pm 4{,}8 \text{ pontos percentuais}^7$$

Exercícios

9) A tabela seguinte mostra os valores das médias e desvios padrão da renda familiar, de uma amostra de 120 famílias, do bairro Saco Grande II, dividida em três localidades. Os dados foram obtidos 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 95% de confiança, para a renda familiar média de cada localidade. Interprete as estimativas.

10) 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).

11) 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 valor de μ, calculado no Exercício 1, pertence a este intervalo.

12) 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*. Usando os dados desta 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 valor do parâmetro μ e verifique se o intervalo que você construiu no item (a) contém o valor deste parâmetro. Consulte seus colegas de sala. Verifique quantos obtiveram intervalos de confiança contendo o valor do parâmetro μ.

---

[7] O intervalo foi colocado em termos da unidade *pontos percentuais* porque era esta a unidade dos dados originais (redução percentual de acidentes de trabalho).

9.4 CORREÇÕES PARA TAMANHO DA 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 introduzir o seguinte fator de redução, na estimativa do erro padrão:

$$\sqrt{\frac{N-n}{N-1}}$$

Temos, então, as seguintes expressões para estimativas de erros padrão:

para estimar uma proporção π	para estimar uma média μ
$S_p = \sqrt{\frac{P.(1-P)}{n}} \cdot \sqrt{\frac{N-n}{N-1}}$	$S_{\bar{x}} = \frac{S}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$

Exemplo 9.4

a) Vamos refazer o Exemplo 9.3, considerando que existam $N = 30$ empresas na região. Neste caso:

$$S_{\bar{x}} = \frac{S}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} = (2{,}10) \cdot \sqrt{\frac{30-10}{30-1}} = (2{,}10) \cdot (0{,}83) = 1{,}74$$

$$t.S_{\bar{x}} = (2{,}262) \cdot (1{,}74) \approx 3{,}9$$

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:

$$S_{\bar{x}} = \frac{S}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} = (2{,}10) \sqrt{\frac{400-10}{400-1}} = (2{,}10) \cdot (0{,}99) = 2{,}08$$

$$t.S_{\bar{x}} = (2{,}262).(2{,}08) = 4{,}7$$

E o intervalo de 95% de confiança para a média μ:

$$18{,}0 \pm 4{,}7 \text{ pontos percentuais.}$$

Comparando o Exemplo 9.4 com o 9.3, verificamos que a inclusão do tamanho da população, N, no cálculo do erro padrão, somente acarretou alteração relevante no caso (a). Quando N é bem superior a n, como no Exemplo 9.4b, podemos usar as mesmas fórmulas desenvolvidas na seção anterior, pois, o resultado final praticamente não vai depender do tamanho, N, da população.

Exercícios

13) 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.

14) Refazer os Exercícios 11 e 12, considerando o tamanho da população.

9.5 TAMANHO MÍNIMO DE UMA AMOSTRA ALEATÓRIA SIMPLES

No Capítulo 3, descrevemos algumas técnicas para 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, considerando o processo de amostragem aleatória simples.

As fórmulas para o cálculo do tamanho, n, da amostra são obtidas das expressões dos intervalos de confiança fixando, *a priori*, o nível de confiança e o erro amostral tolerado. Admitiremos, 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 $\bar{X}$ e de P.

Tendo o valor z da distribuição normal, em função do *nível de confiança desejado*, como também o valor E_0 relativo ao *erro amostral tolerado*, podemos usar o seguinte procedimento para a determinação de n.

Uma primeira aproximação para o cálculo do tamanho da amostra, em função do parâmetro a ser estimado:

• para estimar uma proporção π	• para estimar uma média μ
$n_0 = \dfrac{z^2 \pi(1-\pi)}{E_0^2}$	$n_0 = \dfrac{z^2 \sigma^2}{E_0^2}$

Quando se conhece o tamanho da população, pode-se fazer a seguinte correção para se ter o tamanho da amostra (expressão aproximada):

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

Se a população é grande pode-se adotar o valor de n_0 como o tamanho n da amostra.

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 $\sigma^2 = \pi.(1 - \pi)$.

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. Esta estimativa pode ser obtida de algum estudo anterior, ou com a realização de uma amostra piloto.[8]

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 acusou desvio padrão $S = 1{,}95$ kg. Fixando o nível de confiança em 95%, e tolerando um erro amostral de até 200 gramas (isto é, $E_0 = 0{,}2$ kg), podemos, então, determinar o tamanho da amostra.

Solução: $z = 1{,}96$ (pois, vamos trabalhar com nível de 95% de confiança) e usaremos no lugar de σ^2 o valor da variância amostral: $S^2 = (1{,}95)^2 = 3{,}8$. Donde temos o seguinte cálculo para tamanho mínimo de uma amostra aleatória simples:

$$n = \frac{z^2 \cdot \sigma^2}{E_0^2} \approx \frac{z^2 \cdot S^2}{E_0^2} = \frac{(1{,}96)^2 \cdot (3{,}8)}{(0{,}2)^2} = 365 \text{ crianças}$$

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 no lugar de S^2. No Exemplo 9.5, usando $z = 2$, obtém-se como resultado $n = 380$ crianças.

Argumentos teóricos

Muitas vezes, pela forma de mensuração da variável em estudo, torna-se 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: $\sigma^2 = \pi.(1 - \pi)$, e pode-se provar matematicamente que o valor desta expressão nunca será superior a $\frac{1}{4}$ (um quarto), como mostra a Figura 9.10.

---

[8] O termo "amostra piloto" refere-se a um trabalho inicial de observação de alguns elementos da população, com o objetivo de se obter algumas estimativas iniciais, que possam facilitar o trabalho de planejamento da pesquisa. Por exemplo, o cálculo da variância destes dados, S^2, para usar no lugar de σ^2, no cálculo do tamanho da amostra.

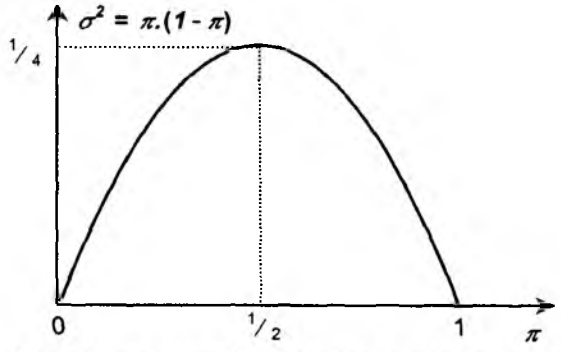


Figura 9.10 O parâmetro σ^2 em função do valor da proporção π.

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 $\frac{1}{2}$, podemos usar, no lugar de σ^2, o seu valor máximo, $\frac{1}{4}$. Donde temos a seguinte expressão para o cálculo do tamanho da amostra:

$$n' = \frac{z^2 \cdot \frac{1}{4}}{E_0^2} = \frac{z^2}{4.E_0^2}$$

O valor de n' deverá ser maior ou igual ao valor de n (tamanho mínimo da amostra). Se o valor do parâmetro π, a ser estimado, estiver próximo de $\frac{1}{2}$, então o valor de n' é uma boa aproximação para o valor de n.

A expressão de n' também é bastante útil quando se deseja obter um tamanho de amostra, capaz de garantir uma certa precisão, para estimativas de várias proporções (vários π's), como geralmente ocorre em pesquisas do tipo levantamento. Usando o nível usual de 95% de confiança, temos $z \approx 2$. A fórmula de n' reduz-se a

$$n' = \frac{1}{E_0^2}$$

Esta expressão já foi vista no Capítulo 3, como uma fórmula usual para o cálculo do tamanho n da amostra.

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: $n \approx n' = \dfrac{1}{E_0^2} = \dfrac{1}{(0{,}02)^2} = 2.500$ 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, admitindo que, na população em estudo, não devam existir mais que 20% de indivíduos infectados?

Solução: Sabemos que $\pi \leq 0{,}20$; então, o valor máximo de σ^2 é (veja a Figura 9.9): $\pi(1-\pi) = (0{,}20).(1-0{,}20) = 0{,}16$. Donde

$$n = \frac{z^2.\pi(1-\pi)}{E_0^2} \approx \frac{(1{,}645)^2.(0{,}16)}{(0{,}01)^2} = 4.330 \text{ indivíduos}$$

Quando o tamanho N da população for conhecido, pode-se fazer uma correção no cálculo do tamanho da amostra. Esta correção foi discutida na Seção 3.4 (Capítulo 3).

Exercícios

15) Com o objetivo de estimar o tempo médio de um caixa eletrônica 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 95% de confiança? Admita que, em estudos anteriores, verificou-se que o desvio padrão não ultrapassa 8 segundos.

16) Deseja-se estudar as percentagens de ocorrências de diversos atributos das famílias de uma comunidade de 600 famílias. Qual deve ser o tamanho de uma amostra aleatória simples, considerando em cada estimativa um erro máximo de 4% e nível de 95% de confiança?

Exercícios complementares

17) Nas situações descritas abaixo, descreva qual é a população, a amostra, o parâmetro de interesse e a estatística que poderia ser usada para estimar o parâmetro de interesse.

a) Para avaliar a proporção de alunos do Curso de Administração favoráveis a 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.

18) Um instituto de pesquisa observou uma amostra aleatória de 800 habitantes de uma grande cidade. Verificou que 320 indivíduos desta amostra apóiam a administração da prefeitura, enquanto que os outros 480 a criticam.

a) O que se pode dizer sobre a percentagem de indivíduos que apóiam a administração da prefeitura, dentre a amostra observada?

b) O que se pode dizer sobre a percentagem de indivíduos que apóiam a administração da prefeitura, dentre os habitantes da cidade?

Obs.: Em caso de estimativa, usar nível de confiança de 95%.

19) 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 deste produto.

a) Qual deve ser o tamanho de uma amostra aleatória simples, que garanta uma estimativa com erro máximo de 5% , ao nível de confiança de 99%?

b) Efetuou-se a amostragem, conforme o tamanho calculado no item (a), e verificou-se que nesta amostra 200 pessoas passariam a usar regularmente o produto. Construa um intervalo 99% de confiança para o parâmetro de interesse. Interprete o intervalo de confiança.

20) 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.

21) 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 não seja superior a 0,2 unidades, com 95% de confiança?

22) 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.

$$-5 \quad -10 \quad 5 \quad -20 \quad -8 \quad 10 \quad 0 \quad -2 \quad -8 \quad -1$$

a) Calcule a média, mediana e desvio padrão da amostra.

b) Construa um intervalo de 95% de confiança para o parâmetro μ (μ = *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 emagrecer os indivíduos?

23) Uma empresa tem 2.400 empregados. Deseja-se extrair uma amostra entre os empregados para verificar o grau de satisfação em relação a qualidade da comida no refeitório. Em uma amostra piloto, numa escala de 0 a 10, o grau de satisfação recebeu nota média 6,5 e desvio padrão de 2,8.

a) Determine o tamanho mínimo da amostra, admitindo um planejamento por amostragem aleatória simples, com erro máximo de 0,5 unidades e nível de 99% de confiança.

b) Considerando que a amostra planejada no item anterior tenha sido executada, donde obteve-se média de 5,3 e desvio padrão de 2,6 pontos. Faça um intervalo de 99% de confiança para o parâmetro μ.

c) Considerando o resultado do item anterior, você diria com um nível mínimo de 99% de confiança, que se a pesquisa fosse aplicada nos 2.400 funcionários, a nota média seria superior a cinco? Justifique.

d) Se na amostra planejada no item (a), 120 atribuíssem 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.

24) Uma pesquisa realizada por pesquisadores da Universidade Federal de Minas Gerais, que 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 européia, 28% é negra e 33% é indígena.[9] Admitindo que a amostragem tenha sido aleatória, qual a margem de erro de cada uma destas estimativas, considerando nível de confiança de 95%?

---

[9] Divulgado no Jornal Hoje – Rede Globo, em 18/04/00.

Capítulo 10

Testes estatísticos de hipóteses

Muitas vezes o pesquisador tem alguma idéia, ou conjetura, sobre o comportamento de uma variável, ou de uma possível associação entre variáveis. Nestes casos, o planejamento da pesquisa deve ser de tal forma que permita, com os dados amostrais, testar a veracidade de suas idéias sobre a população em estudo. Adotamos que a população seja o mundo real e as idéias sejam as hipóteses de pesquisa, que poderão ser testadas por técnicas estatísticas denominadas de *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 a 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. Tomemos, como ilustração, o Exemplo 10.1 (a), em que se busca uma relação entre sexo e tabagismo. Inicialmente 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*, para que esta possa ser observada apropriadamente. Uma maneira razoavelmente simples de mensurar *taba-*

gismo 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.

	Sexo		
Tabagismo	masculino	feminino	Total
fumante	92 (46%)	38 (38%)	130 (43%)
não-fumante	108 (54%)	62 (62%)	170 (57%)
Total	200 (100%)	100 (100%)	300 (100%)

Como na amostra observada, 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 que estamos examinando uma amostra e, conseqüentemente, 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 nestes 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. Esta hipótese é descrita em termos de parâmetros populacionais e é, basicamente, uma negação daquilo que o pesquisador deseja provar. Sob esta 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 e execução 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 desigualdades (maior, menor ou diferente).

No Exemplo 10.1a, 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 Exemplo 10.1b, 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 desta moeda, podemos escrever:

$$H_0: \pi = 0{,}5 \qquad \text{e} \qquad H_1: \pi \neq 0{,}5$$

10.2 CONCEITOS BÁSICOS

Apresentaremos as primeiras idéias sobre testes estatísticos, ou testes de significância, usando como ilustração um experimento binomial. Considere o seguinte problema:

Suspeita-se que uma certa moeda, usada num jogo de azar, é viciada.

Então, se chamarmos π à probabilidade de cara desta moeda, podemos formular as hipóteses da seguinte maneira:

H_0: $\pi = 0{,}5$ (a moeda é honesta) e H_1: $\pi \neq 0{,}5$ (a moeda é viciada)

Suponhamos, inicialmente, H_0 como verdadeira. Ela somente vai ser rejeitada em favor de H_1, se houver evidência suficiente que a contradiga. A existência desta possível evidência será verificada a partir de um conjunto de observações relativas ao problema em estudo. No presente exemplo, o conjunto de observações (amostra) consistirá dos resultados de uma série de lançamentos imparciais da moeda.

Em cada lançamento da moeda, observamos um resultado: *cara* ou *coroa*. Ao observar uma amostra de n lançamentos, podemos computar o valor da estatística:

Y = número total de caras nos n lançamentos

A estatística Y poderá ser usada na definição de um critério de decisão: *aceitar* H_0 ou *rejeitar* H_0 *em favor de* H_1. Neste contexto, a estatística Y é chamada de ***estatística do teste***.

Vamos considerar uma amostra de $n = 10$ lançamentos e as duas seguintes situações.

SITUAÇÃO A – Suponha que nos 10 lançamentos, observamos $Y = 10$ caras. Podemos rejeitar H_0, em favor de H_1?

SITUAÇÃO B – E se tivéssemos observado $Y = 7$ caras?

É intuitivo, que na situação A, existe mais evidência para rejeitar H_0. Contudo, em nenhuma das duas situações, podemos rejeitar H_0 com a certeza de que esta hipótese é realmente falsa, pois, estamos trabalhando com um fenômeno aleatório, onde é plenamente possível, em 10 lançamentos de uma moeda sabidamente honesta (H_0 verdadeira), ocorrerem 7, 8, 9, ou, até mesmo 10 caras! Por outro lado, se a ocorrência de um certo resultado for muito pouco provável para uma moeda honesta, torna-se natural decidirmos por H_1 (moeda viciada).

No presente contexto, torna-se necessário conhecer a probabilidade de ocorrerem $Y = 10$ caras (situação A), ou $Y = 7$ caras (situação B), em 10 lançamentos de uma moeda honesta. Mais geralmente, precisamos da distribuição de probabilidades da estatística do teste Y, admitindo H_0 verdadeira. Esta distribuição de probabilidades será a referência básica para analisarmos o resultado observado na amostra e decidirmos entre H_0 e H_1.

A distribuição de probabilidades de Y (distribuição de referência)

Como o exemplo em questão é um experimento binomial, então, como vimos no Capítulo 7, Y tem distribuição binomial, com parâmetros $n = 10$ e $\pi = 0{,}5$ (supondo H_0 verdadeira). A Figura 10.1 apresenta esta distribuição sob forma gráfica. As probabilidades, p(y), foram obtidas na tabela da distribuição binomial (Tabela II do apêndice). Para facilitar a exposição estas probabilidades foram arredondadas para três decimais.

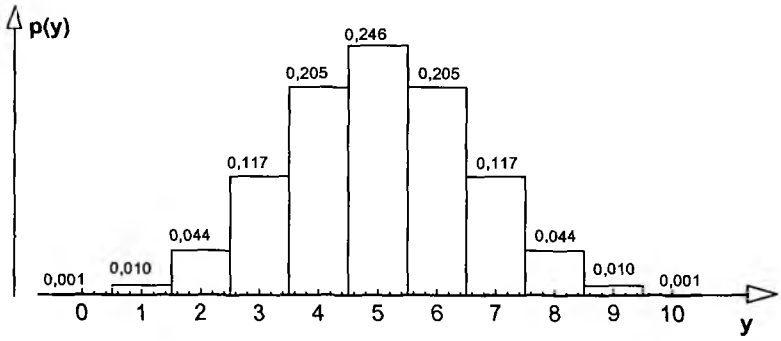


Figura 10.1 Distribuição da estatística Y = *número de caras em 10 lançamentos da moeda*, sob H_0 (binomial com $n = 10$ e $\pi = 0{,}5$).

Com a distribuição de probabilidades da estatística do teste, podemos avaliar melhor a adequação de H_0 com o resultado de Y, observado 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.

Vamos, agora, familiarizar-nos com o conceito de *probabilidade de significância*, que é um valor obtido em função da distribuição de probabilidades da estatística do teste e do resultado observado na amostra. Este valor será o elemento fundamental para a tomada de decisão entre H_0 e H_1.

Probabilidade de significância

Supondo, inicialmente, H_0 como a hipótese verdadeira, a ***probabilidade de significância***, ou ***valor p***, é definida como a probabilidade de a estatística do teste acusar um resultado tanto ou mais distante do esperado como o resultado ocorrido na particular amostra observada. Veja os seguintes exemplos.

Exemplo 10.3 Retomemos a situação A, onde 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, verifica-se que tanto ou mais distante do que o valor observado na amostra ($Y = 10$), encontram-se o valor 0 e o próprio valor 10, como ilustra a Figura 10.2.

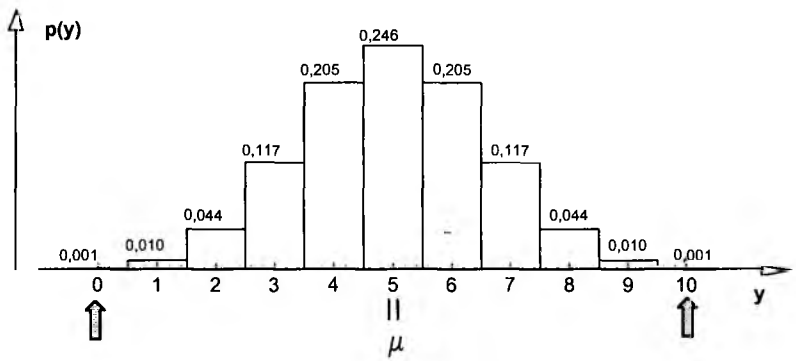


Figura 10.2 Distribuição de Y, sob H_0. As setas indicam os valores que distam do esperado, $\mu = 5$, tanto ou mais do que o valor $Y = 10$, observado na amostra da situação A.

Conseqüentemente, 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 tanto 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, torna-se natural rejeitar a hipótese de que a moeda é honesta (H_0), decidindo-se pela hipótese de que a moeda é viciada (H_1).

Os dados observados mostram evidência suficiente para dizer que a moeda é viciada!

Exemplo 10.4 Vejamos, agora, a situação B, onde observamos $Y = 7$ caras em $n = 10$ lançamentos. Nesta situação, tanto ou mais distante do que o valor $Y = 7$, encontram-se os valores: 7, 8, 9, 10, 0, 1, 2 e 3, como ilustra a Figura 10.3.

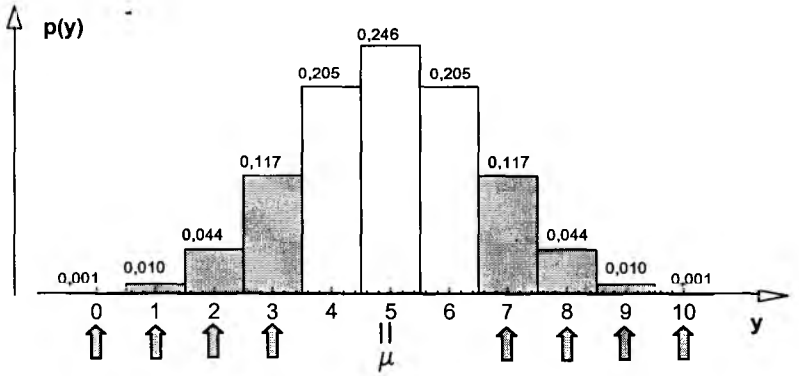


Figura 10.3 Distribuição de Y, sob H_0. As setas indicam os valores que distam do esperado, $\mu = 5$, tanto ou mais do que o valor $Y = 7$, observado na amostra da situação B.

Temos, então, a seguinte probabilidade de significância:

$$p = p(0) + p(1) + p(2) + p(3) + p(7) + p(8) + p(9) + p(10) =$$
$$= 0{,}001 + 0{,}010 + 0{,}044 + 0{,}117 + 0{,}117 + 0{,}044 + 0{,}010 + 0{,}001 =$$
$$= 0{,}344 \text{ (ou, } 34{,}4\%).$$

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, torna-se mais prudente não rejeitar H_0.

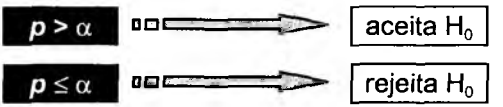
Não há evidência suficiente para afirmar que a moeda é viciada!

O valor p aponta o quão *estranho* foi o resultado observado na amostra à luz de H_0. Logo, 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, 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 estar fazendo uma afirmação errada.

Nível de significância

Na realização de uma pesquisa, quando se deseja confirmar ou refutar alguma hipótese, é comum estabelecer, ainda na fase do planejamento da pesquisa, 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:



Exemplo 10.3 (continuação) Na amostra da situação A, quando observamos 10 caras em 10 lançamentos, se estivermos usando o nível de significância de 5% ($\alpha = 0{,}05$), o teste estatístico *rejeita* H_0, *em favor de* H_1 (pois, a probabilidade de significância, observada na amostra, foi de 0,002 e, portanto, *menor* do que o valor adotado para α).

Exemplo 10.4 (continuação) Usando $\alpha = 0{,}05$ na amostra da situação B, quando observamos 7 caras em 10 lançamentos, o teste estatístico *não* rejeita H_0 (pois, a probabilidade de significância, observada na amostra, foi de 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, tem-se uma 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, apenas, que os dados não mostraram evidência suficiente para rejeitá-la e, por isto, continuamos acreditando em sua veracidade.

> *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.* (R. A. Fisher, 1956, p.16).

Estabelecido um nível de significância α antes da observação dos dados, temos as possibilidades apresentadas no esquema a seguir:

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 este segundo erro como β, mas, ao contrário de α, a probabilidade β não é fixada *a priori*. Em razão disto, 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 II 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 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 exemplo 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. Esta é uma situação típica de *teste bilateral.*

Existem situações em que se pretende rejeitar H_0 somente num dos sentidos. Por exemplo, suspeita-se que a moeda tende a dar mais caras do que coroas. Neste caso, 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 estas 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 um *teste unilateral.* Assim, 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. A Figura 10.4 ilustra a probabilidade de significância deste teste.

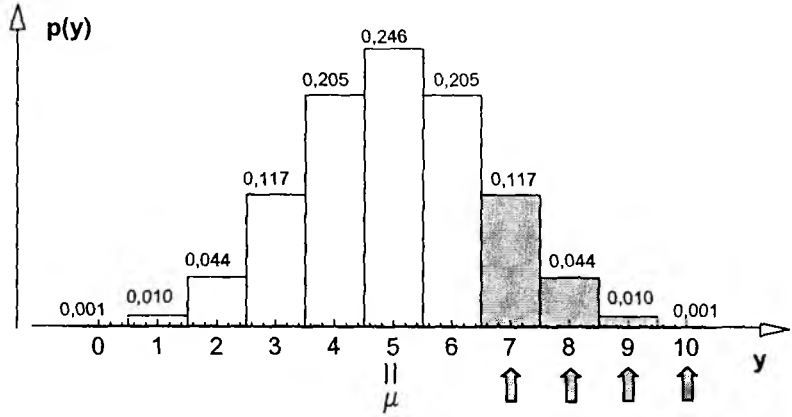


Figura 10.4 Ilustração do cálculo da probabilidade de significância do teste unilateral do-Exemplo 10.5.

Exemplo 10.6 (Teixeira, Meinert e Barbetta, 1987, p.137) Com o objetivo de testar se a diferença de odor em sorvetes de morango é percebida por degustadores, efetuou-se um experimento, como descrito a seguir.

Para cada um dos 8 (oito) degustadores selecionados para o experimento foram dadas, em ordem aleatória e sem identificação, duas amostras de sorvete: uma com odor mais forte e outra normal. As amostras de sorvete foram elaboradas de forma tão similar quanto possível, com exceção da intensidade de odor, que é a característica em estudo.

Chamando de 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 do 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 II 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. Donde 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 a abordagem (unilateral ou bilateral) que 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 II do apêndice), obtenha a probabilidade e 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 resultados 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 para o 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]

$$\mu = n\pi \qquad \text{e} \quad \sigma = \sqrt{n\pi(1-\pi)}$$

---

[1] A aproximação da distribuição normal à binomial foi vista no Capítulo 8. Uma forma muitas vezes usada para verificar a validade da aproximação normal é calculando: (a) $n.\pi$ e (b) $n.(1-\pi)$, alocando para π o valor declarado em H_0. Se as expressões (a) e (b) acusarem valores iguais ou superiores a 5 (cinco), a distribuição normal pode ser usada no lugar da binomial.

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, acusando $Y = 28$ caras. Podemos rejeitar 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.

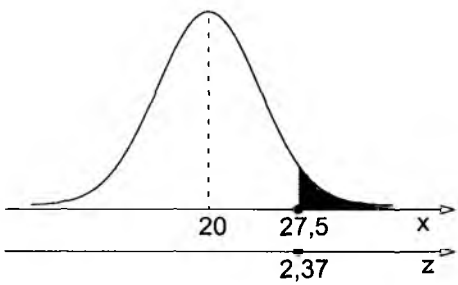


Figura 10.5 Ilustração da obtenção de uma probabilidade de significância, usando um modelo normal.

Para realizar o cálculo da área indicada na Figura 10.5, precisamos calcular os parâmetros do modelo normal:

$$\mu = (40)(0,5) = 20 \text{ e } \quad \sigma = \sqrt{(40)(0,5)(0,5)} = 3,16$$

O valor 27,5 da escala original (escala x) corresponde ao seguinte valor padronizado (escala z):

$$z = \frac{x - \mu}{\sigma} = \frac{27,5 - 20}{3,16} = 2,37$$

Usando a tabela da distribuição normal padrão (Tabela IV do apêndice), encontramos para $z = 2,37$ uma área de 0,0089 na cauda superior da curva. Temos, então, $p = 0,0089$. Como p 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 nenhuma 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 dados que se tem ou que se planeja coletar. Com respeito aos tipos de dados, existem testes voltados para dados quantitativos, onde normalmente as hipóteses são apresentadas em termos de *médias* e testes voltados para dados qualitativos, onde as hipóteses são apresentadas em termos de proporções ou probabilidades de eventos. Os exemplos deste capítulo, usando a distribuição binomial para encontrar o valor p, estão na segunda categoria.

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 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 da moeda foram imparciais e realizados sob as mesmas condições).

A decisão do teste estatístico será sempre a comparação do valor p com o nível de significância α preestabelecido (ver a Seção 10.2), mas a implicação do resultado estatístico depende da aplicação em questão. Por exemplo, num estudo experimental, normalmente a decisão do teste estatístico implica uma relação de causa e efeito, mas num estudo de levantamento, o resultado do teste usualmente leva apenas a uma conclusão de diferença entre grupos.

Hoje em dia, o cálculo da estatística do teste e a obtenção do valor p tornaram uma tarefa relativamente fácil 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. Nessa parte, o ser humano ainda está muito na frente da máquina!

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 os seguintes itens:

a) Num dado teste, a probabilidade de significância foi de $p = 0{,}0001$. Com base no resultado da amostra, qual a conclusão (decide-se pela hipótese nula ou pela hipótese alternativa)? Com base no resultado da amostra, qual o risco de o pesquisador estar tomando a decisão errada?

b) Em outro teste, o nível de significância descritivo foi de $p = 0{,}25$. Qual a conclusão? Qual o risco de o pesquisador estar tomando a decisão errada?

c) Em outros dois testes, as probabilidades de significância foram de 0,0001 e 0,01, respectivamente. Em qual dos testes o pesquisador deve estar mais convicto da decisão de qual hipótese deve ser aceita? Por quê?

12) Com o objetivo de se verificar 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 obteve-se 2 caras (e 6 coroas), qual a conclusão ao nível de significância de 5%?

b) Se em 80 lançamentos obteve-se 20 caras (e 60 coroas), qual a conclusão ao nível de significância de 5%?

13) Para testar se uma criança tem algum conhecimento sobre determinado assunto, elaboraram-se 12 questões do tipo *certo-errado*. A criança acertou 11. Qual é a conclusão ao nível de significância de 5%?

14) Para testar se uma criança tem algum conhecimento sobre determinado assunto, elaboraram-se 12 questões, cada uma com 4 possibilidades de escolha. A criança acertou 5.

a) Formule as hipóteses em termos do parâmetro π = probabilidade de acerto de cada questão.

b) Qual o número esperado de acertos sob H_0.

c) Qual o valor p.

d) Qual a conclusão ao nível de significância de 5%?

15) Para testar se um sistema computacional "inteligente" adquiriu algum conhecimento sobre determinado assunto, elaborou-se 60 questões do tipo *certo-errado*. O sistema acertou 40. Qual é a conclusão 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, com respeito a alguma variável de interesse, ou *variável resposta*. Estes grupos podem diferir quanto a diferentes tratamentos aplicados a seus elementos, ou devido a diferentes populações de onde estes elementos são extraídos. Os Exemplos 11.1 e 11.2 apresentam estas duas situações.

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 em estudo. A comparação entre os dois grupos é realizada a partir de uma avaliação que mensure os conhecimentos de matemática de cada criança (veja a Figura 11.1).

---

[1] A divisão aleatória pode ser feita por sorteio, ou usando uma tabela de números aleatórios. Veja o Exercício 5, Capítulo 3.

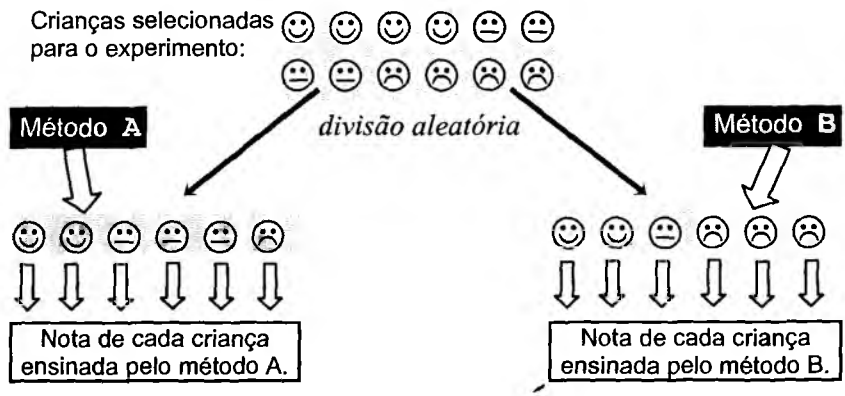


Figura 11.1 Esquema do planejamento de um experimento para comparar dois métodos de ensinar matemática para crianças.

Exemplo 11.2 Para comparar o peso ao nascer de crianças, em duas localidades, podemos extrair uma amostra aleatória de nascimentos em cada localidade, observando os pesos das crianças nas duas amostras (veja a Figura 11.2).

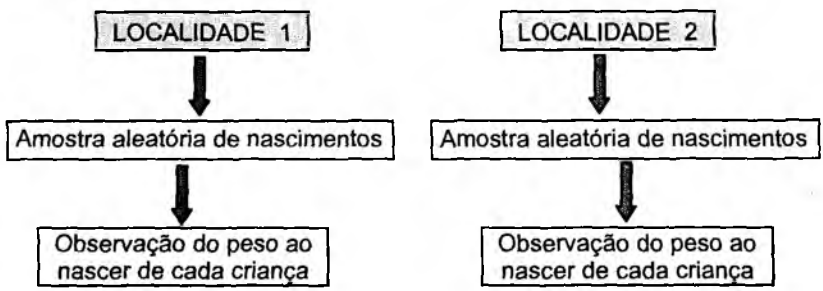


Figura 11.2 Esquema de um planejamento amostral, num estudo tipo levantamento, para comparar o peso ao nascer de crianças, em duas localidades.

O uso de testes estatísticos permite 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, ou *significativas*, podem ser causadas, por exemplo, 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, fez-se 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).

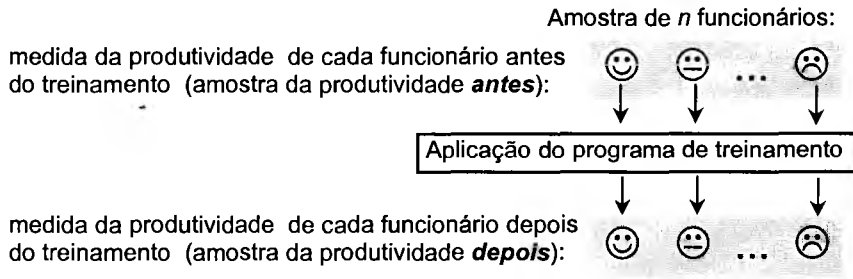


Figura 11.3 Esquema de um estudo, tipo *antes-e-depois*, para avaliar o efeito de um programa de treinamento na produtividade de funcionários de uma empresa.

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.

Ao realizar o planejamento de uma pesquisa, torna-se fundamental planejar, também, o procedimento estatístico que vai ser usado na análise dos dados. Particularmente, em pesquisas confirmatórias, isto é, naquelas pesquisas em que se tem, *a priori*, hipóteses que se deseja colocar à prova, deve-se realizar o planejamento preocupando-se em verificar, por exemplo, se a execução deste planejamento vai gerar dados pareados ou amostras

independentes, dados quantitativos ou categorizados, e assim por diante. Para cada situação, podemos pensar num teste estatístico adequado.

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 execução 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, torna-se mais recomendado um delineamento como proposto no Exemplo 11.1 (amostras independentes).[2]

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.

A aplicação do teste dos sinais é adequada em:

- delineamentos de pesquisa que produzam dados pareados e
- a variável em estudo é observada de forma qualitativa e com apenas duas categorias, tal como: *melhorou* ou *piorou*.[3]

---

[2] Existem muitos outros delineamentos de pesquisa que poderiam ser usados no presente problema. O pesquisador deve verificar cuidadosamente o delineamento mais apropriado para o seu particular problema de pesquisa. Uma boa discussão sobre delineamentos de pesquisa pode ser lida em Selltiz, Wrightsman e Cook, vol. I (1987).

[3] O teste dos sinais também poderia ser usado nas situações em que a variável em observação é mensurada quantitativamente. Contudo, neste caso, existem testes estatísticos mais apropriados, como veremos na Seção 11.3.

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 admitir que ao observar as produtividades de um funcionário, antes e depois da realização do programa de treinamento, a única avaliação possível é: *melhorou* ou *piorou*. Neste contexto, as hipóteses podem ser colocadas em termos do parâmetro π da distribuição binomial, como segue.

$$H_0: \pi = 0{,}5 \quad \text{e} \quad H_1: \pi > 0{,}5$$

onde π = probabilidade do funcionário aumentar a produtividade após o treinamento.

O teste é realizado a partir de uma amostra de n funcionários. Para cada funcionário é observada a sua produtividade *antes* e *depois* da aplicação do programa de treinamento, verificando-se se *melhorou* (+) ou se *piorou* (-). A estatística a ser usada no teste será: Y = *número de funcionários que aumentaram de produtividade.*

Admitindo que:

- todos os funcionários são 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, é o programa de treinamento.

Então, a estatística Y tem distribuição binomial com parâmetros n e π. Desta forma, a probabilidade de significância pode ser computada a partir da distribuição binomial (ou pela distribuição normal, quando n for grande), tal como vimos no capítulo anterior.

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 o total de sinal positivo na amostra: $Y = 7$. A probabilidade de significância para o resultado observado na amostra pode ser obtido pela tabela da distribuição binomial (Tabela II do apêndice), com $n = 10$ e $\pi = 0,5$. Como o teste é unilateral, 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% ($\alpha = 0,05$), que é usual nesses tipos de problemas, o teste dos sinais *não* pode rejeitar H_0 em favor de H_1 (pois, $p > \alpha$). Concluímos, então, que os dados observados no presente estudo *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*. Nesses casos, é comum desprezar estes indivíduos da amostra (veja o Exercício 1d). Contudo, se houver um número grande de indivíduos nesta 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, selecionou-se uma amostra de uma grande platéia, indagando a cada um, sua opinião sobre o candidato, antes e depois da apresentação: se melhorou ou se piorou.

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 delinqüência 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 deve aplicar, nos casos de delinqüência 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 dois conjuntos de dados são pareados, oriundos, por exemplo, de um procedimento tipo *antes-e-depois*.

Exemplo 11.4 Tomemos, novamente, o problema do Exemplo 11.3, mas, agora, vamos admitir 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

$$H_0: \mu_{depois} = \mu_{antes} \quad \text{e} \quad H_1: \mu_{depois} > \mu_{antes},$$

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. 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, relativa a cada funcionário. Estes incrementos (ou reduções) de produtividade estão também apresentados na Figura 11.4, sob forma de um diagrama de pontos.

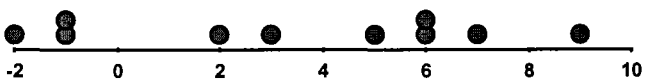


Figura 11.4 Diagrama de pontos das diferenças de produtividade.

Observamos no diagrama de pontos que, na amostra observada, houve uma tendência de ocorrer diferenças positivas (valores de *produtividade depois*, em geral, maiores 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.

A estatística do teste

A estatística do teste baseia-se nos valores observados da variável D, definida por

$$D = (\textit{medida depois}) - (\textit{medida antes})$$

Se a hipótese nula for correta, devemos esperar que os valores observados 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 da amostra, que, neste caso, corresponde ao número de pares (*antes, depois*) observados;
- $\overline{D}$: média das diferenças observadas; e
- S_D: desvio padrão das diferenças observadas.[4]

Exemplo 11.4 (continuação) Diferenças D (última coluna da Tabela 11.2):

3, 7, -2, 6, -1, 6, 2, 9, -1, 5

Donde:

$$n = 10 \qquad \overline{D} = \frac{\sum D}{n} = \frac{34}{10} = 3{,}4$$

$$S_D = \sqrt{\frac{\sum D^2 - n.\overline{D}^2}{n-1}} = \sqrt{\frac{246 - (10)(3{,}4)^2}{10-1}} = 3{,}81$$

[4] O cálculo da média e do desvio padrão foi visto no Capítulo 6.

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 existirá sobre o fenômeno em estudo e, conseqüentemente, 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 do efeito 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.

A 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 a amostra for razoavelmente grande ($n \geq 30$, por exemplo), o teste ainda permanece válido, mesmo que a variável D não tenha uma distribuição normal.

Na prática, recomendamos fazer um histograma de freqüências ou um diagrama de pontos dos valores observados da variável D, para verificar se não existe algum ponto discrepante ou uma forte assimetria, o que poderia comprometer a realização deste teste estatístico. No exemplo em discussão, foi construído um diagrama de pontos (Figura 11.4), em que 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 11.5).

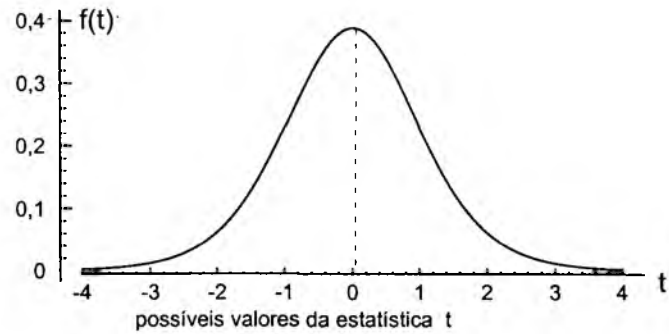


Figura 11.5 Distribuição de referência para o teste *t* do Exemplo 11.4: *Distribuição t de Student com gl = 9 graus de liberdade.*

A Figura 11.5 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 amostrais e calcular o valor da estatística *t*, podemos obter a probabilidade de significância ou valor *p*, a partir de uma tabela da *distribuição t de Student*, conforme é mostrado na continuação do Exemplo 11.4.[5]

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 $gl = 9$ graus de liberdade (pois $gl = n - 1$). Tomemos, então, a linha de $gl = 9$ da Tabela V do apêndice (tabela da *distribuição t de Student*), como mostra a Figura 11.6. Por esta tabela, obtemos a área relativa a um valor maior ou igual a $t = 2{,}82$. Esta área corresponde à probabilidade de significância *p* descrita pelos dados da amostra.

---

[5] Hoje temos no mercado diversos *softwares* computacionais de estatística (*SPSS, SAS, S-PLUS, STATISTICA,* etc.) que calculam o valor da estatística *t* e fornecem o correspondente valor da probabilidade de significância, tornando desnecessário o uso de tabelas da *distribuição t de Student*. Algumas planilhas eletrônicas, como o *Microsoft Excel*, por exemplo, também são supridas pelo teste *t* – veja aplicação na seção seguinte.

dados observados

$t = 2{,}82$

	Área na cauda superior						
gl	0,25	0,10	0,05	0,025	0,010	0,005	...
...							
9	0,703	1,383	1,833	2,262	2,821	3,250	...
...							

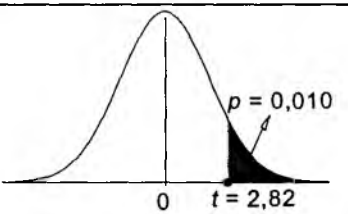


Figura 11.6 Uso da *distribuição t de Student* com $gl = 9$ para a obtenção da probabilidade de significância num teste unilateral, com $n = 10$ e valor observado de t igual a 2,82.

Observando a linha correspondente a $gl = 9$, verificamos, na tabela, que o valor $t = 2{,}82$ (calculado a partir da amostra) está próximo do valor tabulado 2,821. Logo, como ilustra a Figura 11.6, 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 do teste dos sinais e de o teste 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 as informações contidas 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 dos dados se apresentarem de forma parecida com a forma da 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: Em média, a produtividade *aumenta* com o programa de treinamento). Quando o teste é bilateral, isto é, a hipótese alternativa leva o sinal "≠", o procedimento é análogo, mas, no final, o valor da área deverá ser *dobrada*, para que o valor *p* corresponda às áreas das duas caudas da distribuição.

Exemplo 11.5 Desejamos verificar se uma certa alteração no horário do 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:

$$H_0: \mu_{depois} = \mu_{antes} \quad \text{e} \quad H_1: \mu_{depois} \neq \mu_{antes}$$

onde

μ_{antes}: produtividade média dos funcionários da empresa, considerando o horário habitual; e

μ_{depois}: produtividade média dos funcionários da empresa quando há alteração no horário do turno de trabalho.

Por simplicidade, admita que os resultados foram os mesmos do Exemplo 11.4, apresentados na Tabela 11.2, acarretando, como já vimos, um valor de t igual a 2,82, com $gl = 9$. A obtenção da probabilidade de significância é análoga ao caso anterior, considerando, porém, ambos os lados da curva, ou seja, a probabilidade de significância p será o dobro daquele valor observado na Figura 11.6. Portanto: $p = 2.(0{,}010) = 0{,}020$. Ao nível de significância de 5%, o teste rejeitaria H_0, em favor de H_1.

Outras formas de pareamento

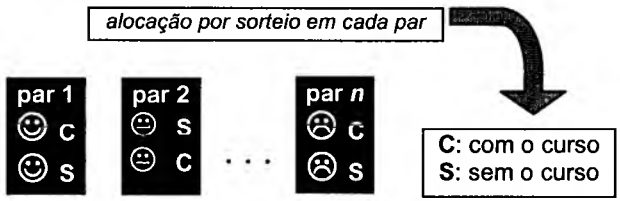
O plano de pesquisa de observar a variável resposta sobre os mesmos elementos, antes e depois de aplicar um certo tratamento, pareceu adequado no problema de avaliar o efeito de um programa de treinamento sobre a produtividade de funcionários (Exemplos 11.3 e 11.4). Contudo, se o programa de treinamento for relativamente longo, de tal forma que, nesse período, outros fatores possam 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 estes fatores intervenientes.

Um planejamento mais adequado para a situação colocada consiste em observar dois grupos de funcionários, sendo que apenas um desses grupos recebe o programa de treinamento. Após a realização deste programa, comparam-se os valores de produtividade entre os dois grupos.[6]

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 (diferença entre os dois elementos do par). O exemplo seguinte apresenta um planejamento de pesquisa com este enfoque.

Exemplo 11.6 Para avaliar o efeito de um curso sobre alimentação e controle de peso, em pessoas obesas, planeja-se tomar pares similares destas pessoas. 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, todas as pessoas participantes do estudo deverão fazer duas visitas ao médico, num prazo de três meses, para medir a variação dos pesos. 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. Veja o Exercício 3.

---

[6] 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.

Exercícios

3) Seja o problema do Exemplo 11.6.

a) Apresente as hipóteses nula e alternativa.

b) Considerando que a execução desta pesquisa produziu os dados constantes na tabela seguinte, qual a conclusão?

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 tal forma que as lojas de um mesmo par são o mais similares possível, em termos, por exemplo, do volume de vendas, localidade, identidade de preços, etc. Em cada par de lojas, uma passou a oferecer o brinde e a outra, não.

a) Apresente as hipóteses nula e alternativa.

b) Os resultados das vendas, em quantidade de unidades vendidas, foram os seguintes:

Par de loja	Vendas sem brinde	Vendas com brinde
1	33	43
2	43	39
3	26	33
4	19	32
5	37	43
6	27	46

Os dados mostram evidência suficiente para se afirmar que a oferta do brinde aumenta as vendas? Use nível de significância de 5%.

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 deste produto nas 12 lojas. Os incrementos (ou reduções) nas vendas foram os seguintes:

7 10 5 -2 9 0 3 -4 8 9 1 3

a) Os dados mostram evidência suficiente para se afirmar que a oferta do brinde aumenta as vendas? Use nível de significância de 5%.

b) Aponte as vantagens e desvantagens deste planejamento de pesquisa, em relação ao apresentado no Exercício 4, considerando o particular problema em discussão.

c) Apresente um terceiro planejamento de pesquisa para este problema, tentando aproveitar as vantagens dos dois procedimentos apresentados.

6) (Mendenhall, 1985, p.359.) Para comparar o uso de duas entradas de uma lanchonete, o gerente anotou o número de pessoas que entravam por uma e por outra entrada, durante sete dias consecutivos. Os dados resultantes estão na tabela a seguir. Esses dados têm evidência suficiente capaz de garantir uma demanda média maior com relação a uma das entradas? Use $\alpha = 0{,}01$.

Dia	Seg	Ter	Qua	Qui	Sex	Sab	Dom
Entrada A	420	374	434	395	637	594	679
Entrada B	391	343	469	412	538	521	625

7) Considerando os dados do anexo do Capítulo 2, podemos afirmar que existe diferença significativa entre: (a) *satisfação dos alunos, com respeito 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 Considere o problema discutido no primeiro exemplo deste capítulo, 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 a realização do teste, precisamos de uma amostra de crianças submetidas ao método A de ensino e outra amostra de crianças submetidas ao método B. 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:

$$H_0: \mu_1 = \mu_2 \quad \text{e} \quad H_1: \mu_1 \neq \mu_2,$$

onde

μ_1: nota média das crianças na avaliação, se elas forem submetidas ao método A de ensino; e

μ_2: nota média das crianças na avaliação, se elas forem submetidas ao método B de ensino.

Neste exemplo, vamos construir os dois grupos, dividindo as crianças aleatoriamente entre eles, como já foi ilustrado na Figura 11.1. Este procedimento deve gerar duas amostras *independentes*, pois, as crianças de um grupo não têm qualquer ligação com as crianças do outro grupo.

> *A aleatorização dos grupos é fundamental para resguardar a validade de um teste de significância* (R. A. Fisher, 1956, p.19).

Entende-se por *aleatorização* não somente a divisão aleatória dos elementos nos grupos, mas também, as condições idênticas em que estes grupos devem ser tratados, a não ser, é claro, pelos diferentes tratamentos em estudo. No exemplo em questão, devemos evitar qualquer interação entre as crianças dos dois grupos, qualquer variação devida aos instrutores, etc.

A Tabela 11.3 mostra os resultados do experimento descrito no Exemplo 11.7, considerando que ambos os grupos foram compostos por dez crianças. E a Figura 11.7 apresenta o diagrama de pontos dos resultados de cada amostra.

Tabela 11.3 Notas em conhecimentos de matemática, 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

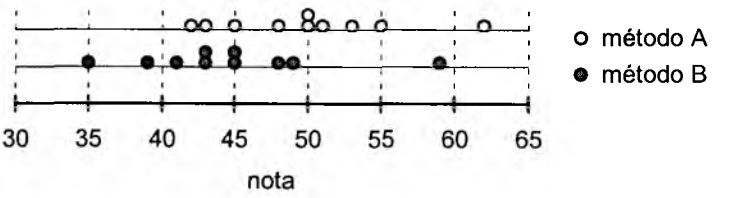


Figura 11.7 Diagrama de pontos das notas obtidas pelas crianças, segundo o método de ensino.

A 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 destas amostras. Quanto maior as amostras, maior a evidência de uma possível evidência de uma diferença real (pense no caso extremo de apenas uma criança em cada grupo, apontando uma diferença de 2 unidades numa escala de 0 a 10 – *não dá para dizer muita coisa!* – mas com 100 crianças em cada grupo, apontando uma diferença de 2 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 11.8.

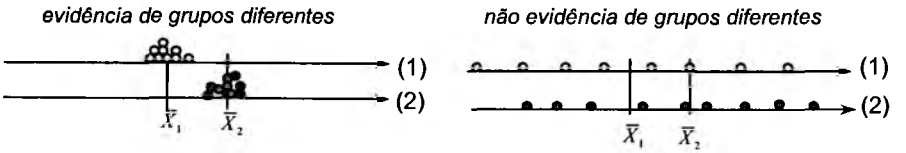


Figura 11.8 A importância de se considerar a variância interna dos grupos.

Considerando o mesmo número de elementos, n, 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:[7]

$$S_a^2 = \frac{S_1^2 + S_2^2}{2}$$

[7] Lembramos ao leitor que a variância (S^2) é o desvio padrão (S) ao quadrado.

E a estatística do teste é dada por

$$t = (\overline{X}_1 - \overline{X}_2) \cdot \sqrt{\frac{n}{2S_a^2}}$$

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) Os cálculos das médias e dos desvios padrão são feitos como foi visto no Capítulo 6.

Amostra 1: $n = 10$, $\overline{X}_1 = 49{,}90$ e $S_1 = 5{,}97$

Amostra 2: $n = 10$, $\overline{X}_2 = 44{,}70$ e $S_2 = 6{,}50$

Variância agregada: $S_a^2 = \dfrac{S_1^2 + S_2^2}{2} = \dfrac{(5{,}97)^2 + (6{,}50)^2}{2} = \dfrac{77{,}89}{2} = 38{,}95$

Estatística do teste:

$$t = (\overline{X}_1 - \overline{X}_2) \cdot \sqrt{\frac{n}{2S_a^2}} = (49{,}90 - 44{,}70) \cdot \sqrt{\frac{10}{2(38{,}95)}} = (5{,}2) \cdot \sqrt{0{,}1284} = (5{,}2) \cdot (0{,}3583)$$

ou seja, $t = 1{,}86$

Para se ter uma distribuição de referência para a estatística t e, assim, proceder o teste estatístico, torna-se necessário que os dados observados satisfaçam as seguintes suposições.

Suposições básicas para a aplicação do teste: (1) os dois conjuntos de dados provêm de distribuições normais e (2) com a mesma variância (mesmo desvio padrão).[8]

---

[8] 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.

Na prática, não é fácil verificar a veracidade destas suposições. Aconselhamos, contudo, construir histogramas de freqüências ou diagramas de pontos para cada amostra. Estes gráficos permitem avaliar se existem fortes violações destas suposições, tais como a presença de pontos discrepantes, distribuições com formas assimétricas ou, ainda, uma distribuição bem mais dispersa do que a outra. No exemplo em discussão, construímos diagramas de pontos para as duas amostras (Figura 11.7), os quais mostram que as amostras em análise parecem compatíveis com as suposições do teste.

Distribuição de referência. Se as médias populacionais forem iguais (H_0 verdadeira) e as suposições básicas puderem ser admitidas, então, a estatística t tem *distribuição t de Student com* $gl = 2n - 2$ *graus de liberdade.*

A continuação do Exemplo 11.7 mostra a obtenção da probabilidade de significância p, usando a distribuição de referência para o valor calculado $t = 1{,}86$ e $gl = 2n - 2 = 2(10) - 2 = 18$.

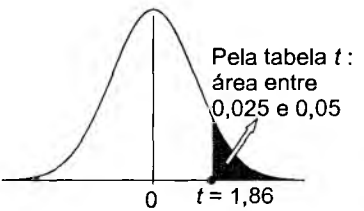
Exemplo 11.7 (continuação) O esquema seguinte ilustra o uso da Tabela V do apêndice (tabela da *distribuição t de Student*) para se obter a probabilidade de significância do valor calculado de t.

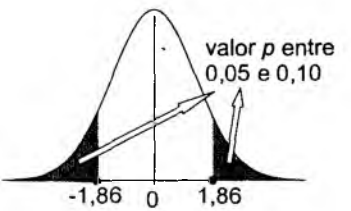
dados observados

$t = 1{,}86$

	Area 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 observados levaram ao valor $t = 1{,}86$, apontando para uma área na cauda superior da curva entre 0,025 e 0,05. Mas, como o teste é bilateral (H_1: $\mu_1 \neq \mu_2$), a área deve ser dobrada para se ter o valor p correto. Veja o esquema a seguir:





Portanto: $0{,}05 < p < 0{,}10$.

Concluímos, então, que ao nível de significância de 5%, os dados não ımprovam uma diferença entre os dois métodos de ensinar matemática. Existe na probabilidade razoável, superior a 5%, de as diferenças observadas nos ıdos experimentais serem provenientes de fatores casuais.

Amostras de tamanhos diferentes

Quando as amostras têm tamanhos diferentes, a variância ɡregada é calculada por

$$S_a^2 = \frac{(n_1 - 1).S_1^2 + (n_2 - 1).S_2^2}{gl}$$

nde

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

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S_a \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

ınde

$\overline{X}_1$: 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 Num estudo realizado sobre alunos do segundo grau de escolas municipais do município de São José – SC, buscou-se verificar se dentre aqueles que já experimentaram algum tipo de droga, homens e mulheres o fizeram pela primeira vez com idades diferentes.[9] Colocando as hipóteses em termos dos valores médios de idades de homens e mulheres, tem-se:

$$H_0: \mu_1 = \mu_2 \quad \text{e} \quad H_1: \mu_1 \neq \mu_2,$$

onde μ_1: dentre os homens, a idade média que experimentaram droga pela primeira vez; e
μ_2: dentre as mulheres, a idade média que experimentaram droga pela primeira vez.

---

[9] Este trabalho foi realizado pelas alunas Kátia Vieira e Roseana Rotta na disciplina de Estatística, sem. 99/1, Curso de Psicologia da UFSC.

A pesquisa foi feita com 56 alunos (32 do sexo masculino e 24 do sexo feminino).[10] As idades em que cada um deles experimentaram droga pela primeira vez e os cálculos para se obter a estatística t são apresentados a seguir.

sexo	idade em que experimentou 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:

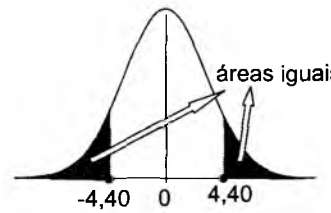
$$S_a^2 = \frac{(n_1 - 1).S_1^2 + (n_2 - 1).S_2^2}{gl} = \frac{(31).(6{,}371) + (23).(4{,}781)}{54} = 5{,}694$$

Desvio padrão agregado: $S_a = \sqrt{5{,}694} = 2{,}386$

Estatística do teste:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S_a \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{10{,}625 - 13{,}458}{(2{,}386).\sqrt{\frac{1}{24} + \frac{1}{32}}} = \frac{2{,}833}{(2{,}386).(0{,}270)} = -4{,}40$$

Como a Tabela V 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. Veja a figura ao lado.



Entrando na tabela com gl = 60 (o mais próximo do gl verdadeiro: igual a 54) e valor de t = 4,40, verificamos pela Tabela V que a área na cauda superior é inferior a 0,0005. Como o teste é bilateral, tem-se que o

[10] Na verdade, a pesquisa foi feita com um número bem maior de estudantes, mas somente 56 declararam já ter usado droga. E para o problema descrito, a amostra ficou restrita a estes 56 estudantes.

valor p é inferior a 0,001 (o dobro da área na cauda superior). O que leva o teste a rejeitar H_0 ao nível de significância de 0,05 ($p < 0{,}001 < 0{,}05 = \alpha$).

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, SIMSTAT,* etc.) que fazem os diversos métodos discutidos na literatura, com uma interface *amigável*. Até mesmo as planilhas eletrônicas estão incorporando técnicas básicas de estatística. A seguir, é listada uma saída do *Microsoft Excel*, com a aplicação do teste *t* aos dados do Exemplo 11.8.[11]

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*	
Stat t	*-4,39732*	
P(T<=t) uni-caudal	*0,000026*	
t crítico uni-caudal	*1,67357*	
P(T<=t) bi-caudal	*0,000052*	
t crítico bi-caudal	*2,00488*	

[11] No *Microsoft Excel*, várias técnicas estatísticas podem ser feitas acionando no menu principal "ferramentas", "suplementos" e solicitando que se instale as "ferramentas de análise". Acionar "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. Para realizar o teste *t* para dados pareados, discutido na seção anterior, escolher "ferramentas", "análise de dados" e "Teste T: duas amostras em par para a média". Para maiores detalhes ver Levine, Berenson, Stephan (2000).

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 linha temos os graus de liberdade e o valor da estatística *t*. Os resultados apresentados nas últimas quatro linhas dependem se estamos fazendo um teste unilateral (*uni-caudal*) ou bilateral (*bi-caudal*). 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.

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 acredita-se que as mães não são bem nutridas. Os dados observados estão apresentados na tabela seguinte.

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 da Localidade 2? Use $\alpha = 0,05$.

b) 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, onde 19 frangos, todos com um mês de vida, foram divididos aleatoriamente em dois grupos. No primeiro, com 12 frangos, usou-se a dieta A, enquanto que no segundo grupo, os 7 frangos foram tratados com a dieta B. No final de um mês encontrou-se 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) Verifique se existe diferença significativa entre alunos bolsistas e não bolsistas, com respeito ao tempo médio para a conclusão dos créditos do Curso de Pós-Graduação em Administração – UFSC, período 1980-84. Os dados estão na tabela seguinte.

Tempo, em meses, para conclusão de créditos de disciplinas dos alunos ingressados no período 1980 a 1984.

bolsistas	não bolsistas
62 24 30 34 54	56 34 60 62 42 63 69 66 44 54 50 61

Fonte: CPGA / UFSC.

11) Numa pesquisa sobre clima organizacional nos departamentos da UFSC, uma amostra de professores respondem a um questionário, onde, num dos itens, o respondente dava 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 Socioeconômico (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 a clareza organizacional dos departamentos são diferentes para os dois centros de ensino?

2) 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 estas localidades, comparando-as duas a duas.[12] 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 observadas	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

[2] 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 Wonnacott, Wonnacott (1981).

11.5 TAMANHO DAS AMOSTRAS

No planejamento de um estudo comparativo, surge a questão de qual o número n de elementos adequado para constituir 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, pode-se 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 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 ao problema de 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

$$\delta = \frac{|\mu_1 - \mu_2|}{\sigma}$$

A quantidade δ é a diferença de magnitude entre as verdadeiras médias em unidades de desvios padrão (σ) das populações em estudo. Supõe-se aqui que as duas populações tenham o mesmo desvio padrão.

Para se avaliar a quantidade n de elementos em cada grupo, o pesquisador precisa ser capaz de fornecer o valor mínimo de δ que leva a conseqüências práticas. Em geral, o pesquisador tem maior facilidade em raciocinar em termos da unidade em que está se medindo a variável em análise, mas, neste caso, torna-se necessário se ter uma avaliação de σ.

A Figura 11.9 indica o número mínimo n para que uma diferença δ seja detectada pelo teste estatístico com probabilidade 0,80 ($\beta = 0,20$) e com probabilidade 0,90 ($\beta = 0,10$).[13]

---

[13] O gráfico da Figura 11.8 foi construído a partir da função poder do teste t bilateral para amostras independentes, usando nível de significância de 5%. Procedeu-se um processo iterativo sobre as expressões apresentadas em Cochran e Cox (1957, Capítulo 2).

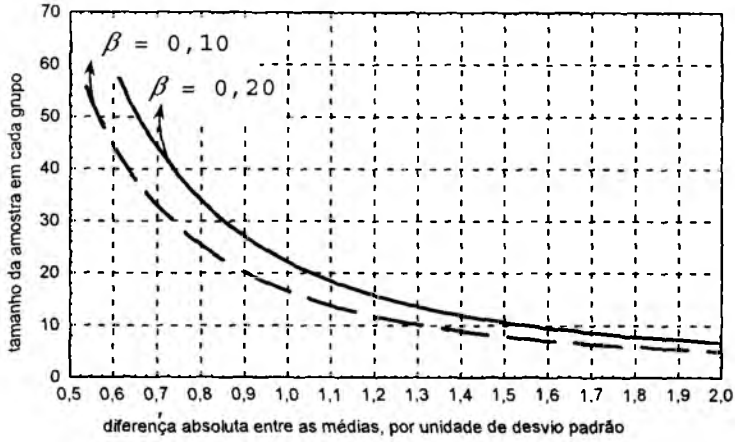
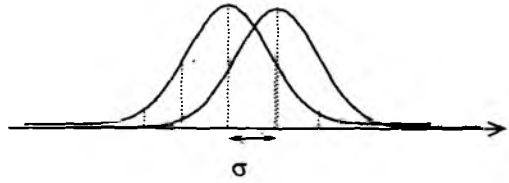


Figura 11.9 Tamanho mínimo da amostra, n, em cada grupo, em função da distância $\delta = |\mu_1 - \mu_2|/\sigma$ que se deseja detectar no teste estatístico.

Como exemplo, considere-se 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 em cada grupo). No final do estudo, o aprendizado de cada criança será avaliado numa escala de 0 a 10. Admita-se que os pesquisadores consideram relevantes uma diferença de 1,5 pontos entre as médias e, com base em estudos anteriores, o desvio padrão nesta escala não deve passar de 2 unidades. Logo, $\delta = {}^{1,5}/_{2} = 0{,}75$. Pelo gráfico da Figura 11.9, o número mínimo de crianças em cada grupo deve ser de aproximadamente $n = 37$ para $\beta = 0{,}10$ ou $n = 28$ para $\beta = 0{,}20$.

Exercício

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 destes grupos será aplicado um método de ensino. Quantas crianças deve-se ter em cada grupo, para garantir que um teste t para amostras independentes, ao nível de significância de 5%, detecte uma diferença de 1 desvio padrão com 90% de probabilidade? Admitindo distribuição normal, a diferença mínima que se quer detectar está representada na figura a seguir.



11.6 COMENTÁRIOS FINAIS

Na Seção 11.3 apresentou-se 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, conseqüentemente, tem-se um projeto de pesquisa melhor. Considere, por exemplo, o problema de se comparar dois tipos de materiais em termos do desgaste na sola de tênis de criança. Pode-se planejar um experimento, onde 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, decide-se por sorteio qual material vai ser usado (*aleatorização*). Depois de algum tempo, mede-se o desgaste das solas de todas as crianças do experimento e comparam-se 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 (teste *t* para dados pareados), destacando uma possível diferença entre os tipos de materiais. A Figura 11.10 ilustra a diferença de se considerar pares e de se considerar as duas amostras independentes na análise dos dados.

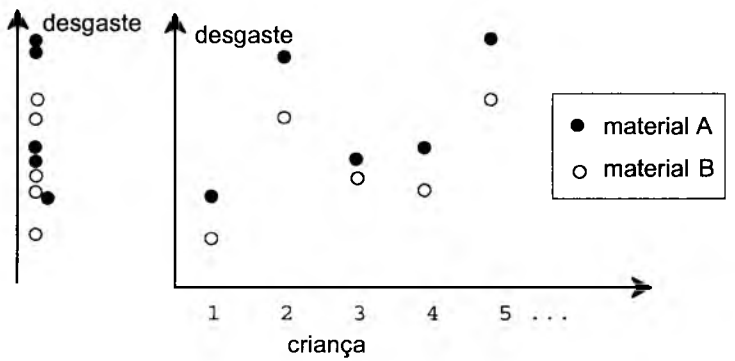


Figura 11.10 Ilustração de um conjunto de dados visto de forma pareada (à direita) e de forma independente (à esquerda).

Analisando a Figura 11.10, 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 fica ofuscada pelas diferenças entre as crianças.

A aplicação de testes *t* pode ser feita em estudos experimentais ou em estudos de levantamento. No exemplo precedente, tem-se 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, concluímos, também, que esta diferença é devida 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 ao nascer de crianças em duas localidades, podemos fazer um levantamento por amostragem, analisando os nascimentos nestas localidade. 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 sócio-econômicas, 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 *significativa estatisticamente*. 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 estas diferenças, que podem ser 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 destes 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 observados.[14] O teste dos sinais, visto no início deste capítulo, é um exemplo de teste não-paramétrico. Outro teste não-paramétrico é o qui- quadrado, a ser visto no capítulo seguinte.[15]

Exercícios complementares

14) Uma empresa de cerveja, após uma grande fusão, estuda a possibilidade de alterar o rótulo de uma de suas marcas, usando formas e cores mais vivas. Para avaliar se existe vantagem em alterar o rótulo, a empresa levou a cabo uma pesquisa de *marketing*. Enlatou a cerveja com o rótulo tradicional e com o rótulo novo. A pesquisa foi feita em 8 estabelecimentos comerciais. Em 4 deles, extraídos por sorteio, colocou-se o produto com o rótulo novo e, nos outros 4, manteve-se o produto com o rótulo tradicional. Após um mês, avaliou-se a quantidade vendida em cada estabelecimento. Os estabelecimentos que usaram o rótulo tradicional tiveram os seguintes resultados nas vendas (em milhares de unidades): 6, 5, 2, 2. Os estabelecimentos que usaram o rótulo novo tiveram os seguintes resultados nas vendas (em milhares de unidades): 4, 9, 5, 6. Os dados mostram evidência suficiente de que a média de vendas é superior com o rótulo novo? Responda usando um teste estatístico apropriado ao nível de significância de 5%.

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 6 estabelecimentos comerciais dispostos a colaborar com a pesquisa, colocaram-se as duas embalagens (de rótulo tradicional e de rótulo novo) da mesma cerveja. Tomou-se o cuidado para que em cada estabelecimento, a apresentação das duas embalagens do produto fosse feita de forma idêntica. Os resultados das vendas mensais (em milhares de unidades), para cada estabelecimento e cada embalagem foram as seguintes:

Estabelecimentos:	1	2	3	4	5	6
Rótulo tradicional:	16	12	28	32	19	25
Rótulo novo:	20	11	33	40	21	31

Os dados mostram evidência suficiente de que a média de vendas é superior com o rótulo novo? Responda usando um teste estatístico apropriado ao nível de significância de 5%.

---

[14] Os testes *t* supõem que os dados provenham de distribuições normais e, no caso do teste *t* para amostras independentes, supõem também que as populações tenham, aproximadamente, a mesma variância.

[15] Outros testes não-paramétricos podem ser vistos em Noether (1983) ou em Siegel (1975).

16) Com respeito a questão anterior, suponha que os gerentes dos estabelecimentos comerciais se recusaram a fornecer os valores das vendas, mas informaram com qual rótulo obteve-se maiores vendas. Nos estabelecimentos 1, 3, 4, 5 e 6 as vendas foram maiores com o rótulo novo e no estabelecimento 2 as vendas foram maiores com o rótulo tradicional. Estes dados são suficientes para afirmar que a maioria dos estabelecimentos devem vender mais cerveja com o rótulo novo? Responda usando um teste estatístico apropriado ao nível de significância de 5%.

17) Com o objetivo de avaliar o efeito de uma certa merenda escolar "reforçada", fez-se um estudo com dois grupos de crianças, que tinham princípios de desnutrição. Fizeram parte do estudo 7 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 abaixo apresentam o ganho de peso, em kg, durante seis meses.

	Par de criança						
Grupo	1	2	3	4	5	6	7
A	6	5	8	2	5	4	4
B	2	4	5	3	4	3	5

Esses dados têm evidência suficiente, capaz de 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) Num estudo sobre a Identidade Social dos professores com o Departamento a que pertencem, mostrou os seguintes resultados. Quanto maior o escore significa maior Identidade Social com o Departamento.

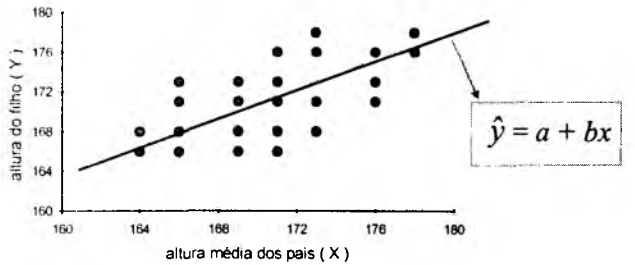
Depto. de Arquitetura: amostra de 24 professores, média de 40,8 e desvio padrão de 5,9.

Depto. de Psicologia: amostra de 19 professores, média de 42,5 e desvio padrão de 5,4.

Estes dados mostram evidências suficientes de que, em média, a Identidade Social com o Departamento é diferente quando comparamos os Deptos. de Arquitetura e Psicologia? Explique.

Parte V

Relacionamento 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 são mensuradas numericamente, mas, indicam certas qualidades, ou atributos, de tal forma que podemos alocar cada elemento numa categoria preestabelecida, resultando em *dados categorizados*. Por exemplo, ao observar a variável *sexo*, cada indivíduo pesquisado deve ser alocado, ou na categoria *masculino*, ou na categoria *feminino*. Lembramos que as variáveis devem estar bem definidas, tal que cada elemento pesquisado se encaixe *em uma e apenas em uma* categoria.

Um dos grandes propósitos em pesquisas nas Ciências Sociais é verificar se duas ou mais variáveis se apresentam *associadas*. Dizemos que duas variáveis estão associadas, se o conhecimento de uma altera a probabilidade de algum resultado da outra. Podemos dizer, por exemplo, que existe associação entre *a propensão de uma pessoa ir à praia* e o *clima*, pois, existe maior 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 se existe associação entre duas variáveis qualitativas, com base numa amostra de observações. Veremos, também, uma maneira de medir o grau de associação descrito pela amostra.

---

[1] Observamos que dizer que existe associação entre X e Y não implica, necessariamente, que X causa Y, ou que Y causa X. Desde que o conhecimento de uma delas altera a probabilidade dos resultados da outra, já se tem uma associação.

12.1 O TESTE DE ASSOCIAÇÃO QUI-QUADRADO

O teste de associação 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 ilustra o exemplo seguinte.

Exemplo 12.1 Para estudar a associação entre sexo (*masculino* ou *feminino*) e tabagismo (*fumante* ou *não-fumante*), numa certa população, observou-se uma amostra aleatória de 300 pessoas adultas desta 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.

	Sexo		
Tabagismo	masculino	feminino	Total
fumante	92	38	130
(%)	(46,00)	(38,00)	(43,33)
não-fumante	108	62	170
(%)	(54,00)	(62,00)	(56,67)
Total	200	100	300

Nota: As percentagens, entre parênteses, referem-se aos totais da variável sexo (totais das colunas).

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*. Deseja-se verificar se os dados da amostra mostram evidência suficiente para afirmarmos que, na população em estudo, existe associação entre sexo e tabagismo. Ou, equivalentemente, se existe diferença significativa entre a proporção de homens fumantes e a proporção de mulheres fumantes. Formalmente, temos 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 na população em estudo.[2]

Se $\pi_h = \pi_m$, então o conhecimento do sexo do indivíduo não fornece qualquer conhecimento sobre o fato de ele ser fumante ou não. 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 do indivíduo aumenta (ou diminui) a chance de ele ser fumante. Logo, a hipótese alternativa pode ser escrita como

H_1: Existe *associação* entre as variáveis *sexo* e *tabagismo*, na população em estudo.

O *teste qui-quadrado* também pode ser usado para comparar duas ou mais amostras, quando os resultados da variável resposta estão dispostos em categorias. O exemplo seguinte mostra esta situação.

Exemplo 12.2 Com o objetivo de verificar se três localidades são diferentes em termos do *grau de instrução do chefe da casa*, foram selecionadas amostras aleatórias de famílias nestas localidades, fazendo-se a classificação segundo o grau de instrução do chefe da casa. Os resultados estão apresentados na Tabela 12.2.

A Tabela 12.2 foi apresentada no Capítulo 4, onde interpretamos que, na *amostra observada*, existem diferenças entre as três localidades quanto ao perfil do grau de instrução do chefe da casa. Considerando, porém, que os dados referem-se a amostras, resta saber se estas diferenças são significativas, ou seja, se os dados mostram evidência suficiente para inferirmos que estas diferenças também existem nas populações de onde os dados foram extraídos.

---

[2] Neste livro, para testar as hipóteses em questão, adotaremos um procedimento bastante geral, conhecido como *teste qui-quadrado*. Mas, no presente exemplo, 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) e Triola (1999, p.226).

Tabela 12.2 Distribuição de freqüências do grau de instrução do chefe da casa, segundo a localidade da residência. Amostra de 120 famílias do Bairro Saco Grande II, Florianópolis – SC, 1988.

Grau de Instrução	Localidade		
	Monte Verde	Parque da Figueira	Encosta do Morro
nenhum	6	14	18
(%)	(15,0)	(32,6)	(48,7)
primeiro grau	11	14	13
(%)	(27,5)	(32,6)	(35,1)
segundo grau	23	15	6
(%)	(57,5)	(34,8)	(16,2)
Total	40	43	37
(%)	(100,0)	(100,0)	(100,0)

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

Formalmente, queremos testar as seguintes hipóteses:

H_0: As distribuições de freqüências do grau de instrução do chefe da casa *são iguais* nas três localidades;

H_1: As distribuições de freqüências do grau de instrução do chefe da casa *não* são iguais nas três localidades.

Se considerarmos que as três localidades formam categorias de uma variável, que chamaremos de *localidade da residência*, podemos colocar as hipóteses em termos de *independência* (H_0) e *associação* (H_1) entre as variáveis *localidade da residência* e *nível de instrução do chefe da casa*.[3]

De um modo geral, dadas duas variáveis qualitativas, as hipóteses do *teste qui-quadrado* podem ser formuladas da seguinte maneira:

H_0: As duas variáveis são *independentes*;

H_1: Existe *associação* entre as duas variáveis.

---

[3] Muitos autores preferem considerar a presente situação como um teste de *homogeneidade* entre as amostras das diferentes localidades, pois, na verdade a *localidade da residência* não é propriamente uma variável, mas sim uma referência aos subgrupos da população em estudo. Porém, o teste qui-quadrado é aplicado da mesma maneira.

No que segue, apresentaremos os procedimentos para a realização do teste qui-quadrado.

A estatística do teste

A estatística do teste, que designaremos por χ^2, é uma espécie de medida de distância entre as freqüências observadas, O, e as freqüências que esperaríamos encontrar em cada casela, E, na suposição das variáveis serem independentes. Ilustraremos a obtenção das freqüências esperadas (E) e da estatística χ^2, usando os dados da Tabela 12.1.

Exemplo 12.1 (continuação) Para obter as freqüências esperadas, consideraremos a distribuição percentual de fumantes e não fumantes em toda a amostra (43,33% de fumantes e 56,67% de não fumantes). Se tabagismo e sexo forem variáveis *independentes* (H_0), 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:

43,33% de 200 homens fumantes (E = (0,433).(200) = 86,67) e
56,67% de 200 homens não-fumantes (E = (0,5667).(200) = 113,33).

De forma análoga, podemos obter as freqüências esperadas no estrato das mulheres.

O cálculo das freqüências esperadas pode ser simplificado com a aplicação da seguinte fórmula, aplicada a *cada casela* da tabela de contingência:

$$E = \frac{(total\ da\ linha) \times (total\ da\ coluna)}{(total\ geral)}$$

Calcula-se, assim, as freqüências esperadas em cada casela:

	Sexo		
Tabagismo	masculino	feminino	Total
fumante	$E = {}^{(130).(200)}/_{300} = 86,67$	$E = {}^{(130).(100)}/_{300} = 43,33$	130
não-fumante	$E = {}^{(170).(200)}/_{300} = 113,33$	$E = {}^{(170).(100)}/_{300} = 56,67$	170
Total	200	100	300

A estatística do teste qui-quadrado, χ^2, é definida por

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

onde a soma se estende a todas as caselas da tabela de contingência.

O esquema seguinte mostra o cálculo das parcelas: $(O-E)^2/E$, que compõem a estatística χ^2, também conhecidas como *contribuições do* χ^2.

	Sexo	
Tabagismo	masculino	feminino
fumante	$(92 - 86{,}67)^2 / 86{,}67 = 0{,}328$	$(38 - 43{,}33)^2 / 43{,}33 = 0{,}656$
não fumante	$(108 - 113{,}33)^2 / 113{,}33 = 0{,}251$	$(62 - 56{,}67)^2 / 56{,}67 = 0{,}501$

E, portanto, $\chi^2 = 0{,}328 + 0{,}656 + 0{,}251 + 0{,}501 = 1{,}74$.

Quando as variáveis são *independentes* (H_0), as freqüências observadas tendem a ficar perto das freqüências esperadas. (*Apenas variações casuais!*) Neste caso, o valor de χ^2 deve ser pequeno. Em outras palavras, um valor pequeno de χ^2 indica que as variáveis podem ser independentes. Por outro lado, um valor grande na estatística χ^2, sinaliza que as diferenças entre as freqüências observadas e freqüências esperadas não devem ser meramente casuais, ou seja, deve haver *associação* entre as duas variáveis.

Como em todo teste estatístico, 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. Esta distribuição existe, desde que:

a) os dados estejam dispostos numa tabela de contingência propriamente dita, isto é, cada elemento observado é alocado *numa e apenas numa* casela; e

b) as amostras sejam grandes.

A verificação da adequação dos tamanhos das amostras é usualmente feita em termos das freqüências esperadas. A maioria dos autores consideram adequada a aplicação do teste qui-quadrado quando *todas* as freqüências esperadas forem maiores ou iguais a 5 (cinco).[4]

---

[4] Quando ocorrer alguma freqüência esperada menor do que cinco, pode-se aplicar o chamado teste exato de Fisher. Veja, por exemplo, Levin (1985, p.221).

No exemplo em discussão, as freqüências esperadas em cada uma das 4 caselas foram iguais a 86,67, 43,33, 113,33 e 56,67, portanto, todas superiores a 5, o que permite a realização do teste qui-quadrado.

A distribuição do teste (distribuição de referência)

Se as duas variáveis forem realmente independentes (H_0) e admitindo as condições (a) e (b), então os possíveis valores da estatística χ^2 seguem a chamada *distribuição qui-quadrado com* $gl = (\ell - 1).(c - 1)$ *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 $gl = (2 - 1).(2 - 1) = 1$. Logo, se H_0 for verdadeira, os possíveis valores da estatística χ^2 devem seguir uma distribuição qui-quadrado com $gl = 1$ grau de liberdade, como mostra a Figura 12.1a.

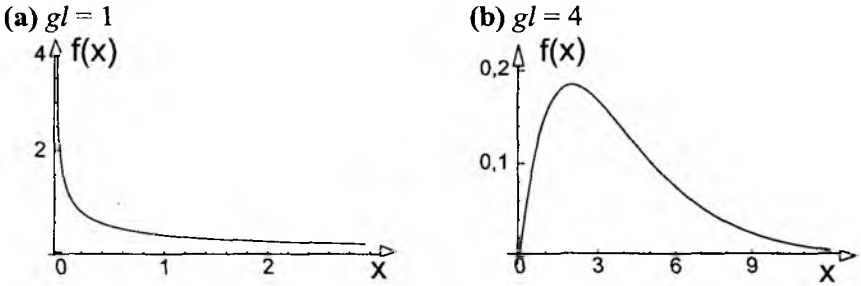


Figura 12.1 Distribuições qui-quadrado com $gl = 1$ e $gl = 4$.

A forma da distribuição qui-quadrado torna-se menos assimétrica à medida em que cresce o número de graus de liberdade (veja a Figura 12.1b).

Probabilidade de significância

A Figura 12.2 ilustra uma probabilidade de significância (valor p), como uma área sob a curva da distribuição qui-quadrado. Supondo que as duas variáveis sejam realmente independentes, o valor p representa a probabilidade de a estatística χ^2 acusar um valor maior ou igual do que o valor do χ^2 calculado a partir dos dados em análise.

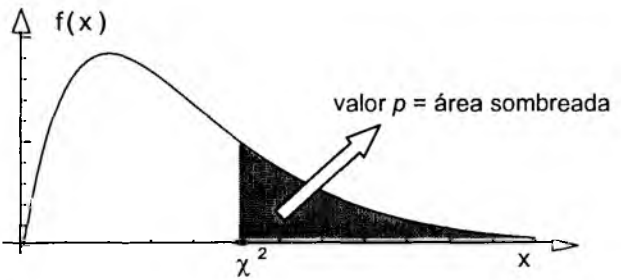
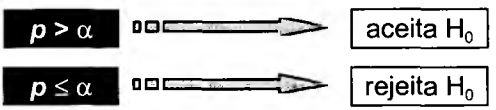


Figura 12.2 A probabilidade de significância *p*, como uma área sob a curva da distribuição qui-quadrado.

Quando os dados observados derivam um χ^2 grande (e, em conseqüência, um *p* pequeno – veja a Figura 12.2), 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 conseqüência, um *p* grande), o teste *não* pode rejeitar H_0, pois, o valor calculado de χ^2 está condizente com a distribuição dos possíveis valores de χ^2 construída à luz de H_0.

O limite entre aceitar H_0 e rejeitar H_0 pode ser feito pela comparação do valor *p* com o nível de significância α arbitrado. Lembramos que o nível de significância representa o risco tolerável do erro de *rejeitar* H_0, *quando* H_0 *é verdadeira* e é usual arbitrar $\alpha = 0{,}05$. Conforme vimos no Capítulo 10, a regra geral da decisão de um teste estatístico é

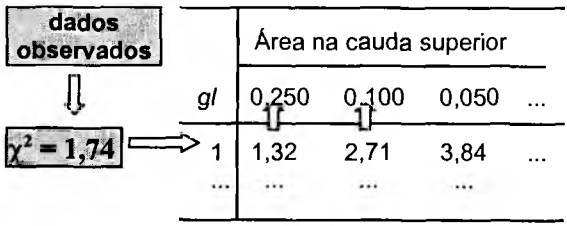


A tabela da distribuição qui-quadrado

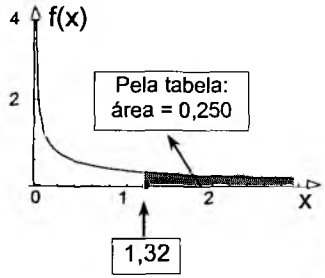
Depois de calculado o valor da estatística χ^2, podemos obter a probabilidade de significância *p*, usando uma tabela da distribuição qui-quadrado (Tabela VI do apêndice). A continuação do Exemplo 12.1 ilustra o uso desta tabela.

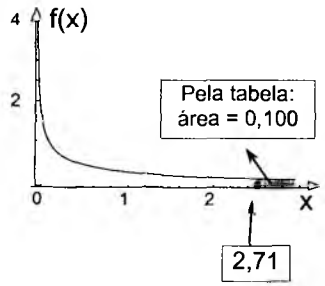
Exemplo 12.1 (continuação) Usando a Tabela VI do apêndice, entramos na linha correspondente a com $gl = 1$. Verificamos que o valor calculado $\chi^2 = 1{,}74$ está em torno dos valores 1,32 e 2,71 da tabela, os quais estão

associados às áreas na cauda superior de 0,25 e 0,10, respectivamente, conforme ilustra o seguinte esquema:

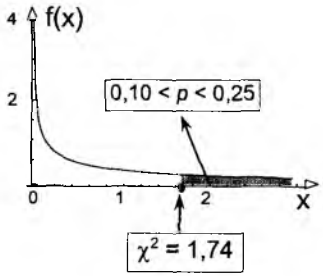


Graficamente:





Logo, para o valor calculado ($\chi^2 = 1,74$), tem-se:



Portanto, o valor *p* está 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 2x2

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 2x2, especialmente quando as amostras não forem muito grandes (por exemplo, quando existir alguma freqüência esperada entre 5 e 10), recomendamos aplicar a chamada *correção de continuidade de Yates*, que consiste em reduzir 0,5 unidades nas diferenças absolutas entre as freqüências observadas e esperadas.[5] E a fórmula da estatística χ^2 para tabelas de contingência 2x2, com correção de continuidade, é dada por

$$\chi^2 = \sum \frac{(|O-E|-0,5)^2}{E}$$

onde o símbolo das duas barras verticais, $|\ |$, significa valor absoluto. Então, depois de calcular a diferença entre O e E, devemos desprezar o sinal (+ ou -) e reduzir 0,5 unidades.

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 casela:

	Sexo	
Tabagismo	masculino	feminino
fumante	$(\|92 - 86,67\| - 0,5)^2/86,67$ = 0,269	$(\|38 - 43,33\| - 0,5)^2/43,33$ = 0,538
não-fumante	$(\|108 - 113,33\| - 0,5)^2/113,33$ =0,206	$(\|62 - 56,67\| - 0,5)^2/56,67$ = 0,412

Donde: $\chi^2 = 0,269 + 0,538 + 0,206 + 0,412 = 1,43$.

Usando a Tabela VI 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 nestes casos que a correção é mais recomendada.

---

[5] Numa tabela 2x2, a distribuição dos possíveis valores da estatística χ^2, quando calculada *com* a correção de continuidade, aproxima-se mais da distribuição qui-quadrado com $gl = 1$ do que quando calculada *sem* esta correção.

Uma fórmula mais rápida para o cálculo do χ^2 em tabelas 2x2

Em tabelas 2x2, 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*

$$\chi^2 = \frac{n.\left(|ad - bc| - \frac{n}{2}\right)^2}{(a+b).(c+d).(a+c).(b+d)}$$

Vamos ilustrar o uso desta fórmula com os dados da Tabela 12.1:

a = 92	*b* = 38	130
c =108	*d* = 62	170
200	100	300

$$\chi^2 = \frac{(300).\left(|(92).(62) - (38).(108)| - \frac{300}{2}\right)^2}{(130).(170).(200).(100)}$$

Donde: $\chi^2 = \frac{(300).[1600 - 150]^2}{442000000} = \frac{(300).(2102500)}{442000000} = 1{,}43$

Para calcular a estatística χ^2 sem a correção de continuidade, basta excluir a fração $^n/_2$ do numerador da expressão apresentada neste tópico.

Aplicação do teste qui-quadrado em tabelas de grande dimensão

Exemplo 12.3 (Box, Hunter e Hunter, 1978, p.145) Considere um estudo exploratório em que se está 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 que o Hospital E é um hospital de referência, que recebe os casos mais graves. A Tabela 12.3 mostra os resultados de um levantamento por amostragem, realizado nos cinco hospitais.

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 fornece os resultados das freqüências esperadas e as parcelas de cada casela no cálculo da estatística χ^2, conforme a formulação apresentada na seção anterior.

Tabela 12.3 Resultados (freqüências 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	5	8	21	43
(%)	(27,7)	(16,1)	(10,1)	(16,4)	(52,4)
parcial	18	10	36	56	29
(%)	(38,3)	(32,3)	(45,6)	(43,8)	(35,4)
completa	16	16	35	51	10
(%)	(34,0)	(51,6)	(44,3)	(39,8)	(12,2)

Tabela 12.4 Resultados do procedimento cirúrgico: freqüências observadas (centro), freqüências 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 o valor da estatística do teste: $\chi^2 = 56{,}7$.

Usando a tabela da distribuição qui-quadrado (Tabela VI do apêndice), com $gl = (\ell - 1).(c - 1) = (3 - 1).(5 - 1) = 8$, verificamos que a probabilidade de significância p é inferior a 0,001. Então, para qualquer nível usual de significância (por exemplo, $\alpha = 0{,}05$), o teste detecta uma 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, submetidos à cirurgia em questão.

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. Um caso muito comum é comparar grupos de categorias agregadas segundo algum critério e, posteriormente, estudar separadamente as categorias que estavam agrupadas. Na seqüência do Exemplo 12.3, ilustramos este procedimento.

Exemplo 12.3 (continuação) Observando as parcelas da estatística χ^2 (canto inferior direito das caselas da Tabela 12.4), verificamos que as maiores contribuições partiram do Hospital E, que é um hospital de referência e, portanto, recebe os casos mais graves. Podemos, então, fazer uma análise estatística mais elaborada, para verificar se a significância foi devida a 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. Os valores das freqüências observadas na coluna dos hospitais comuns corresponde à soma das freqüências observadas dos hospitais A, B, C e D da Tabela 12.4. As freqüências esperadas e as parcelas do χ^2 foram calculadas novamente.

Tabela 12.5 Comparação do hospital de referência com os demais. Freqüências observadas (centro), freqüências 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 VI, chegamos a conclusão que $p < 0{,}001$, mostrando haver uma diferença significativa entre os hospitais comuns e o hospital de referência.

Finalmente, a Tabela 12.6 analisa os hospitais comuns entre si. As freqüências observadas desta tabela correspondem às freqüências observadas da Tabela 12.4, eliminando o Hospital E.

Tabela 12.6 Comparação entre os hospitais comuns. Freqüências observadas (centro), freqüências 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 *0,18*	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, buscou-se verificar uma possível associação entre o local da residência e a utilização de programas de alimentação popular. Segue uma saída do pacote computacional SIMSTAT.[6]

---

6 Ver www.simstat.com

```
CROSSTAB:
      PAP          Programas de alimentação popular
   by LOCAL        Local da residência

  LOCAL->
             Count | Monte    | Pq. da   | encosta |
           Col Pct | Verde    | Figueira | do morro|
                   |      1   |      2   |      3  |   Total
PAP        ---------+----------+----------+---------+---------
                 0 |     18   |     12   |     12  |     42
  não usa          |   45,0   |   27,9   |   32,4  |   35,0
                   +----------+----------+---------+
                 1 |     22   |     31   |     25  |     78
  usa              |   55,0   |   72,1   |   67,6  |   65,0
           --------+----------+----------+---------+
             Column |     40   |     43   |     37  |    120
              Total |   33,3   |   35,8   |   30,8  |  100.0

Chi-Square              Value       D.F.      Significance
-----------             -----       ----      ------------
Pearson                 2,8164      2          0,2446
Likelihood ratio        2,7915      2          0,2477

Smallest expected frequency = 12,950
Cells with expected frequency less than 5 =    0 of 6 (0,0%)

Statistic                              Value          Significance
---------                              -----          ------------
Contingency Coefficient                0,15143

VALID CASES: 120      MISSING CASES: 0
```

A partir dos dados brutos, é construída uma tabela de contingência. O teste qui-quadrado (*Chi-square of Pearson*) é apresentado logo abaixo da tabela de contingência com os resultados $\chi^2 = 2{,}8164$, $gl = 2$ e $p = 0{,}2446$, mostrando não haver associação (aceitando H_0). O pacote apresenta, também, outra abordagem do teste qui-quadrado (*Likelihood ratio*), conduzindo à mesma conclusão ($p = 0{,}2477$). Em seguida, é apresentado o menor valor das freqüências esperadas e em quantas caselas obteve-se freqüências esperadas menores do que 5. No presente exemplo, como a menor freqüência esperada é 12,95 e, portanto, não há freqüências esperadas inferiores a 5, o teste é válido. Finalmente, é apresentado o coeficiente de contingência igual a 0,1514, que será comentado na próxima seção.

Exercícios

1) Seja a seguinte amostra:

Classificação de uma amostra de 38 indivíduos, quanto a 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 *freqüentadores*) 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 *freqüência às aulas* e *aprovação no exame final*? Use $\alpha = 0{,}05$.

3) a) Faça um teste qui-quadrado sobre os dados da Tabela 12.2, para verificar se existe diferença significativa entre as distribuições do nível de instrução do chefe da casa, nas três localidades estudadas. Use $\alpha = 0{,}01$.

b) Verifique se existe diferença significativa na distribuição do nível de instrução do chefe da casa 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 do chefe da casa 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 grau de instrução do chefe da casa.[7]

---

[7] 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 *grau 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, a partir de um conjunto de observações. É um processo de inferência, em que se parte dos dados para se tirar conclusões sobre o universo de onde estes 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, torna-se 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 estas medidas são descritivas, isto é, referem-se apenas aos dados observados.

O cálculo destes coeficientes de associação também costuma ser realizado após a aplicação de um teste estatístico, quando estes detectam associação. Neste caso, um coeficiente de associação fornece uma *estimativa do grau de associação* entre as duas variáveis.

Exemplo 12.4 Vamos contrapor dois conjuntos de pessoas, classificadas segundo o sexo (*homem* ou *mulher*) e tabagismo (*fumante* ou *não fumante*). Os resultados destas duas amostras estão apresentados nas Tabelas 12.7 e 12.8. Na amostra A, os dados indicam uma situação de completa *independência*, pois o conhecimento do sexo do respondente não fornece qualquer informação sobre à variável tabagismo (veja que a percentagem de homens fumantes é igual a percentagem de mulheres fumantes). Por outro lado, a amostra B ilustra um caso de *associação perfeita* (pois, os fumantes são todos homens e os não-fumantes são todos mulheres).

Duas amostras de 300 pessoas cada, classificadas segundo o sexo (homem ou mulher) e tabagismo (fumante ou não fumante).

Tabela 12.7 Amostra A.

	Sexo	
Tabagismo	homem	mulher
fumante	80 (40%)	40 (40%)
não-fumante	120 (60%)	60 (60%)

Tabela 12.8 Amostra B.

	Sexo	
Tabagismo	homem	mulher
fumante	200	0
não-fumante	0	100

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. E, ainda, o *quanto* se aproximam.

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 Tabelas 12.7 e 12.8, sem a correção de continuidade, encontramos os seguintes valores: $\chi^2 = 0$ (para a Tabela 12.7) e $\chi^2 = 300$ (para a Tabela 12.8). 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.

O 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 a partir da estatística χ^2 e do número total de elementos observados, n, da seguinte forma:[8]

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}}$$

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, como a Tabela 12.4, 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 completa independência. Será 1 somente quando houver associação perfeita. Valores de C^* próximos de 1 descrevem

---

[8] Para calcular o coeficiente de contingência é conveniente calcular χ^2 sem a correção de continuidade.

uma *associação forte*, enquanto que 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) Na Tabela 12.7, temos: $n = 300$, $k = 2$ e $\chi^2 = 0$. Então:

$$C^* = \sqrt{\frac{(2).(0)}{(2-1).(0+300)}} = 0 \Longrightarrow \textit{completa independência!}$$

Na Tabela 12.8, temos: $n = 300$, $k = 2$ e $\chi^2 = 300$. Então:

$$C^* = \sqrt{\frac{(2).(300)}{(2-1).(300+300)}} = 1 \Longrightarrow \textit{associação perfeita!}$$

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 (donde, $k = 3$), acusando um $\chi^2 = 56{,}7$. Então:

$$C^* = \sqrt{\frac{3.(56{,}7)}{2.(367+56{,}7)}} = 0{,}45$$

Donde concluímos que a amostra descreve uma associação moderada entre *hospital* e *recuperação funcional de pacientes*.

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 o seu grau de instrução?

( ) nenhum
( ) primeiro grau incompleto
( ) primeiro grau completo
( ) segundo grau (completo ou incompleto)
( ) superior (completo ou incompleto)

(b) Qual 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 na verificação da existência de associação, mas também no seu sentido (positiva ou negativa). Dizemos que existe *associação* (ou *correlação*) *positiva* quando, na medida em que o nível de uma variável aumenta, cresce a chance de ocorrer níveis elevados na outra variável; *associação* (ou *correlação*) *negativa* ocorre quando, ao aumentar o nível de uma variável, diminui a chance de ocorrer níveis elevados na outra variável. No presente contexto, preferimos usar o termo *correlação* no lugar de *associação*.

O coeficiente de correlação que apresentaremos a seguir baseia-se 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: *alto* → *baixo*; peso: *pesado* → *leve*). E de Maria para João, ambas as variáveis mudam para níveis superiores (estatura: *baixo* → *alto*; peso: *leve* → *pesado*).

Pedro é baixo e pesado;
José é alto e leve

Pedro e José, por outro lado, formam um par discordante, pois, ao passar do Pedro para o José, a estatura aumenta, enquanto que o peso diminui (estatura: *baixo* → *alto*; 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, agora, 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 associação entre *nível de instrução* e *posição com relação ao aborto*, relativo aos dados da Tabela 12.9.

Tabela 12.9 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 já estão dispostas numa mesma ordem (ambas estão em ordem crescente), passamos a contar o número de concordâncias e o número de discordâncias.

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)

Portanto: n_d = 109.063

O coeficiente γ de Goodman e Kruskal

O coeficiente γ considera a diferença entre o número de concordâncias e o número de discordância (n_c - n_d), dividida pelo número total de pares concordantes ou discordantes (n_c + n_d). Ou seja:

$$\gamma = \frac{n_c - n_d}{n_c + n_d}$$

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 está superando o número de discordâncias (correlação positiva forte). Simetricamente, quanto mais próximo de -1 estiver γ, mais o número de discordâncias está superando o número de concordâncias (correlação negativa forte).

Exemplo 12.6 (continuação) Calculamos n_c = 246.960 e n_d = 109.063. Donde:

$$\gamma = \frac{246960 - 109063}{246960 + 109063} = 0{,}39$$

Concluímos, então, que a amostra apresenta uma correlação positiva moderada entre *grau 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 grau de instrução e a renda familiar. Segue uma saída do pacote computacional SIMSTAT.

CROSSTAB:

RENDA_C Categorias de renda em salários mínimos
by GI Grau de instrução
GI->

RENDA_C	Count / Col Pct	nenhum completo 1	primeiro grau 2	segundo grau 3	Total
até 4,9	1	24	18	10	52
		64,9	47,4	22,7	43,7
de 5,0 a 9,9	2	11	14	22	47
		29,7	36,8	50,0	39,5
10 ou mais	3	2	6	12	20
		5,4	15,8	27,3	16,8
	Column Total	37	38	44	119
		31,1	31,9	37,0	100.0

Chi-Square	Value	D.F.	Significance
Pearson	16,2822	4	0,0027
Likelihood ratio	17,3020	4	0,0017

Smallest expected frequency = 6,218
Cells with expected frequency less than 5 = 0 of 9 (0,0%)

Statistic	Value	Significance
Contingency Coefficient	0,34693	
Kendall's Tau-b	0,33006	0,0001
Gamma	0,49507	

VALID CASES: 119 MISSING CASES: 1

O resultado do teste qui-quadrado de Pearson ($\chi^2 = 16{,}28$, $gl = 4$ e $p = 0{,}0027$) leva a rejeição de H_0, isto é mostra haver associação entre renda e grau de instrução. O coeficiente de contingência igual a 0,347 indica uma associação moderada. O coeficiente γ, em torno de 0,5, indica uma correlação positiva moderada.

Não existe um teste estatístico direto sobre o coeficiente γ, mas existem outros coeficientes baseados na idéia de pares concordantes e discordantes, dentre eles o τ_b de Kendall, que no exemplo apresentou os

resultados $\tau_b = 0{,}33$ com $p \approx 0{,}0001$, indicando que a correlação positiva é significativa.[9] Cabe a observação que houve um caso inválido (falta de resposta), ou seja, a análise foi realizada com 119 famílias e não com as 120 famílias amostradas.

Na literatura, encontram-se vários coeficientes de associação para variáveis qualitativas. Uma boa discussão sobre estes coeficientes pode ser encontrada em Leach (1979).

Exercícios

5) Calcule o coeficiente C^* para os dados da Tabela 12.1 e interprete o resultado.

6) Calcule o coeficiente C^* para os dados da Tabela 12.2 e interprete o resultado.

7) Noventa crianças foram classificadas segundo suas habilidades em matemática e música, resultando nos seguintes dados.

Habilidade	Habilidade para matemática		
para música	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.

---

[9] No teste sobre o coeficiente τ_b, a hipótese nula afirma ausência de correlação e a hipótese alternativa a presença de correlação. Como no exemplo em questão, encontrou-se $p < 0{,}05$, o teste rejeitou H_0, provando estatisticamente a presença da correlação na população em estudo.

Exercícios complementares

10) A tabela que segue apresenta uma classificação de pessoas classificadas em termos do grau de instrução e em termos da colaboração com a coleta seletiva de lixo. Estes dados fazem parte de uma pesquisa realizada em Florianópolis – SC, em 1999.[10] Verifique se existe associação significativa entre estas duas variáveis.

Grau de instrução	Colabora com a coleta seletiva de lixo	
	sim	não
até o 1º grau	22	13
2º grau (compl. ou incompl.)	33	34
superior (compl. ou incompl.)	39	36

11) Os dados abaixo referem-se ao tipo de escola que o aluno estudou o segundo grau (0 = *pública* e 1 =*particular*) e o resultado do vestibular (0 = *não passou* e 1 = *passou*) de uma amostra de 30 alunos.

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

Construa uma distribuição de freqüências conjunta para as variáveis *tipo de escola* e *resultado do vestibular*. Apresente esta 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 existe associação entre três áreas de estudo (humanas, biológica e exatas) e a favorabilidade em relação ao exame de final de curso proposto pelo governo (favorável ou contrário), em estudantes universitários, observaram-se 120 estudantes aleatoriamente. Dos 40 estudantes da área de humanas, 10 eram favoráveis (e os restantes contrários). Dos 30 estudantes da área biológica, 10 eram favoráveis (e os restantes contrários). E dos 50 da área exatas, 20 eram favoráveis (e os restantes contrários). Pode-se dizer que existe

---

[10] Os dados foram coletados pelos alunos João Fáveri e Ângela Queiroz do Curso de Psicologia da UFSC, semestre 99/1.

associação entre estas 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, feminino);
- Local da residência (na própria cidade, em outra cidade);
- Grau 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 grau de satisfação com o local de residência do cliente.

Capítulo 13

Correlação e regressão

Neste capítulo, vamos dar seqüência ao estudo de associação entre duas variáveis, mas agora, supondo que ambas as variáveis sejam mensuradas *quantitativamente*. Usaremos, neste caso, o termo *correlação* no lugar de *associação*.

Variáveis correlacionadas

Dizemos que duas variáveis, X e Y, estão *positivamente correlacionadas* quando elas *caminham num mesmo sentido*, ou seja, elementos com valores pequenos de X tendem a ter valores pequenos de Y e elementos com valores grandes de X tendem a ter valores grandes de Y. Estão *negativamente correlacionadas* quando elas *caminham em sentidos opostos*, ou seja, elementos com valores pequenos de X tendem a ter valores grandes de Y e elementos com valores grandes de X tendem a ter valores pequenos de Y.

As variáveis *peso* e *altura*, por exemplo, apresentam-se, em geral, *correlacionadas positivamente*, pois a maioria dos indivíduos altos também são pesados, enquanto que a maioria dos indivíduos baixos são leves. Por outro lado, no Brasil, as variáveis *renda familiar* e *número de elementos da família* costumam se apresentar *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 municípios catarinenses.

Tabela 13.1 - Alguns dados de doze importantes municípios catarinenses.

município	população (em 1000 hab.)	pop. urbana (em 1000 hab.)	% de pop. urbana	taxa de cresc. demográfico	taxa de mort. infantil	taxa de alfabetização
Itajaí	101	94	93	3,19	37	85
Blumenau	193	181	94	4,60	27	90
Rio do Sul	42	39	94	2,78	38	85
Joinville	304	292	96	6,46	25	87
Curitibanos	42	32	76	1,99	67	75
Lages	152	126	83	1,89	63	78
Canoinhas	55	36	66	2,92	41	81
Chapecó	105	77	73	5,32	13	75
Concórdia	68	25	37	2,71	28	84
Florianópolis	219	186	85	3,11	17	87
Criciúma	129	116	90	3,11	32	85
Laguna	42	33	78	1,21	32	77

Fonte: Municípios Catarinenses - Dados Básicos, GAPLAN-SC (1987).

Notas sobre as variáveis:

(1) *população*: população estimada residente no município, em mil habitantes, ano de 1986.
(2) *pop. urbana*: população estimada residente em áreas urbanas, em mil hab., ano de 1986.
(3) *% de pop. urbana* = (pop. urbana / população).(100).
(4) *taxa de cresc. demográfico*: taxa média geométrica de incremento anual da população, 1970/ 80.
(5) *taxa de mort. infantil*: coeficiente de mortalidade infantil por 1000 nascidos vivos, 1982.
(6) *taxa de alfabetização*: percentagem de adultos alfabetizados.

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 observado de uma variável e y é o correspondente valor da outra variável. A Figura 13.1 ilustra a construção de um diagrama de dispersão.

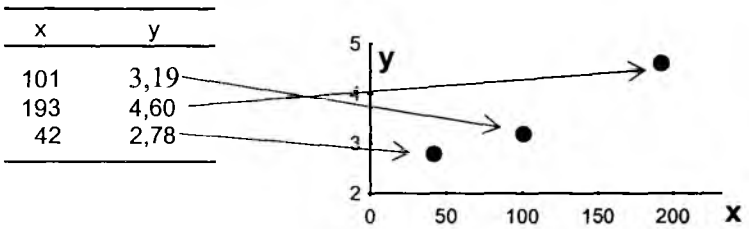


Figura 13.1 Construção de um diagrama de dispersão. Representação das três primeiras observações de X = *população residente* e Y = *taxa de crescimento demográfico*, referente aos dados da Tabela 13.1.

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 parte de *dados pareados*.

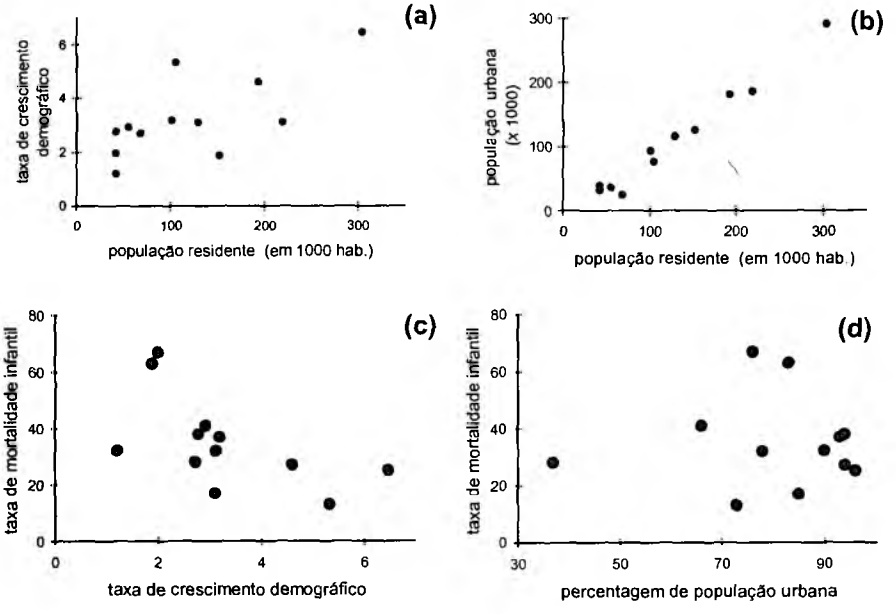


Figura 13.2 Alguns diagramas de dispersão, construídos a partir dos dados da Tabela 13.1.

Os diagramas da Figura 13.1a e 13.1b mostram duas situações de *correlações positivas*, pois, em ambos os casos, os pontos estão em torno de uma linha imaginária *ascendente*. Valores pequenos de uma variável tendem a estar associados a valores pequenos da outra, o mesmo acontecendo para valores grandes. Em (b) os dados apresentam-se *mais próximos* de uma linha ascendente do que em (a), o que caracteriza uma correlação *mais forte*.

A Figura 13.1c mostra que os dados observados de *taxa de crescimento demográfico* e *taxa de mortalidade infantil* têm correlação negativa, pois os pontos estão em torno de uma linha imaginária descendente.

Os dados observados da *percentagem de população urbana* e *taxa de mortalidade infantil*, Figura 13.1d, não sugerem um relacionamento entre estas duas variáveis, nos municípios em estudo, 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.

Os diagramas de dispersão, além de permitirem visualizar uma possível correlação nos dados observados, podem, também, indicar alguns outros aspectos relevantes na análise exploratória de dados. Na Figura 13.1d, por exemplo, observamos a presença de um ponto discrepante dos demais (coordenadas $X = 37$ e $Y = 28$). O município referente a este ponto discrepante (Concórdia) poderia ser estudado isoladamente dos demais.

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 brasileira* e *venda de carros japoneses* 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 interpretação relevante.

A análise de dados para verificar correlações é usualmente feita em termos exploratórios, onde 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.

Exercícios

1) Considerando os dados da Tabela 13.1, construir um diagrama de dispersão para as variáveis *taxa de alfabetização* e *taxa de mortalidade infantil*. 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 observados destas duas variáveis.

   b) Existe algum aluno que *foge* ao comportamento geral dos demais (ponto discrepante)?

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 localidade da Encosta do Morro, faça um diagrama de dispersão com os dados de: X = *renda familiar* e Y = *número de moradores no domicílio*. Interprete.

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*.

A idéia da construção do coeficiente de correlação de Pearson

O valor do coeficiente de correlação não deve depender da unidade de medida dos dados. Por exemplo, o coeficiente de correlação entre as variáveis *peso* e *altura*, observadas num certo conjunto de indivíduos, deve acusar o mesmo valor, independentemente se o peso for medido em *gramas* ou *quilogramas* e a altura em *metros* ou *centímetros*.

Para evitar o efeito da unidade de medida, 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:

$$r = \frac{\Sigma(x'.y')}{n-1}$$

onde:

n é o tamanho da amostra, isto é, o número de pares (x, y) observados e
$\Sigma(x'.y')$ é a soma dos produtos $x'.y'$ dos pares de valores padronizados, isto é, para cada par (x', y'), faz-se o produto $x'.y'$ e, depois, somam-se os resultados destes produtos.

Exemplo 13.1 Vamos mostrar o cálculo do coeficiente de correlação de Pearson, usando os dados das variáveis X = *população residente* e Y = *taxa de crescimento populacional*, relativas aos municípios da Tabela 13.1. A Tabela 13.2 mostra alguns cálculos intermediários.

Tabela 13.2 Obtenção de valores padronizados e produtos $x'.y'$ para o cálculo de r.

valores originais		valores padronizados		produtos
x	y	x'	y'	$x'.y'$
101	3,2	-0,24	-0,05	0,012
193	4,6	0,87	0,88	0,766
42	2,8	-0,95	-0,32	0,304
304	6,5	2,20	2,15	4,730
42	2,0	-0,95	-0,85	0,808
152	1,9	0,37	-0,91	-0,337
55	2,9	-0,79	-0,25	0,198
105	5,3	-0,19	1,35	-0,257
68	2,7	-0,63	-0,38	0,239
219	3,1	1,18	-0,12	-0,142
129	3,1	0,10	-0,12	-0,012
42	1,2	-0,95	-1,38	1,311
$\overline{X}$ = 121,0	$\overline{Y}$ = 3,275			$\Sigma(x'.y')$ = 7,620
S_x = 83,037	S_y = 1,503			

$$r = \frac{\Sigma(x'.y')}{n-1} = \frac{7{,}620}{11} = 0{,}69$$

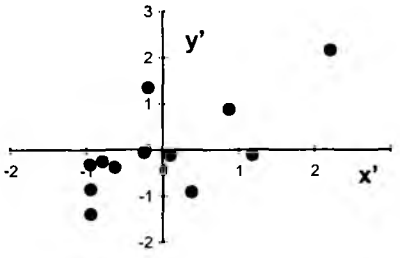


Figura 13.4 Diagrama de dispersão dos valores padronizados do Exemplo 13.1.

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 $x'.y'$ resultam em valores positivos e, em conseqüência, tem-se o coeficiente r positivo. A Figura 13.4 ilustra esta situação. Os quadrantes I e III (onde x' e y' têm o mesmo sinal), estão com maior concentração de pontos longe da origem, acarretando num valor de r positivo.

O exemplo seguinte mostra o cálculo do coeficiente r para uma situação de correlação negativa.

Exemplo 13.2 Cálculo do coeficiente de correlação de Pearson com os dados das variáveis X = *taxa de crescimento populacional* e Y = *taxa de mortalidade infantil*, relativas aos municípios da Tabela 13.1. A Tabela 13.3 mostra os cálculos intermediários.

Tabela 13.3 Obtenção de valores padronizados e produtos x'.y' para o cálculo de r.

valores originais		valores padronizados		produtos
x	y	x'	y'	$x'.y'$
3,2	37	-0,05	0,12	-0,006
4,6	27	0,88	-0,49	-0,431
2,8	38	-0,32	0,18	-0,058
6,5	25	2,15	-0,61	-1,312
2,0	67	-0,85	1,97	-1,675
1,9	63	-0,91	1,73	-1,574
2,9	41	-0,25	0,37	-0,093
5,3	13	1,35	-1,36	-1,836
2,7	28	-0,38	-0,43	0,163
3,1	17	-0,12	-1,11	0,133
3,1	32	-0,12	-0,18	0,022
1,2	32	-1,38	-0,18	0,248
$\overline{X}$ = 3,275	$\overline{Y}$ = 35,0		$\Sigma(x'.y')$ = -6,419	
S_x = 1,503	S_y =16,226			

$$r = \frac{\Sigma(x'.y')}{n-1} = \frac{-6{,}419}{11} = -0{,}58$$

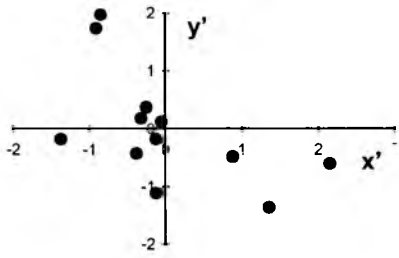


Figura 13.5 Diagrama de dispersão dos valores padronizados do Exemplo 13.2.

Quando estamos trabalhando com dados correlacionados negativamente, como no Exemplo 13.2, os pares (x', y') tendem a ter sinais trocados, especialmente para aqueles pontos longe da origem. Isto tende a levar os produtos $x'.y'$ a resultarem em valores negativos e, em conseqüência, tem-se o 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.

Para qualquer conjunto de dados, o valor do coeficiente de correlação de Pearson, r, estará no intervalo de -1 a 1. Será *positivo* quando os dados apresentarem correlação linear positiva; será *negativo* quando os dados apresentarem correlação linear negativa.

O valor de r será *tão mais próximo* de 1 (ou -1) quanto mais *forte* for a correlação nos dados observados. Teremos $r = +1$ se os pontos estiverem exatamente sobre uma reta ascendente (*correlação positiva perfeita*). Por outro lado, teremos $r = -1$ se os pontos estiverem exatamente sobre uma reta descendente (*correlação negativa perfeita*). Quando não houver correlação nos dados, r acusará um valor próximo de 0 (zero).

A Figura 13.6 mostra os possíveis valores de r e a interpretação em termos do sentido (positivo ou negativo) e da força (fraca, moderada ou forte) da correlação. E a Figura 13.7 compara formas de diagramas de dispersão com valores de r.

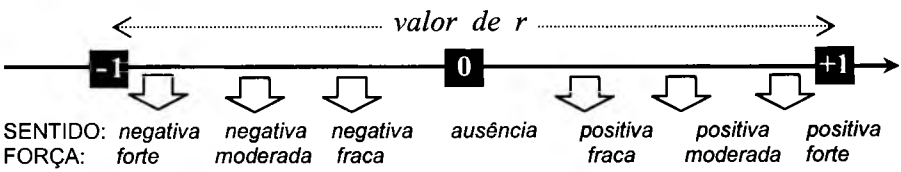
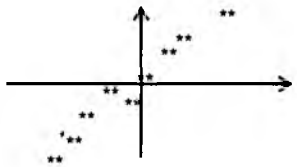
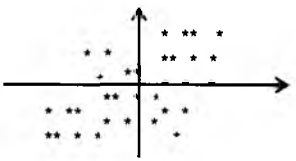


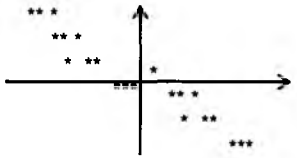
Figura 13.6 Sentido e força da correlação em função do valor de r.



correlação positiva forte (r ≈ 0,90)



correlação positiva fraca (r ≈ 0,30)



correlação negativa forte (r ≈ -0,90)

correlação negativa fraca (r ≈ -0,30)

Figura 13.7 Representações de pontos em diagramas de dispersão, em termos do sentido e força da correlação.

O método usual para se calcular r

Efetuar o cálculo do coeficiente de correlação *r* pela maneira que apresentamos no tópico anterior, além de ser bastante trabalhoso, tem o inconveniente de incorporar erros de arredondamentos no cálculo dos valores padronizados, podendo comprometer o resultado final. Neste contexto, sugerimos usar a seguinte fórmula alternativa para o cálculo de *r*, baseada nas observações originais.[1]

$$r = \frac{n.\Sigma(X.Y) - (\Sigma X).(\Sigma Y)}{\sqrt{n.\Sigma X^2 - (\Sigma X)^2} \cdot \sqrt{n.\Sigma Y^2 - (\Sigma Y)^2}}$$

Para obter os somatórios, procede-se da seguinte maneira.

$\Sigma(X.Y)$: fazem-se os produtos $x.y$, referente a cada par de observações e, depois, efetua-se a soma;

ΣX: somam-se os valores da variável X;

ΣY: somam-se os valores da variável Y;

ΣX^2: eleva-se ao quadrado cada valor de X e, depois, efetua-se a soma; e

ΣY^2: eleva-se ao quadrado cada valor de Y e, depois, efetua-se a soma.

---

[1] Pode-se provar matematicamente a equivalência das duas fórmulas para o cálculo de *r*.

Para ilustrar o uso da última expressão para o cálculo de r, vamos refazer o Exemplo 13.1. A Tabela 13.4 apresenta alguns cálculos intermediários.

Tabela 13.4 Cálculos intermediários para a obtenção de r.

	dados		cálculos intermediários		
	X	Y	X^2	Y^2	$X.Y$
	101	3,2	10201	10,24	323,2
	193	4,6	37249	21,16	887,8
	42	2,8	1764	7,84	117,6
	304	6,5	92416	42,25	1976,0
	42	2,0	1764	4,00	84,0
	152	1,9	23104	3,61	288,8
	55	2,9	3025	8,41	159,5
	105	5,3	11025	28,09	556,5
	68	2,7	4624	7,29	183,6
	219	3,1	47961	9,61	678,9
	129	3,1	16641	9,61	399,9
	42	1,2	1764	1,44	50,4
SOMA:	1452	39,3	251538	153,55	5706,2
Notação:	ΣX	ΣY	ΣX^2	ΣY^2	$\Sigma(X.Y)$

$$r = \frac{n.\Sigma(X.Y) - (\Sigma X)(\Sigma Y)}{\sqrt{n.\Sigma X^2 - (\Sigma X)^2} \cdot \sqrt{n.\Sigma Y^2 - (\Sigma Y)^2}}$$

Logo,

$$r = \frac{12.(5706,2) - 1452.(39,3)}{\sqrt{12.(251538) - (1452)^2} \cdot \sqrt{12.(153,55) - (39,3)^2}} =$$

$$= \frac{68474,4 - 57063,6}{\sqrt{3018456 - 2108304} \cdot \sqrt{1842,6 - 1544,49}} =$$

$$= \frac{11410,8}{\sqrt{910152} \cdot \sqrt{298,11}} = \frac{11410,8}{16472,0} = 0,69$$

Encontramos o mesmo resultado obtido no tópico anterior. E isto era de se esperar, pois as fórmulas são matematicamente equivalentes.

Teste de significância sobre r

Muitas vezes, temos o interesse em testar a existência de correlação entre duas variáveis, X e Y, a partir de uma amostra de observações pareadas (x, y). Nestes casos, além de mensurar o grau de correlação observado nos dados, queremos, também, testar as seguintes hipóteses, relativas à população em estudo.

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 o coeficiente de correlação com determinado sinal (+ ou -).

Restringindo-se à verificação de correlação linear e supondo que os dados de X e de Y provenham de distribuições normais, pode-se realizar o teste de correlação através da distribuição *t de Student com* $gl = n - 2$.[2] A Tabela VII do apêndice apresenta, para cada n, o valor mínimo de r para ser significativo, isto é, o valor absoluto mínimo de r para se rejeitar H_0.

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.

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 a *aptidão em matemática* e a *aptidão em música*.

---

[2] Para se verificar as suposições do teste de correlação, sugerimos construir: (1) um diagrama de pontos para os dados de cada variável para verificar se não existem fortes evidências de desvio da distribuição normal e (2) um diagrama de dispersão para verificar se os dados sugerem um relacionamento *não-linear*, em que não seria adequada a presente análise.

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, obteve-se o valor $r = 0{,}17$. Observando a Tabela VII do apêndice, verifica-se que, ao nível de significância usual de 5%, o valor mínimo de r para ser significativo 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 afirmar a existência de correlação positiva entre as duas variáveis em estudo.

A Tabela VII também pode ser usada para se ter uma avaliação da probabilidade de significância (valor p). No exemplo em questão, pode-se verificar que o valor encontrado ($r = 0{,}17$) é inferior a todos os valores tabelados para $n = 12$, ou seja, a probabilidade de significância é $p > 0{,}10$ (teste unilateral). Assim, mesmo que estivéssemos fazendo o teste ao nível de significância de $\alpha = 10\%$, o teste aceitaria H_0.

Uso do computador

A tabela a seguir é a saída do procedimento "correlação" do *Microsoft Excel*, com os dados da *percentagem de população urbana, taxa de crescimento demográfico, taxa de mortalidade infantil* e *taxa de alfabetização* da Tabela 13.1.[4]

---

[3] Observe que o problema sugere um teste unilateral ("correlação positiva" e não somente "existência de correlação"). Cabe observar, também, 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.

[4] Para acionar este procedimento, entre em "ferramentas", "análise de dados" e "correlação".

	%POP URB	CRESC	ALFAB	MORT
%POP URB	1,00			
CRESC	0,29	1,00		
ALFAB	0,34	0,40	1,00	
MORT	0,00	-0,59	-0,43	1,00

Observa-se que a saída do *Excel* fornece a correlação entre todos os pares das variáveis em questão. Usando pacotes computacionais mais especializados em estatística, o coeficiente de correlação costuma vir acompanhado do valor p associado ao teste estatístico bilateral. A seguir, é apresentada uma saída do *STATISTICA*[5].

	%POP URB	CRESC	ALFAB	MORT
%POP URB	1,00			
CRESC	0,29 p=0,363	1,00		
ALFAB	0,34 p=0,276	0,40 p=0,200	1,00	
MORT	0,00 p=0,999	-0,59 p=0,044	-0,43 p=0,168	1,00

Com estes resultados, concluímos que a única correlação significativa ao nível de significância de 5% é a correlação entre a *taxa de crescimento demográfico* e a *taxa de mortalidade infantil* (r = -0,59 com p = 0,044), indicando uma tendência moderada de quanto maior for a *taxa de crescimento demográfico* do município, menor deve ser a sua *taxa de mortalidade infantil.*[6]

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 usar a abordagem não-paramétrica, conforme discutido no

---

5 Ver *www.statcom.br*

6 Devemos lembrar que a existência de correlação não implica uma relação de causa-e-efeito. Provavelmente a presente correlação é causada pelas condições socioeconômicas dos municípios.

capítulo anterior. Nesta linha, um coeficiente muito usado é o coeficiente de correlação r_s *de Spearman*, que se utiliza apenas da ordenação dos valores.

A Tabela 13.5 apresenta os dados usados no Exemplo 13.3 e, para facilitar, já ordenados em relação à variável *aptidão em matemática*. Para cada variável, são atribuídos postos (*ranks*) da seguinte maneira: ao maior valor é atribuído o posto 1, ao segundo maior valor é atribuído o posto 2, e assim por diante. Quando ocorre algum empate, ou seja, quando se tem uma repetição de valor, considera-se que isto tenha acontecido por deficiência do instrumento de medida e atribuem-se postos seqüenciais e, em seguida, calcula-se a média dos postos com valores empatados. Por exemplo, na variável *aptidão em matemática*, tem-se para a criança 10 o valor 83 (o maior), logo, seu posto é 1. Em seguida vêm as crianças 6 e 9 com valores empatados em 75. Uma recebe posto 2 e a outra posto 3. Como o instrumento de medida não detecta qual está na frente, aloca-se posto 2,5 (média entre 2 e 3) para ambas. Em seguida, tem-se a criança 3, com valor 73, a qual recebe posto 4. E assim por diante.

Tabela 13.5 Alocação de postos para o cálculo de r_s *de Spearman*.

criança	aptidão em matemática (X)	posto em X	aptidão em música (Y)	posto em Y
10	83	1	82	4
6	75	2,5[1]	92	1
9	75	2,5[1]	54	12
3	73	4	70	7
8	72	5	88	2
11	62	6	64	9
1	60	7	80	5
2	58	8	62	10,5[2]
5	54	9	62	10,5[2]
12	52	10	69	8
4	51	11	83	3
7	48	12	79	6

Notas: [1] Média dos postos 2 e 3 referente ao valor empatado 75.
[2] Média dos postos 10 e 11 referente ao valor empatado 62.

Para se obter o coeficiente r_s, pode-se aplicar a fórmula de Pearson (seção anterior) sobre os postos de X e Y. Porém, com algumas simplificações, obtém-se a expressão a seguir:

$$r_s = 1 - \frac{6\sum D^2}{n(n^2-1)}$$

onde

D: diferença entre os postos das duas variáveis, calculado para dada elemento;

ΣD^2: soma dos quadrados dos valores de D; e

n: número de elementos observados (tamanho da amostra).

Tabela 13.6 Esquema de cálculo do coeficiente r_s *de Spearman*.

criança	posto em X	posto em Y	D	D^2
10	1	4	3	9
6	2,5	1	1,5	2,25
9	2,5	12	9,5	90,25
3	4	7	3	9
8	5	2	–3	9
11	6	9	3	9
1	7	5	–2	4
2	8	10,5	2,5	6,25
5	9	10,5	1,5	2.25
12	10	8	–2	4
4	11	3	–8	64
7	12	6	–6	36

Somando-se a última coluna, tem-se: $\sum D^2 = 245{,}25$. E o coeficiente r_s *de Spearman*:

$$r_s = 1 - \frac{6\sum D^2}{n(n^2-1)} = 1 - \frac{6\cdot(245{,}25)}{12\cdot(12^2-1)} = 1 - 0{,}86 = 0{,}14$$

indicando uma correlação positiva muito fraca nos dados observados.[7]

---

[7] 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.

A Tabela VIII do apêndice apresenta os valores absolutos mínimos de r_s para ser significativo (rejeitar a hipótese nula de ausência de correlação na população de onde foi extraída a amostra), em função do tamanho da amostra e do nível de significância α adotado. Verifica-se que, para $n = 12$ e nível de significância de 5%, o valor mínimo de r_s para ser significativo é de 0,503 (teste unilateral). Como o valor encontrado (r_s = 0,14) é menor que o valor tabelado, o teste não acusa significância.

Exercícios

5) Faça o cálculo do coeficiente *r* com os dados do Exemplo 13.3 e confira o resultado encontrado.

6) Considerando os dados da Tabela 13.1, calcule o coeficiente de correlação de Pearson entre as variáveis *taxa de alfabetização* e *taxa de mortalidade infantil*. Interprete o resultado obtido.

7) Considere os dados do Exercício 2.

a) Calcule a correlação entre a *nota no vestibular de matemática* e a *nota na disciplina de cálculo*.

b) Retire o valor discrepante detectado no Exercício 2b e calcule novamente o coeficiente *r*. Interprete.

c) Verifique se a correlação encontrada no item anterior é significativa. Faça o teste ao nível de significância de 5% e interprete o resultado.

8) Com respeito aos 23 alunos de uma turma de estatística, foram observadas as variáveis *número de faltas* e *nota final na disciplina*. Estes dados acusaram a seguinte correlação, descrita pelo coeficiente de correlação de Pearson: $r = -0,56$. Comente as seguintes frases relativas à turma em estudo e ao coeficiente obtido.

a) "Como $r = -0,56$ (correlação negativa moderada), nenhum aluno com grande número de faltas tirou nota alta".

b) "Como as duas variáveis são correlacionadas, bastaria usar uma delas como critério de avaliação, pois uma acarreta a outra."

c) "Os dados observados mostraram uma leve tendência de que a nota final se relaciona inversamente com o número de faltas, então, os alunos *freqüentadores* tiveram, em geral, melhor desempenho nas avaliações, do que os alunos que faltaram muito."

9) Numa amostra aleatória de $n = 212$ livros da Biblioteca Central da UFSC, encontramos $r = 0,207$ entre a *idade da edição* e o *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 passado. Estes 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 a determinada característica, também possuíam esta característica, porém, numa intensidade, em média, menor do que a média de seus pais.

Os estudos de Galton baseavam-se em observações empíricas. Em um destes trabalhos ele relacionou centenas de alturas de indivíduos, com as respectivas alturas médias de seus pais. O Exemplo 13.4 apresenta algumas destas observações.

Exemplo 13.4 Vamos considerar uma parte do problema que gerou o primeiro estudo de regressão, realizado por Galton, por volta de 1885. A Tabela 13.7 apresenta algumas observações coletadas por Galton.

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 estrutura de relacionamento entre X e Y, a Figura 13.9 ilustra dois modelos matemáticos para esta estrutura. A reta (A): $y = x$ indica que, *em média*, os filhos têm alturas iguais a altura média de seus pais, enquanto que a reta (B) representa

a hipótese de Galton, a qual afirma que *existe uma tendência de que filhos de pais altos tenham alturas inferiores às alturas médias de seus pais, enquanto os filhos de pais baixos tenham alturas superiores às alturas médias de seus pais.*

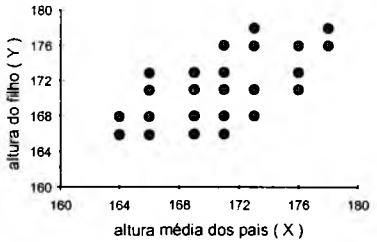


Figura 13.8 Diagrama de dispersão dos dados da Tabela 13.7.

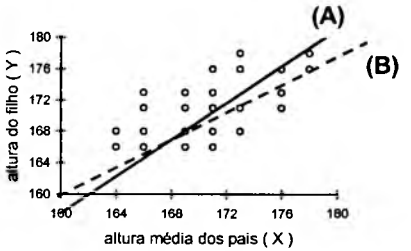


Figura 13.9 Ilustração de modelos matemáticos relacionando X e Y.

O Exemplo 13.4 se distingue dos exemplos anteriores por supor uma relação de causalidade entre X e Y, descrita em termos de uma relação matemática. É esta a diferença básica de um estudo de correlações e uma análise de regressão. A aplicação da análise de regressão é geralmente feita sob um referencial teórico, que justifique uma relação matemática de causalidade.

O modelo da regressão linear simples

O modelo estatístico-matemático de regressão, em sua formulação mais simples, relaciona uma variável Y, chamada de variável *resposta* ou *dependente*, com uma variável X, denominada de 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 (m^2)	→	preço do imóvel (r

Assim como num estudo de correlações, a análise de regressão também parte de 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 esta 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 $y = 1 + 2x$ representa uma reta, num par de eixos cartesianos. Para desenharmos esta reta basta atribuir dois valores para x e calcular os correspondentes valores de y. Digamos: $x = 0 \Rightarrow y = 1 + 2.(0) = 1$ e $x = 1 \Rightarrow y = 1 + 2.(1) = 3$. Com estes dois pontos, podemos traçar a reta da Figura 13.10.

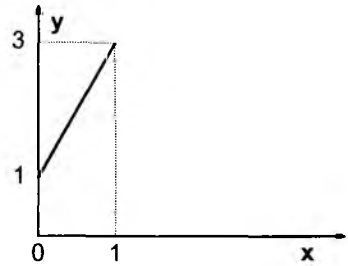


Figura 13.10 Representação gráfica da equação $y = 1 + 2x$.

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 mais adequado para um par de observações é

$$y = \alpha + \beta x + \varepsilon$$

onde ε representa o *efeito 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, representados no modelo por ε.

No modelo $y = \alpha + \beta x + \varepsilon$, chamaremos de *parte estrutural* a parcela de y determinada por x, isto é, $\alpha + \beta x$. E o procedimento inicial da análise de regressão é produzir uma estimativa para esta parte, a partir de uma amostra de observações (x, y).

Estimativas dos parâmetros α e β

A idéia básica da construção da parte estrutural do modelo, supostamente linear, é encontrar a reta que passe *mais próxima 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 13.11.

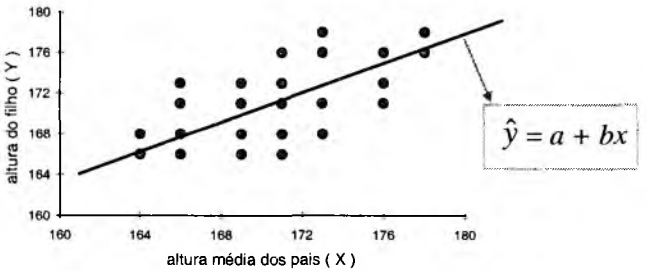


Figura 13.11 Representação da equação de regressão do Exemplo 13.4.

O chamado *método de mínimos quadrados* fornece as seguintes expressões para a equação de regressão.[8]

$$b = \frac{n.\Sigma(X.Y) - (\Sigma X).(\Sigma Y)}{n.\Sigma X^2 - (\Sigma X)^2} \quad \text{e} \quad a = \frac{\Sigma Y - b.\Sigma X}{n}$$

onde

n: número de pares (x, y) observados (tamanho da amostra);

$\Sigma(X.Y)$: somatório dos produtos $x.y$ (primeiramente fazem-se os produtos $x.y$, relativos a todos os pares observados e, depois, efetua-se a soma dos resultados destes produtos);

ΣX: soma dos valores observados da variável X;

ΣY: soma dos valores observados da variável Y; e

ΣX^2: soma dos quadrados dos valores de X (primeiro elevam-se os valores de X ao quadrado e, depois, efetua-se a soma).

---

8 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 que estamos apresentando. Veja, por exemplo, Wonnacott e Wonnacott (1991, p.287).

Exemplo 13.5 Ilustraremos a obtenção 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 obtenção da reta de regressão.

Dados		Cálculos intermediários	
X	Y	X^2	$X.Y$
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
ΣX = 1.539	ΣY = 1.540	ΣX^2 = 263.333	$\Sigma(X.Y)$ = 263.483

$$b = \frac{9.(263483) - (1539).(1540)}{9.(263333) - (1539)^2} = \frac{1287}{1476} = 0{,}872$$

$$a = \frac{1540 - (0{,}872).(1539)}{9} = 22{,}00$$

Donde 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.*[9] Veja a Figura 13.12.

---

[9] Por exemplo, para um dado valor $x = 164 \Rightarrow \hat{y} = 22 + (0{,}872).(164) = 165{,}0$ e para $x = 178 \Rightarrow \hat{y} = 22 + (0{,}872).(178) = 177{,}2$. Marcam-se os pontos (164; 165) e (178; 177,2) no plano formado pelos eixos X e Y e traça-se a reta que passa por estes dois pontos.

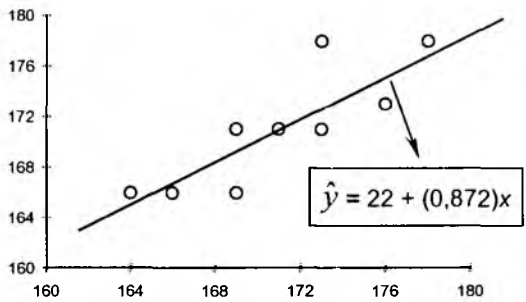


Figura 13.12 Diagrama de dispersão dos dados da Tabela 13.5 e a reta de regressão ajustada a estes dados.

Interpretação: Com respeito aos nove indivíduos observados, podemos predizer a altura de um filho ($\hat{y}$), a partir de uma dada altura média de seus pais, x, através da equação: $\hat{y} = 22 + (0{,}872)x$. Por exemplo, para uma altura média dos pais de $x = 175$ cm, temos uma estimativa para a altura do filho de $\hat{y} = 22 + (0{,}872).(175) = 174$ cm.

O coeficiente b, que no caso é 0,872, fornece uma estimativa da variação esperada de Y, a partir da variação de *uma* unidade em X. O sinal deste coeficiente indica o sentido do relacionamento. Como é positivo, indica uma correlação positiva entre as variáveis X e Y, para os nove indivíduos em estudo.[10]

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 a amostra observada. Vamos, então, desenvolver alguns procedimentos que permitem fazer este tipo de análise.

---

[10] 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 a partir dos dados observados e a reta teórica $y = x$ pode ser meramente casual. Para dar maior embasamento a esta discussão pode ser feito um teste estatístico sobre os parâmetros do modelo. Este tipo de teste estatístico pode ser estudado, por exemplo, em Chatterjee e Price (1977).

Para cada valor x observado (ou estabelecido), temos o correspondente valor observado da variável Y, representado por y, e o valor predito pelo modelo: $\hat{y} = a + bx$. Por exemplo, para o par observado ($x = 176$; $y = 173$), temos o próprio valor observado de Y ($y = 173$) e o valor predito pela equação de regressão: $\hat{y} = 22 + (0{,}872).(176) = 175{,}47$. A Figura 13.13 ilustra esta correspondência.

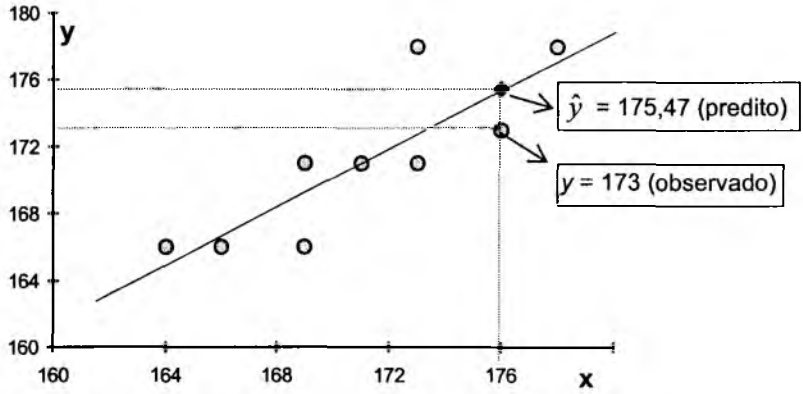


Figura 13.13 Valores observado e predito para $x = 176$.

Sendo $\bar{y}$ a média aritmética dos valores de Y e sendo $\hat{y}$ os valores preditos pela equação de regressão, vamos considerar os seguintes desvios:

a) $y - \bar{y}$ (desvios em relação à média dos valores de Y e, portanto, não leva em consideração a relação entre Y e X);

b) $y - \hat{y}$ (desvios em relação aos valores preditos pela equação de regressão – são os chamados *resíduos*, pois, mesmo levando em conta a relação entre Y e X, ainda não se tem uma predição exata dos valores observados devido ao efeito aleatório); e

c) $\hat{y} - \bar{y}$ (desvios dos valores preditos em relação à média dos valores de Y – é a diferença entre os dois desvios anteriores e corresponde à parcela do desvio total, $y - \bar{y}$, *explicada* pelo modelo de regressão). Veja a Figura 13.14.

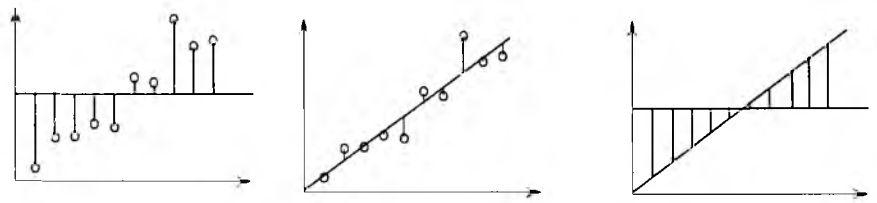


Figura 13.14 Ilustração dos desvios numa situação hipotética.

As somas dos quadrados dos desvios aqui considerados têm interpretações interessantes, conforme apontadas a seguir:

a) $\sum(y-\bar{y})^2$ (soma dos quadrados dos desvios de cada valor em relação à média) é uma medida da *variação total dos valores de Y.* [11]

b) $\sum(y-\hat{y})^2$ (soma quadrática dos resíduos) pode ser interpretada como uma medida da *variação não explicada pelo modelo de regressão* ou *variação residual* e

c) $\sum(\hat{y}-\bar{y})^2$ (soma dos quadrados dos desvios dos valores preditos em relação à média): é uma medida da parcela da *variação de Y explicada* pelo modelo de regressão. A Tabela 13.9 mostra o cálculo destas somas de quadrados.

Tabela 13.9 Obtenção dos valores preditos e cálculos das somas de quadrados dos desvios com os dados do Exemplo 13.5.

x	y	$\hat{y}$ =22 + (0,872)x	$(y-\bar{y})^2$	$(\hat{y}-\bar{y})^2$	$(y-\hat{y})^2$
164	166	165,01	26,11	37,11	0,98
166	166	166,75	26,11	19,01	0,56
169	171	169,37	0,01	3,03	2,66
169	166	169,37	26,11	3,03	11,36
171	171	171,11	0,01	0,00	0,01
173	171	172,86	0,01	3,06	3,46
173	178	172,86	47,47	3,06	26,42
176	173	175,47	3,57	19,01	6,10
178	178	177,22	47,47	37,33	0,61
		Soma:	177	125	52
$\bar{y}$ =	171,11	Notação:	$\sum(y-\bar{y})^2$	$\sum(\hat{y}-\bar{y})^2$	$\sum(y-\hat{y})^2$

[11] Note que $\sum(y-\bar{y})^2$ corresponde ao o numerador da fórmula da variância (Capítulo 6).

A Tabela 13.10 sintetiza os cálculos das somas de quadrados. Observe que a *variação total* corresponde à soma das *variações explicada e residual.*

Tabela 13.10 Decomposição da variação das variações de y.

Fonte de variação	Somas de quadrados
explicada por X, segundo o modelo (*variação explicada*)	$\sum(\hat{y}-\bar{y})^2 = 125$
variação residual ou *variação não explicada*	$\sum(y-\hat{y})^2 = 52$
variação total	$\sum(y-\bar{y})^2 = 177$

Chamaremos de *coeficiente de determinação* à seguinte razão:

$$R^2 = \frac{\sum(\hat{y}-\bar{y})^2}{\sum(y-\bar{y})^2} = \frac{\text{variação explicada}}{\text{variação total}}$$

O coeficiente de determinação é uma medida descritiva da proporção da variação de Y que pode ser explicada por X, segundo o modelo especificado. Em relação ao exemplo 13.5, temos:

$$R^2 = 125/177 \approx 0{,}70 \text{ (ou, } R^2 \approx 70\%)$$

Interpretação: Dentre os nove indivíduos estudados, as variações de suas alturas são explicadas, em parte, pela variação das alturas de seus pais (R^2 = 70% de explicação), e outra parte (1 - R^2 = 30%) devido a outros fatores.

Pode-se mostrar matematicamente que, no caso do modelo da regressão linear simples, o coeficiente de determinação R^2 coincide com o quadrado do coeficiente de correlação *r de Pearson*, estudado na Seção 13.2

Uso do computador

Exemplo 13.6 O anexo deste capítulo contém dados relativos a venda de 142 automóveis *seminovos*, incluindo o modelo, o preço de revenda (R$), o preço do modelo novo (R$), o tempo de uso do automóvel (anos completos) e a quilometragem (em km).

O preço de venda de um carro *seminovo* depende do preço deste modelo de carro 0 km. Assim, procura-se estabelecer um modelo de regressão entre o preço de revenda (Y) e o preço do correspondente modelo 0 km (X). Usando a planilha *Excel* (*ferramentas, análise- de dados, regressão*) obtivemos os seguintes resultados:

Estatística de regressão	
R múltiplo	0,889
R-Quadrado	0,791
R-quadrado ajustado	0,789
Erro padrão	1778,484
Observações	142

ANOVA

	gl	*SQ*	*MQ*	*F*	*F de significação*
Regressão	1	1,67E+09	1,67E+09	528,5782	2,22E-49
Resíduo	140	4,43E+08	3163004		
Total	141	2,11E+09			

	Coeficientes	Erro padrão	Stat t	valor-P	Inferior 95,0%	Superior 95,0%
Interseção	2654,11	431,22	6,155	7,46E-09	1801,56	3506,67
valor novo	0,476	0,021	22,991	2,22E-49	0,43	0,52

A primeira tabela de resultados mostra algumas estatísticas e, em particular, o R^2 (*R-quadrado*) igual a 0,791. Este resultado indica que na amostra observada, cerca de 79% da variação do preço de revenda pode ser "explicada" por uma relação linear com o preço do automóvel 0 km. Os demais 21% podem ser considerados como a variação provocada por outros fatores não considerados no modelo de regressão.

A segunda tabela apresenta a análise de variância (ANOVA) do modelo. A coluna SQ apresenta a soma de quadrados dos desvios, conforme discutido na Tabela 13.10. E, baseado nestas somas de quadrados, tem os resultados de um teste estatístico para as hipóteses

H_0: o coeficiente da variável independente X pode ser considerado nulo; e

H_1: o coeficiente da variável independente X é significativamente diferente de zero.

O teste, conhecido como *teste F da análise de variância d modelo*, resultou, no presente caso, na estatística $F = 528$, com correspon dente *valor p* = 2,22E-49 (ou seja, $p = 2{,}22$ com a vírgula 49 posições esquerda). Como o *valor p* é extremamente pequeno, o teste estatístic rejeita H_0, indicando que o valor do carro novo (X) é significativo par *explicar* o preço do carro *seminovo* (Y).

A terceira tabela fornece várias informações relevantes. primeira coluna apresenta as estimativas dos coeficientes, donde, n presente exemplo, temos a seguinte equação de predição para o preço d revenda (Y) em função do preço do automóvel novo (X):

$$\hat{y} = 2654{,}11 + (0{,}476)x$$

ou seja, tendo o preço do carro novo, x, pode-se obter uma previsão para preço de revenda, $\hat{y}$. Por exemplo, um modelo no qual o preço de novo R$16.000,00, seu preço de revenda, predito pelo modelo, é de

$$\hat{y} = 2654{,}11 + (0{,}476)(16000) = 10270$$

ou seja, R$ 10.270,00.

Com a equação de regressão, observa-se, também, que a cada real d diferença no carro novo, espera-se uma diferença de 0,476 reais na revenda.[12]

A última tabela também fornece os resultados de testes estatístico sobre cada um dos parâmetros do modelo. Em particular, na regressã simples, o teste sobre o parâmetro β (inclinação) é equivalente ao teste F d análise de variância sobre o modelo, discutido anteriormente. As dua últimas colunas desta tabela apresentam um intervalo de 95% de confianç para os dois parâmetros do modelo (o intercepto α e a inclinação β), com mesmo sentido dos intervalos de confiança discutidos no Capítulo 9.

Exercícios

10) Nos últimos anos, em várias regiões, houve um movimento migratório que fe crescer bastante a população urbana nos municípios médios e grandes. Nest contexto, vamos tentar *explicar* o crescimento demográfico de um município em função de sua população urbana, para os municípios da Tabela 13.1.

---

[12] É claro que um bom modelo para o preço de revenda deve levar em cont outros fatores, tais como a idade do veículo, estado de conservação, etc. Na Seçã 13.6 Usaremos um modelo mais elaborado.

a) Qual deve ser a variável dependente e a independente?
b) Estabeleça a equação de regressão.
c) Faça um gráfico com os pontos observados e a reta de regressão estimada.
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 *efeito aleatório*. Ou seja, estamos assumindo que X causa Y através de uma relação linear e toda a variação em torno desta relação deve-se ao efeito aleatório. Além disso, para a validade dos intervalos de confiança e testes estatísticos discutidos no Exemplo 13.6, torna-se 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 corrigir as distorções.

Um primeiro gráfico pode ser feito antes de se aplicar a análise de regressão. É o diagrama de dispersão, conforme discutido na Seção 13.1. Por este gráfico, pode-se 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, pode-se calcular os *resíduos*, através da diferença entre os valores observados y e os valores preditos $\hat{y}$, associados à cada x usado na análise. Ou seja, *resíduo* $= y - \hat{y}$. Um gráfico apresentando os pares (x, *resíduo*) é bastante útil na avaliação do modelo de regressão. Veja o gráfico à direita da Figura 13.15.

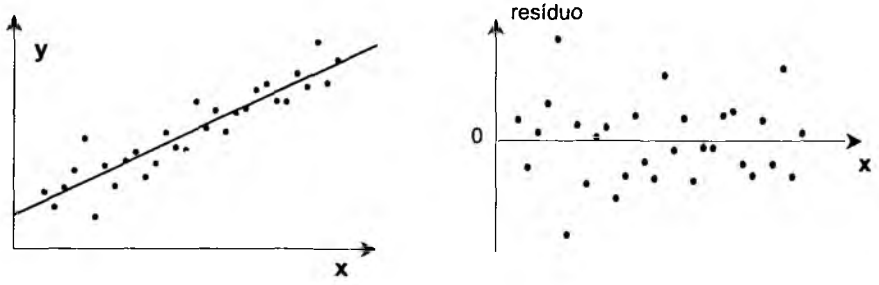


Figura 13.15 Gráficos indicando adequação do modelo.

Os gráficos da Figura 13.15 indicam uma situação onde as suposições do modelo estão aparentemente satisfeitas, pois os resíduos apresentam-se distribuídos de forma aleatória em torno da reta de regressão. No gráfico dos resíduos, a reta de regressão corresponde à linha horizontal sobre o valor zero. Já a Figura 13.16 apresenta uma situação onde existe um ponto discrepante. Este ponto é visível nos dois gráficos, mas no gráfico dos resíduos ele aparece mais nitidamente.

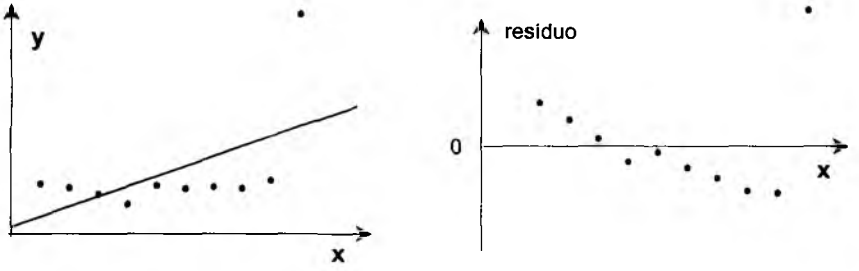


Figura 13.16 Gráficos indicando a presença de um valor discrepante.

A Figura 13.16 mostra como um ponto discrepante pode *forçar* uma inclinação na reta, sugerindo uma tendência não compatível com as demais observações. Este problema surge, principalmente, quando se tem uma amostra de observações pequena e o ponto discrepante estiver numa das extremidades do intervalo de observação de X. É prudente, neste caso, buscar a razão da existência deste ponto discrepante. Se a sua causa for algum erro, alguma falha no experimento ou, ainda, puder ser considerada como uma situação pouco provável, devemos efetuar nova análise sem esta 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, pode-se estabelecer diferentes tempos de cozimento e verificar o resultado Y. Nestes casos, recomenda-se variar X uniformemente sobre o intervalo de estudo. Por exemplo, se pretende fazer a análise entre 20 e 30 minutos de cozimento, pode-se fazer ensaios com os tempos de cozimentos de 20, 21, 22, ...,30 minutos.

Em estudos de levantamento, normalmente X e Y são observadas, donde torna-se 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, conseqüentemente, tendem a consumir pouco, provocando distribuições assimétricas para X e Y. Nesta situação, os dados devem se distribuir conforme mostra a Figura 13.17.

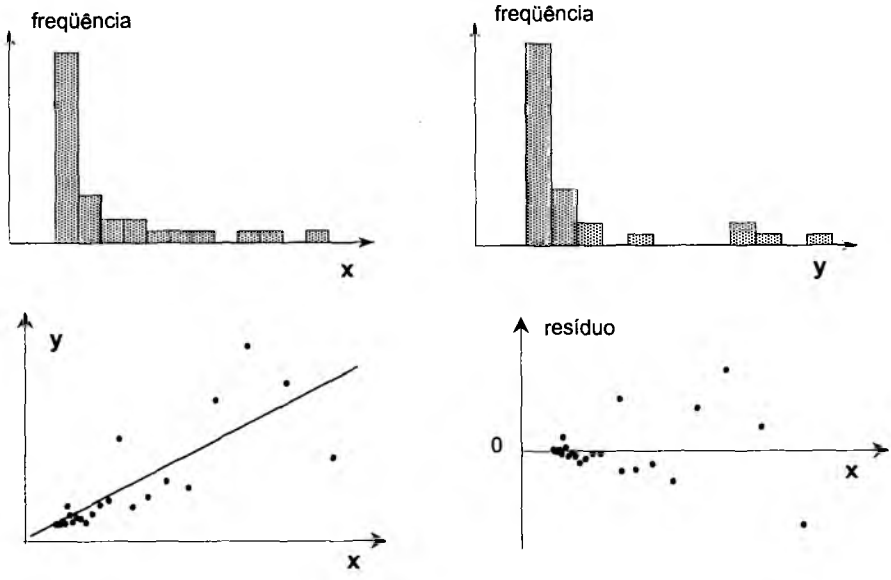


Figura 13.17 Gráficos indicando distribuições assimétricas de X e de Y e variância de Y aumentando proporcionalmente com X.

Em situações como indicado na Figura 13.17, os valores grandes de X vão ter mais peso na determinação da inclinação da reta. Neste caso,

recomenda-se a aplicação da transformação logarítmica tanto nos valores de X como nos valores de Y, estabelecendo o seguinte modelo: [13]

$$log(y) = \alpha + \beta\ log(x) + \varepsilon$$

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 aproximadamente simétricas. Com isto, tem-se uma situação mais adequada para estabelecer a reta de regressão. Em termos computacionais, deve-se:

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 três problemas para a aplicação de uma regressão linear: (1) uma relação *não-linear* para a parte estrutural do modelo; (2) uma redução da variância à medida que X aumenta; e (3) maior número de observações para níveis pequenos de X. É uma situação típica onde se recomenda uma transformação logarítmica (ou raiz quadrada) somente nos valores da variável X, ou seja, passa-se a considerar o seguinte modelo para os dados:

$$y = \alpha + \beta log(x) + \varepsilon$$

Note que este modelo pode ser considerado linear em termos das variáveis $log(x)$ e y (não mais entre x e y). Em termos computacionais, deve-se:

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 neste novo modelo.

---

[13] 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. Observa-se que estas transformações são possíveis somente quando todos os valores são positivos.

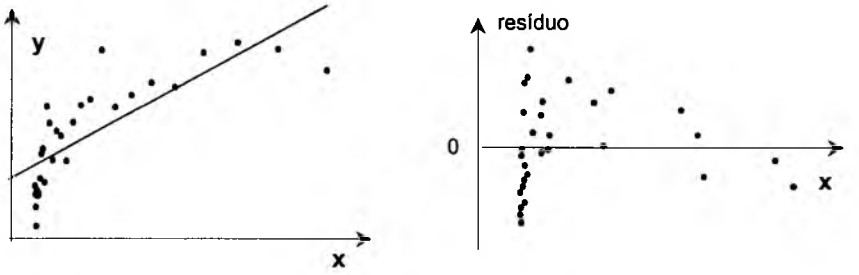


Figura 13.18 Gráficos indicando uma relação *não-linear* – aparentemente logarítmica – e variância não constante.

A Figura 13.19 apresenta uma situação com problemas análogos ao caso anterior, mais especificamente, apresenta os seguintes problemas: (1) uma relação *não-linear* para a parte estrutural do modelo; (2) um aumento da variância à medida que X aumenta; e (3) uma concentração maior de valores grandes de X. Em casos como este, recomenda-se uma transformação logarítmica nos valores da variável Y, ajustando o seguinte modelo aos dados:

$$log\,(y) = \alpha + \beta x + \varepsilon$$

Ou seja,

a) calcula-se o logaritmo de cada valor y;

b) aplica-se a análise de regressão linear sobre os dados $(x, log(y))$; e

c) constrói-se novamente o gráfico de resíduos para verificar se o novo modelo é mais adequado aos dados.

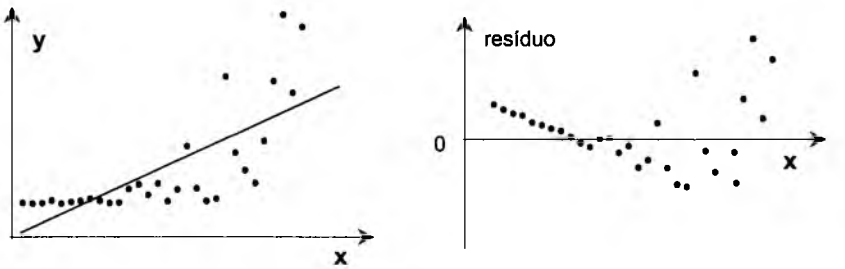


Figura 13.19 Gráficos indicando uma relação *não-linear* – aparentemente exponencial – e variância não constante.

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, pois 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 se ouvir falar em aumentos percentuais. 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, estes ganhos tornam-se iguais. Por esta razão, é muito comum usar a escala (ou transformação) logarítmica em variáveis econômicas ou medidas de tamanho em geral.

Exemplo 13.6 (continuação) Na seção anterior, realizou-se uma regressão entre o preço de revenda de carros *seminovos* (Y) e o preço do correspondente modelo 0 km (X), considerando uma amostra de 142 automóveis apresentada no anexo deste capítulo. A Figura 13.20 apresenta o diagrama de dispersão e o gráfico dos resíduos deste modelo, obtidos pela planilha *Excel*.

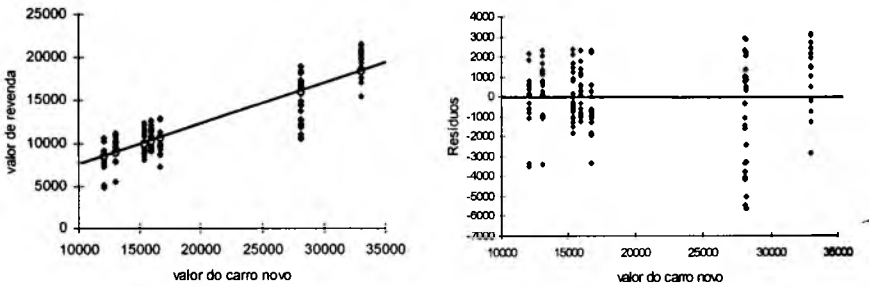


Figura 13.20 Gráfico de dispersão com o ajuste da reta de regressão e gráfico dos resíduos.

Observa-se na Figura 13.20 que X só assume alguns determinados valores. Isto porque os automóveis em estudo são de 7 modelos e, para cada modelo, o preço 0 km é único. Por outro lado, não parece haver fortes violações nas suposições do modelo de regressão, a não ser a ocorrência maior de valores pequenos com respeito às duas variáveis, o que sugere tentarmos uma transformação logarítmica em X e em Y.

Realizamos a transformação logarítmica nos valores das duas variáveis e refizemos a análise de regressão, contudo o R^2 reduziu e o gráfico dos resíduos apontou uma distribuição assimétrica, com cauda mais longa à esquerda. Em função destes resultados, preferimos manter o modelo original. Na verdade, o preço de um carro *seminovo* depende de vários outros fatores, levando a um modelo de regressão múltipla, o qual discutiremos na próxima seção.

13.6 INTRODUÇÃO À REGRESSÃO MÚLTIPLA

Em geral, ao considerarmos uma variável dependente Y, esta costuma depender de várias variáveis independentes (X_1, X_2, ..., X_k). Na análise de regressão múltipla, procura-se construir um modelo estatístico-matemático para se estudar objetivamente a relação entre as variáveis independentes e a variável dependente e, a partir do modelo, conhecer a influência de cada variável independente, como também, 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 = renda (R$) X_2 = poupança (R$) X_3 = taxa de juros (%)	→	Y = consumo (R$)
X_1 = memória RAM (Gb) X_2 = sistema operacional X_3 = tipo de processador	→	Y = tempo de resposta do sistema computacional (segundos)
X_1 = área construída do imóvel (m^2) X_2 = padrão de qualidade (custo do m^2, R$) X_3 = localização	→	Y = preço de um imóvel novo (R$)
X_1 = valor do modelo novo (R$) X_2 = quilometragem (km) X_3 = idade do veículo (anos) X_4 = estado de conservação X_5 = opcionais	→	Y = valor de revenda de carro *seminovo* (R$)

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_4 = *estado de conservação do veículo* pode ter valor 1 quando este for considerado "bom" e 0 quando for considerado "ruim". Também será considerado que Y é uma variável aleatória, isto é, somente será conhecida após a observação do elemento (indivíduo, carro, 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, em parte, dos correspondentes valores x_1, x_2,..., x_k e de uma infinidade de outros fatores, representados por ε. 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 e β_1, β_2, ... ,β_k são parâmetros a serem estimados com os dados e ε representa o *efeito aleatório*. As demais suposições são análogas à regressão simples, acrescentando a suposição de que as variáveis independentes X_1, X_2, ..., X_k não devem ter correlações altas entre si.

Exemplo 13.7 Considerando os dados de 142 automóveis (anexo), vamos construir um modelo de regressão para tentar *explicar Y = preço de revenda de automóveis seminovos* (em R$), em função de:

X_1 = preço do correspondente modelo 0 km (em R$);

X_2 = tempo de uso (em anos completos); e

X_3 = quilometragem (em milhares de km).

Usando a planilha *Excel* (*ferramentas, análise de dados, regressão*), obtivemos os seguintes resultados:

Estatística de regressão	
R múltiplo	0,961
R-Quadrado	0,923
R-quadrado ajustado	0,921
Erro padrão	1087
Observações	142

ANOVA

	gl	*SQ*	*MQ*	*F*	*F de significação*
Regressão	3	1,95E+09	6,51E+08	550,27	1,52E-76
Resíduo	138	1,63E+08	1182186		
Total	141	2,11E+09			

	Coeficientes	Erro padrão	Stat t	valor-P	Inferior 95,0%	Superior 95,0%
Interseção	6240,13	352,11	17,722	2,25E-37	5543,89	6936,36
valor novo	0,48	0,01	37,448	3,61E-74	0,45	0,50
tempo uso	-432,92	136,64	-3,168	0,0019	-703,10	-162,75
quilometragem	-45,11	9,00	-5,014	1,61E-06	-62,90	-27,32

Observamos, na primeira tabela, o valor de R^2 (*R-quadrado*) igual a 0,923. Este resultado indica que na amostra observada, cerca de 92% da variação do preço de revenda pode ser "explicada" por uma relação linear que envolve o preço do automóvel 0 km (X_1), tempo de uso (X_2) e a quilometragem (X_3). Um resultado expressivamente maior do que os 71% obtido no Exemplo 13.6, quando se considerou apenas X_1 como variável independente.[14]

A segunda tabela (ANOVA) fornece o resultado estatístico da seguinte hipótese nula:

$$H_0: \beta_1 = \beta_2 = ... = \beta_k = 0$$

relativa ao modelo

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$$

ou seja, por esta hipótese, o conjunto de variáveis independentes em estudo não tem poder de *explicação* sobre a variável dependente.[15] Este teste, conhecido como *teste F da análise de variância do modelo*, resultou na estatística F = 550,27, com correspondente *valor p* = 1,52E-76 (ou seja, *p* corresponde a 1,52 com a vírgula 76 posições à esquerda). Como o *valor p* é extremamente pequeno, o teste estatístico rejeita H_0, indicando que as variáveis independentes escolhidas são significativas para *explicar Y*.

---

[14] O cálculo do R^2 na regressão múltipla é equivalente ao da regressão simples.

[15] Cabe observar que o teste estatístico refere-se à população, ou seja, quando se tem uma amostra muito pequena, pode-se obter um valor alto de R^2 e o teste aceitar H_0: $\beta_1 = \beta_2 = ... = \beta_k = 0$.

A terceira tabela fornece as estimativas dos coeficientes, incluindo intervalos de confiança e testes estatísticos para cada particular coeficiente. A primeira coluna apresenta as estimativas dos coeficientes, donde, no presente exemplo, temos a seguinte equação de predição para o preço de revenda (Y) em função do preço do automóvel 0 km (X_1), do tempo de uso (X_2) e da quilometragem (X_3):

$$\hat{y} = 6240 + 0{,}48x_1 - 433x_2 - 45{,}1x_3$$

Assim, tendo o preço do carro novo (x_1), o tempo de uso (x_2) e a quilometragem (x_3) de um carro pode-se obter uma predição para o seu preço de revenda, $\hat{y}$. Por exemplo, um modelo, cujo preço do carro novo é R$16.000,00, que tenha 2 anos de uso e 50 mil quilômetros rodados, seu preço de revenda, predito pelo modelo, é de

$$\hat{y} = 6240 + (0{,}48)(16000) - (433)(2) - (45{,}1)(50) = 10779$$

ou seja, R$ 10.779,00.

Com a equação de regressão, observa-se, também, que a cada real de diferença no carro novo, espera-se uma diferença de 48 centavos de reais na revenda (mantendo-se constantes o tempo de uso e a quilometragem). A cada ano de envelhecimento do automóvel, espera-se R$433,00 a menos na revenda (mantendo-se constantes o valor do carro novo e a quilometragem). E, também, a cada mil quilômetros rodados, espera-se R$45,11 a menos na revenda (mantendo-se constantes o valor de novo e o tempo de uso).[16]

A última tabela também fornece os resultados de testes estatísticos individuais, relativos a cada um dos coeficientes da equação de regressão. Ou seja, tem-se os resultados dos quatro seguintes testes:

---

16 Dois comentários são pertinentes no momento:

a) É sabido que a desvalorização do automóvel não é linear com o tempo de uso. Uma transformação logarítmica em Y deve tornar o modelo mais realista.

b) As variáveis independentes, nesta aplicação, são correlacionadas. Por exemplo, um automóvel mais velho deve ter maior quilometragem. Logo, a interpretação "mantendo as demais variáveis constantes" fica prejudicada. Além disso, os valores dos coeficientes de variáveis independentes correlacionadas não são bem estimados (observe a magnitude dos intervalos de confiança nas duas últimas colunas da terceira tabela).

1) H_0: $\alpha = 0$;
2) H_0: $\beta_1 = 0$;
3) H_0: $\beta_2 = 0$; e
4) H_0: $\beta_3 = 0$.

Como em todos os quatro casos, os *valores p* foram inferiores ao nível de significância usual de 0,05, rejeitam-se as quatro hipóteses nulas, concluindo que nenhuma das variáveis independentes pode ser excluída do modelo.

Assim como na regressão simples, podem-se calcular os *resíduos* para verificar a adequação do modelo de regressão. Calculam-se, inicialmente, os valores preditos, $\hat{y}$, associados a cada conjunto de valores (x_1, x_2, ..., x_k) usado na análise. No exemplo dos automóveis, os valores preditos seriam calculados pela expressão $\hat{y} = 6240 + 0{,}48x_1 - 433x_2 - 45{,}1x_3$, com x_1, x_2 e x_3 associados a cada um dos 142 automóveis avaliados. Os resíduos são obtidos através da diferença entre os valores observados e os valores preditos: $resíduo = y - \hat{y}$.

Os resíduos podem ser apresentados num diagrama de dispersão com cada variável independente ou com os valores preditos, os quais correspondem a uma combinação das variáveis independentes. A Figura 13.21 apresenta o diagrama de dispersão dos pares ordenados (*predito*, *resíduo*), construído com apoio do *STATISTICA*.[17]

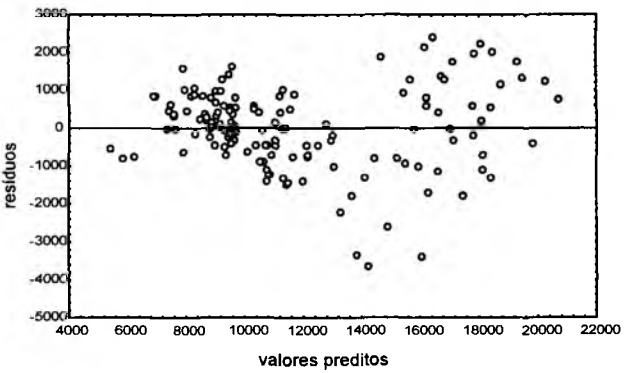


Figura 13.21 Gráfico dos resíduos com os valores preditos.

[17] www.statsoft.com.br

A análise do gráfico de resíduos (Figura 13.21) mostra um certo padrão. Para valores preditos pequenos, os resíduos tendem a ser positivos, depois eles tendem a ser negativos e, para valores preditos grandes, eles tendem a ser positivos de novo. Além disso, observa-se que a dispersão aumenta para os valores preditos maiores. Conforme visto na seção anterior, estas características sugerem a aplicação de uma transformação logarítmica na variável dependente.

Raciocinando em termos da relação entre tempo de uso (X_2) e o valor do automóvel (Y), é mais natural considerar que a cada ano de uso, o automóvel tenha uma redução percentual do seu valor, reforçando a transformação sugerida pelo gráfico dos resíduos. Contudo, a construção de um modelo mais adequado para estes dados é deixada para o leitor (Exercício 17).

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 nesta amostra, obtendo-se $r = -0{,}20$. Pode-se concluir que realmente existe correlação entre X e Y, ao nível de significância de 5%? Faça um teste estatístico apropriado.

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 pequena 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 em três níveis de peso e altura, apresentada na tabela abaixo:

	altura		
peso	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 que 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 freqüentaram (e tiveram aprovação) em escolas regulares (Y) e o número médio de anos que os seus pais freqüentaram (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 freqüentaram escolas regulares (Y) e o número médio de anos que os seus pais freqüentaram 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 ^{0}C (X), e o *volume de vendas diária de sorvete*, em kg (Y). Com os dados, estabeleceu uma equação de regressão, resultando em:

$$y = 0{,}5 + 1{,}8x, \quad \text{com } R^2 = 0{,}80$$

Pergunta-se:

a) Qual o consumo esperado de sorvete num dia de 27^{0}C?

b) Qual o incremento esperado nas vendas de sorvete a cada 1^{0}C de aumento da temperatura?

16) A tabela a seguir relaciona os pesos (em centenas de kg) e as taxas de consumo 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
consumo	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 consumo, na amostra observada?

c) Para estabelecer uma equação de regressão, qual deve ser a variável dependente e qual deve ser a variável independente? Justifique a sua resposta.

d) Estabeleça a equação de regressão, considerando a resposta do item (c).

e) Apresente o diagrama de dispersão e a reta de regressão obtida em (d).

f) Você considera adequado o ajuste do modelo de regressão do item (d)? Dê uma medida desta adequação interpretando-a.

g) Qual o consumo esperado para um carro de 2000 kg? Justifique sua resposta. Lembrete: os dados de consumo na tabela estão em *centenas* de kg.

h) Você considera seu estudo capaz de predizer o consumo esperado de um veículo com peso de 7000 kg? Justifique sua resposta.

17) Com o auxílio de um computador, refaça o Exemplo 13.7, mas considerando como variável dependente o $log(Y)$, onde Y = valor de revenda do automóvel. Observe o gráfico dos resíduos. Exclua três observações que aparecem como discrepantes. Refaça novamente a análise.

ANEXO

Os dados que seguem foram coletados pelo Prof. Manoel R. Lino (INE / CTC / UFSC) e fornecem informações sobre a venda de 142 automóveis *seminovos*, incluindo o modelo, o preço de revenda (R$), o preço do modelo novo (R$), o tempo de uso do automóvel (anos completos) e a quilometragem (em km).

Auto	modelo	preço de rev.	preço novo	tempo de uso	km	Auto	modelo	preço de rev.	preço novo	tempo de uso	km
1	Mille	4890	12081	5	98	72	Gol	10340	15945	3	39
2	Mille	5064	12081	5	88	73	Gol	9680	15945	3	39
3	Mille	7820	12081	4	73	74	Gol	11640	15945	2	39
4	Mille	7320	12081	4	65	75	Gol	11350	15945	2	36
5	Mille	8100	12081	4	62	76	Gol	11380	15945	2	36
6	Mille	7590	12081	4	59	77	Gol	12050	15945	2	32
7	Mille	8950	12081	3	61	78	Gol	11430	15945	2	18
8	Mille	8590	12081	3	42	79	Gol	12570	15945	1	38
9	Mille	8530	12081	3	38	80	Gol	12040	15945	1	20
10	Mille	9040	12081	2	50	81	Gol	12580	15945	1	11
11	Mille	8790	12081	2	41	82	Fiorino	7270	16711	5	92
12	Mille	9200	12081	2	38	83	Fiorino	8790	16711	5	72
13	Mille	10240	12081	1	19	84	Fiorino	9510	16711	4	75
14	Mille	10560	12081	1	12	85	Fiorino	8659	16711	4	69
15	Fiesta	5500	13050	5	90	86	Fiorino	9660	16711	4	66
16	Fiesta	7780	13050	5	75	87	Fiorino	9870	16711	3	57
17	Fiesta	7850	13050	5	64	88	Fiorino	9749	16711	3	50
18	Fiesta	7900	13050	5	60	89	Fiorino	9340	16711	3	48
19	Fiesta	7980	13050	5	60	90	Fiorino	9643	16711	3	45
20	Fiesta	9450	13050	4	63	91	Fiorino	11230	16711	2	46
21	Fiesta	9040	13050	4	56	92	Fiorino	9970	16711	2	42
22	Fiesta	8900	13050	4	45	93	Fiorino	10900	16711	2	37
23	Fiesta	8970	13050	4	45	94	Fiorino	10589	16711	2	30
24	Fiesta	9990	13050	3	48	95	Fiorino	12910	16711	1	22
25	Fiesta	10150	13050	3	44	96	Fiorino	12830	16711	1	17
26	Fiesta	9150	13050	3	36	97	Parati	12000	28137	5	99
27	Fiesta	10200	13050	3	33	98	Parati	11880	28137	5	85
28	Fiesta	10530	13050	2	52	99	Parati	10590	28137	4	82
29	Fiesta	10900	13050	2	47	100	Parati	12280	28137	4	67
30	Fiesta	11200	13050	2	45	101	Parati	14410	28137	4	60

continua ...

Auto	modelo	preço de rev.	preço novo	tempo de uso	km	Auto	modelo	preço de rev.	preço novo	tempo de uso	km
31	Fiesta	9680	13050	2	23	102	Parati	14580	28137	4	54
32	Fiesta	10200	13050	2	15	103	Parati	15750	28137	4	48
33	Fiesta	9580	13050	1	27	104	Parati	14960	28137	3	53
34	Fiesta	9980	13050	1	16	105	Parati	18340	28137	3	48
35	Fiesta	10050	13050	1	12	106	Parati	14580	28137	3	46
36	Corsa	8450	15337	5	75	107	Parati	17020	28137	3	39
37	Corsa	8120	15337	5	69	108	Parati	12680	28137	2	60
38	Corsa	8680	15337	5	65	109	Parati	17020	28137	2	39
39	Corsa	8900	15337	5	56	110	Parati	16800	28137	2	37
40	Corsa	9200	15337	4	78	111	Parati	16800	28137	2	37
41	Corsa	8960	15337	4	65	112	Parati	15680	28137	2	29
42	Corsa	9350	15337	4	62	113	Parati	18360	28137	1	32
43	Corsa	9180	15337	4	59	114	Parati	18960	28137	1	18
44	Corsa	9260	15337	4	59	115	Parati	17090	28137	1	18
45	Corsa	9250	15337	4	56	116	Escort	11050	28168	5	94
46	Corsa	9680	15337	3	75	117	Escort	10480	28168	5	82
47	Corsa	10100	15337	3	60	118	Escort	13650	28168	5	68
48	Corsa	9950	15337	3	59	119	Escort	12800	28168	4	85
49	Corsa	9580	15337	3	57	120	Escort	16570	28168	4	72
50	Corsa	9460	15337	3	48	121	Escort	16400	28168	4	55
51	Corsa	10900	15337	2	49	122	Escort	16950	28168	3	60
52	Corsa	11200	15337	2	36	123	Escort	16860	28168	3	47
53	Corsa	10750	15337	2	36	124	Escort	17050	28168	3	47
54	Corsa	12050	15337	2	33	125	Escort	18120	28168	2	44
55	Corsa	12350	15337	1	40	126	Escort	18900	28168	2	37
56	Corsa	11640	15337	1	23	127	Escort	18280	28168	1	26
57	Corsa	11400	15337	1	22	128	Escort	17400	28168	1	25
58	Gol	9200	15945	5	75	129	Vectra	18830	32995	5	75
59	Gol	9340	15945	5	75	130	Vectra	18120	32995	5	68
60	Gol	9000	15945	5	68	131	Vectra	15490	32995	4	80
61	Gol	9340	15945	5	45	132	Vectra	17600	32995	4	54
62	Gol	9450	15945	4	78	133	Vectra	17050	32995	4	47
63	Gol	9680	15945	4	69	134	Vectra	19800	32995	3	63
64	Gol	9920	15945	4	62	135	Vectra	20300	32995	3	58
65	Gol	9320	15945	4	59	136	Vectra	20500	32995	3	49
66	Gol	9950	15945	4	58	137	Vectra	19880	32995	3	43
67	Gol	9680	15945	4	55	138	Vectra	21050	32995	2	40
68	Gol	10500	15945	3	63	139	Vectra	20810	32995	2	36
69	Gol	10860	15945	3	50	140	Vectra	19400	32995	2	29
70	Gol	10780	15945	3	50	141	Vectra	21500	32995	1	28
71	Gol	10560	15945	3	43	142	Vectra	21440	32995	1	19

REFERÊNCIAS BIBLIOGRÁFICAS

AGRESTI, A. *Analysis of ordinal categorical data*. USA: John-Wiley, 1984.

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*. 4 ed. Coleção Métodos Quantitativos. São Paulo: Editora Atual, 1987.

CHATTERJEE, S., PRICE, B. *Regression analysis by examples*. USA: John Wiley, 1977.

COCHRAN, W. G. *Sampling techniques*. 3 ed. USA: John Wiley, 1977.

COCHRAN, W. G., COX, G. M. *Experimental designs*. 2 ed. New York: John Wiley, 1957.

FISHER, R. A. *The design of experiments*. 6 ed. Londres, 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: Editora Harbra, 1985.

LEVINE, D. M., BERENSON, M. L., STEPHAN, D. *Estatística: teoria e aplicações usando o Excel*. Rio de Janeiro: LTC, 2000

MENDENHALL, N. *Probabilidade e estatística*, v. 1 e 2. Rio de Janeiro: Editora Campos, 1985.

NOETHER, G. F. *Introdução à estatística. Uma abordagem não-paramétrica*. 2 ed. Rio de Janeiro: Editora Guanabara Dois, 1983.

SELLTIZ, WRIGHTSMAN, COOK *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 mensurement of uncertainty before 1900*. USA, Harward, 1986.

STEVENSON, W. J. *Estatística aplicada à administração*. São Paulo: Editora 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*. Rio de Janeiro: LTC, 1999.

WONNACOTT, T. H., WONNACOTT, R. J. *Estatística aplicada à economia e à administração*. Rio de Janeiro: Livros Técnicos e Científicos, 1981.

APÊNDICE

TABELA I Números aleatórios.

98 08 62 48 26	45 24 02 84 04	44 99 90 88 96	39 09 47 34 07	35 44 13 18 80
33 18 51 62 32	41 94 15 09 49	89 43 54 85 81	88 69 54 19 94	37 54 87 30 43
80 95 10 04 06	96 38 27 07 74	20 15 12 33 87	25 01 62 52 98	94 62 46 11 71
79 75 24 91 40	71 96 12 82 96	69 86 10 25 91	74 85 22 05 39	00 38 75 95 79
18 63 33 25 37	98 14 50 65 71	31 01 02 46 74	05 45 56 14 27	77 93 89 19 36
74 02 94 39 02	77 55 73 22 70	97 79 01 71 19	52 52 75 80 21	80 81 45 17 48
54 17 84 56 11	80 99 33 71 43	05 33 51 29 69	56 12 71 92 55	36 04 09 03 24
11 66 44 98 83	52 07 98 48 27	59 38 17 15 39	09 97 33 34 40	88 46 12 33 56
48 32 47 79 28	31 24 96 47 10	02 29 53 68 70	32 30 75 75 46	15 02 00 99 94
69 07 49 41 38	87 63 79 19 76	35 58 40 44 01	10 51 82 16 15	01 84 87 69 38
09 18 82 00 97	32 82 53 95 27	04 22 08 63 04	83 38 98 73 74	64 27 85 80 44
90 04 58 54 97	51 98 15 06 54	98 93 88 19 97	91 87 07 61 50	68 47 66 46 59
73 18 95 02 07	47 67 72 52 69	62 29 06 44 64	27 12 46 70 18	41 36 18 27 60
75 76 89 64 90	20 97 18 17 49	90 42 91 22 72	95 37 50 58 71	93 82 34 31 78
54 01 64 40 56	66 28 13 10 03	00 68 22 73 98	20 71 45 32 95	07 70 61 78 13
08 35 86 99 10	78 54 24 27 85	13 66 15 88 73	04 61 89 75 53	21 22 30 84 20
28 30 60 32 64	81 33 31 05 91	40 51 00 78 93	32 60 46 04 75	94 11 90 18 40
53 84 08 62 33	81 59 41 36 28	51 21 59 02 90	28 46 66 87 95	77 76 22 07 91
91 75 75 37 41	61 61 36 22 69	50 26 39 02 12	55 78 17 65 14	83 48 34 70 55
89 41 59 26 94	00 39 75 83 91	12 60 71 76 46	48 94 97 23 06	94 54 13 74 08
77 51 30 38 20	86 83 42 99 01	68 41 48 27 74	51 90 81 39 80	72 89 35 55 07
19 50 23 71 74	69 97 92 02 88	55 21 02 97 73	74 28 77 52 51	65 34 46 74 15
21 81 85 93 13	93 27 88 17 57	05 68 67 31 56	07 08 28 50 46	31 85 33 84 52
51 47 46 64 99	68 10 72 36 21	94 04 99 13 45	42 83 60 91 91	08 00 74 54 49
99 55 96 83 31	62 53 52 41 70	69 77 71 28 30	74 81 97 81 42	43 86 07 28 34
33 71 34 80 07	93 58 47 28 69	51 92 66 47 21	58 30 32 98 22	93 17 49 39 72
85 27 48 68 93	11 30 32 92 70	28 83 43 41 37	73 51 59 04 00	71 14 84 36 43
84 13 38 96 40	44 03 55 21 66	73 85 27 00 91	61 22 26 05 61	62 32 71 84 23
56 73 21 62 34	17 39 59 61 31	10 12 39 16 22	85 49 65 75 60	81 60 41 88 80
65 13 85 68 06	87 64 88 52 61	34 31 36 58 61	45 87 52 10 69	85 64 44 72 77
38 00 10 21 76	81 71 91 17 11	71 60 29 29 37	74 21 96 40 49	65 58 44 96 98
37 40 29 63 97	01 30 47 75 86	56 27 11 00 86	47 32 46 26 05	40 03 03 74 38
97 12 54 03 48	87 08 33 14 17	21 81 53 92 50	75 23 76 20 47	15 50 12 95 78
21 82 64 11 34	47 14 33 40 72	64 63 88 59 02	49 13 90 64 41	03 85 65 45 52
73 13 54 27 42	95 71 90 90 35	85 79 47 42 96	08 78 98 81 56	64 69 11 92 02
07 63 87 79 29	03 06 11 80 72	96 20 74 41 56	23 82 19 95 38	04 71 36 69 94
60 52 88 34 41	07 95 41 98 14	59 17 52 06 95	05 53 35 21 39	61 21 20 64 55
83 59 63 56 55	06 95 89 29 83	05 12 80 97 19	77 43 35 37 83	92 30 15 04 98
10 85 06 27 46	99 59 91 05 07	13 49 90 63 19	53 07 57 18 39	06 41 01 93 62
39 82 09 89 52	43 62 26 31 47	64 42 18 08 14	43 80 00 93 51	31 02 47 31 67

Fonte: Blalock (1960).

TABELA II Distribuição binomial: probabilidade de cada valor x em função de n e π.

		π									
n	x	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

continua ...

Tabela II (continuação)

		π								
n	x	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

continua ...

Tabela II (continuação)

		π									
n	x	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	0,0006
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

continua ...

Tabela II (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	0,0063	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	0,3028	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

continua ...

Tabela II (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	0,1488	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

continua ...

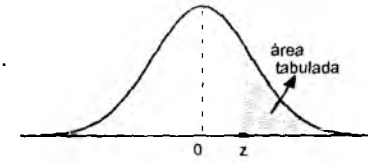
Tabela II (continuação)

		π								
n	*x*	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 III Coeficientes binomiais.

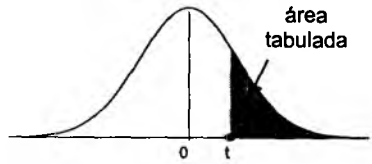
n	$\binom{n}{0}$	$\binom{n}{1}$	$\binom{n}{2}$	$\binom{n}{3}$	$\binom{n}{4}$	$\binom{n}{5}$	$\binom{n}{6}$	$\binom{n}{7}$	$\binom{n}{8}$	$\binom{n}{9}$	$\binom{n}{10}$
0	1										
1	1	1									
2	1	2	1								
3	1	3	3	1							
4	1	4	6	4	1						
5	1	5	10	10	5	1					
6	1	6	15	20	15	6	1				
7	1	7	21	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 IV Distribuição normal padrão.



	segunda decimal de z									
z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
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									

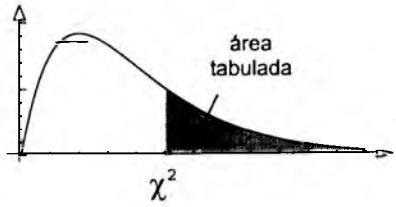
TABELA V Distribuição *t* de *Student*



	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 VI 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 VII 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 a partir da estatística $t = r.(n-2)/(1-r)$ 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 VIII 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 a partir da estatística $z = r_s \sqrt{(n-1)}$, 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 a preferências dos eleitores quanto aos candidatos, sem que o entrevistado interfira no processo, ou seja, procura-se levantar os dados naturalmente, com eles se apresentam no momento da pesquisa.

4) a) altura em centímetros (quantitativa); d) sexo, possíveis respostas: masculin e feminino (qualitativa).

6) Quando um respondente depara com um questionário muito longo, este s cansa de responder e pode deixar parte do questionário em branco, o responder apressadamente, comprometendo as respostas.

CAPÍTULO 3

1) {Getúlio, Paulo Cesar, Fabrício, Ermílio, Hiraldo, Mauro, Ercílio, Bartolomeu Cardoso, Josefina}
2) {2, 2, 5, 13, 9, 11, 10, 1, 16, 5}
3) {S, L, I, H}
4) Não, basta extrair 100 números da tabela, com quatro algarismos, pertencente ao conjunto {1650, 1651, ...,8840}, sem repetição.
11) $n = 2.500$
12) $n = 286$

CAPÍTULO 4

2) Tabela de freqüências múltipla: Distribuição de uma amostra de famílias quanto ao uso de programas de alimentação popular, por localidade de 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 freqüências: 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
freqüência	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 freqüências: 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 nenhuma residência com número de moradores muito diferente das demais.

8)

1*	3
1•	5678899
2*	0000011111122233333444
2•	55555556667999
3*	00111224
3•	5556666

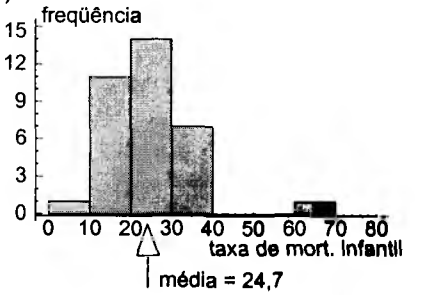
CAPÍTULO 6

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**

11) $M_d = 4$; $Q_i = 3{,}5$ e $Q_s = 5$

13) $E_i = 1$; $Q_i = 2$; $M_d = 4$; $Q_s = 5$ e $E_s = 12$

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) 0,5001

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)

Resultados	0,0	0,2	0,4	0,6	0,8	1,0
Probabilidades	0,0778	0,2592	0,3456	0,2304	0,0768	0,0102

13) 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) 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) 0,955 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) 65,0% ± 8,6%

8) a) 55,0% ± 15,7% b) 72,1% ± 13,7% c) 67,6% ± 15,2%

9) 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 Figuelra	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 amostrai 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.

10) 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.

13) 33,3% ± 7,3%

15) $n = 64$ (usando $z = 2$)

16) $n = 306$ (usando $z = 2$)

17) 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.

18) a) 40%

b) Com 95% de confiança, o intervalo 40,0% ± 3,4% contém a percentagem dos habitantes da cidade que apóiam a administração da prefeitura.

19) a) $n = 664$ b) 30,1% ± 4,6%. Com 90% de confiança, o intervalo 30,1% ± 4,6% contém a percentagem de pessoas que passariam a usar o produto.

20) 13,6% ± 2,6%

21) a) 3,50 ± 0,64 b) n = 84 (usou-se $z = 2$ e $N = 500$)

22) 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).

23) 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%

23) 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 do que a produtividade sem treinamento. H_1: Em média, a produtividade com treinamento é maior do que a produtividade sem treinamento. (teste unilateral)

b) H_0: Em média, a velocidade é igual ao valor anunciado. H_1: Em média, a velocidade é menor do que o valor anunciado. (teste unilateral).

c) H_0: As produtividades médias são iguais para os dois métodos de treinamento. H_1: As produtividades médias são diferentes para os dois métodos de treinamento. (teste bilateral).

11) a) Decide-se por H_1, pois o valor p é menor do que o nível de significância adotado. O risco dele 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 conseqüente 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: a percentagem de ouvintes que melhoram de impressão é a mesma da que piora; H_1: a maior parte dos ouvintes melhora de impressão.

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 $t = 2{,}70$ ==> $0{,}01 < p < 0{,}025$ (teste unilateral).

5) a) Rejeita H_0 ao nível de 5%, pois, $t = 3{,}04$ ==> $0{,}005 < p < 0{,}010$ (teste unilateral).

7) Sim, rejeita H_0 ao nível de 5%, pois, $t = 3{,}09$ ==> $0{,}001 < p < 0{,}005$ (teste bilateral).

8) a) Não (aceita H_0 ao nível de 5%), pois $t = 1{,}33$ ==> $0{,}05 < p < 0{,}10$ (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 defiram 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 $t = 2,92$ $\Rightarrow p \approx 0,002$.

Monte Verde x Encosta do Morro: existe diferença significativa, pois $t = 3,07$ $\Rightarrow \quad 0,002 < p < 0,005$.

Pq. da Figueira x encosta do Morro: não existe diferença significativa, pois, $t = 0,99 \Rightarrow 0,20 < p < 0,50$.

13) 17 (usando o gráfico da Figura 11.9).

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: $t = 1,62$ e $0,05 < p < 0,10$.

18) Não. Teste *t* bilateral para amostras independentes: $t = 0,97$ e $0,20 < p < 0,50$. Portanto, a diferença entre as médias amostrais pode ser explicadas meramente pelo acaso.

CAPÍTULO 12

1) a) 3,53 b) 2,40 c) Não $(0,10 < p < 0,25)$

2) Sim, pois $\chi^2 = 6,84 \Rightarrow 0,005 < p < 0,01$

3) a) Sim (rejeita H_0), pois $\chi^2 = 16,25 \Rightarrow 0,001 < p < 0,005$.

b) Sim (rejeita H_0), pois $\chi^2 = 11,18 \Rightarrow 0,001 < p < 0,005$.

c) Não (aceita H_0), pois $\chi^2 = 5,14 \Rightarrow 0,10 < p < 0,05$.

4) Admitindo a = 0,05. (a) Não (aceita H_0), pois $\chi^2 = 2,82 \Rightarrow 0,10 < p < 0,25$.

b) Sim (rejeita H_0), pois $\chi^2 = 16,25 \Rightarrow 0,001 < p < 0,005$.

5) $C^* = 0,107$. Os dados observados apresentam uma fraca associação entre sexo e tabagismo.

6) $C^* = 0,423$.

7) $\gamma = 0,3356$.

8) a) 0,214 b) -0,185

9) 0,665

10) Não. $(\chi^2 = 1,77,\ p > 0,25)$

11)

	Aprovação no vestibular	
Tipo de escola	não	sim
pública	13 (72%)	4 (33%)
particular	5 (28%)	8 (67%)
Total	18 (100%)	12 (100%)

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%.

12) Não. ($\chi^2 = 2{,}25$, $p > 0{,}25$)

13) a) Teste qui quadrado com correção de Yates.

b) Teste *t* para amostras independentes.

c) Teste *t* para amostras independentes.

CAPÍTULO 13

1)

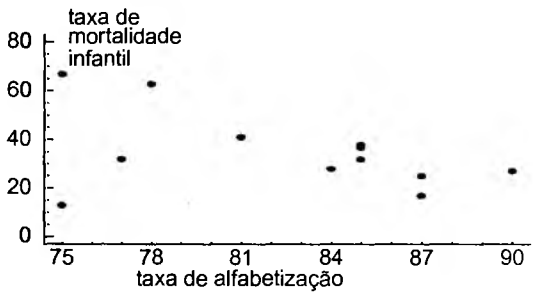


Diagrama de dispersão entre taxa de mortalidade infantil e taxa de crescimento demográfico em 12 municípios de SC, 1970/80.

6) $r = -0{,}43$. Em termos dos 12 municípios pesquisados, e na época de observação dos dados, verificou-se uma correlação negativa moderada entre "taxa de alfabetização" e "taxa de mortalidade infantil"; isto é, quanto maior o nível de alfabetização, tem-se uma leve tendência de redução na taxa de mortalidade infantil.

7) a) 0,69 b) 0,86

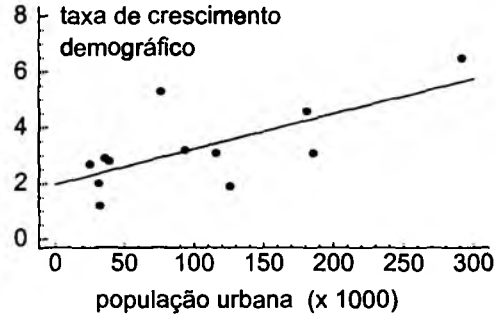
c) correlação positiva significativa.

10) a) Variável dependente: taxa de crescimento demográfico; e

variável independente: população urbana

b) (taxa de cresc. dem.) = 1,97 + (0,013).(pop. urbana). Obs.: População urbana está em unidades de 1.000 habitantes.

c)



d) Predição: taxa de crescimento de 5,8.

e) $R^2 = 48\%$

12) Não. Pela tabela VII o valor absoluto de *r* deveria ser no mínimo igual a 0,444 para ser significativo.

13) a) $r = -0,85$. Para as 6 famílias pesquisadas, tem-se uma correlação negativa forte entre *renda familiar* e *número de filhos.*

b) $\gamma = 0,33$. Em relação aos 10 indivíduos pesquisados, verifica-se uma correlação positiva fraca.

c) $C^* = 0,09$. Em relação aos 100 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 VII)

c) $y = 1,19 + 1,70\ x$

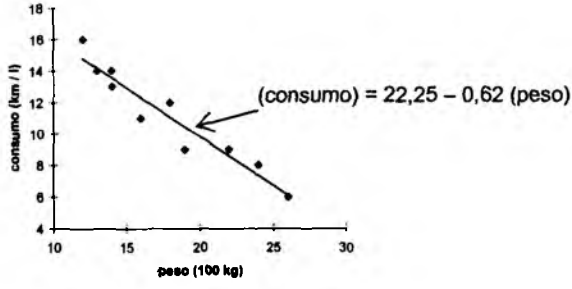
15) a) 49,1 kg b) 1,8 kg

16) a) $r = 0,96$ b) Correlação positiva 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 1200 a 2600 kg e, portanto, a equação de regressão deve ser usada apenas nesta faixa.

CONFECCIONADO NAS OFICINAS GRÁFICAS DA
IMPRENSA UNIVERSITÁRIA DA UNIVERSIDADE
FEDERAL DE SANTA CATARINA
SETEMBRO/2002
FLORIANÓPOLIS - SANTA CATARINA - BRASIL