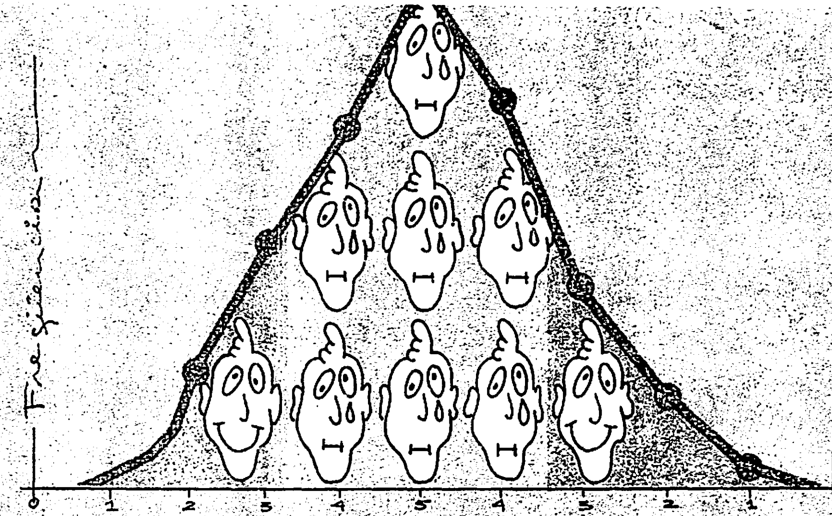


ESTATÍSTICA APLICADA A CIÊNCIAS HUMANAS

2ª edição

JACK LEVIN

HARBRA

CURVA NORMAL

A editora HARBRA ltda. coloca à disposição do leitor um gabarito de *curva normal*. Para recebê-lo, basta preencher os dados abaixo, destacar esta folha e enviar para:

editora HARBRA ltda.
Rua Joaquim Távora, 609 - Vila Mariana
04015 - São Paulo - SP

Atenção: Esta oferta só é válida com o envio deste *original*.

Nome: . . . .
Endereço: . . . .
Cidade: . . . . Estado: . . . .
Instituição em que ( ) leciona/( ) estuda: . . . .
. . . .
Endereço: . . . .
Cidade: . . . . Estado: . . . .

Estatística Aplicada a Ciências Humanas

2.ª edição

Estatística Aplicada a Ciências Humanas

2ª edição

Jack Levin
Northeastern University

Tradução e Adaptação:
Sérgio Francisco Costa

Títular de Estatística da
Faculdade de Filosofia, Ciências e Letras
São Caetano do Sul (SP)

editora HARBRA ltda.

A Flea, Michael,
Bonnie e Andrea.

Direção Geral: Julio E. Emöd
Supervisão Editorial: Maria Pia Castiglia
Coordenação e Revisão de Texto: Maria Elizabeth Santo
Revisão de Provas: Maria Braga de Jesus
Capa: Isabel
Impressão e Acabamento: Gráfica Editora Hamburg Ltda.

Rua Joaquim Távora, 609 – Vila Mariana – SP – SP
Telefones: 549-2244 e 571-0276

Tradução de Elementary Statistics in Social Research, 2ª edição

Publicado originalmente nos Estados Unidos por Harper & Row, Publishers, Inc.

Impresso no Brasil — *Printed in Brazil*

Conteúdo

Prefácio

O objetivo de *Estatística Aplicada a Ciências Humanas*, segunda edição, é igual ao da primeira edição: oferecer uma introdução à Estatística para alunos da área de Ciências Humanas (o que engloba campos tais como Sociologia, Política, Serviço Social, Psicologia, Administração, Educação) que não tenham tido a oportunidade de receber treinamento aprofundado em Matemática, e que estejam enfrentando seu primeiro curso de Estatística. Como antes, este livro *não* tem a intenção de ser um material de referência exaustivo; não deve também ser tomado como um texto adequado a cursos avançados de Métodos Estatísticos. Ao contrário, foi escrito e revisado com vistas a preencher a sentida necessidade de um enfoque significativo e compreensível daquilo que é básico em Estatística. Mais uma vez, as ilustrações "passo a passo" dos procedimentos estatísticos foram localizadas em pontos estratégicos do texto, além de terem sido apresentados numerosos problemas, tirados de pesquisas reais.

Como na edição anterior, este volume foi organizado em três partes: a Parte I (Capítulos 2 a 5) leva o estudante a compreender e a manejar alguns métodos úteis para a descrição e comparação de dados brutos. A Parte II (Capítulos 6 e 7) tem o propósito de servir de transição, uma vez que leva o leitor do tópico relacionado com a curva normal – importante instrumento descritivo – ao primeiro capítulo em que essa mesma curva é usada como fundamento para generalizações (partindo de amostras para populações). Ainda com o enfoque centrado na tomada de decisões, a Parte III (Capítulo 8 a 12) contém vários testes de significância bem conhecidos, além de procedimentos para a obtenção de coeficientes de correlação e uma introdução à análise de regressão.

Esta edição apresenta algumas mudanças dignas de nota. Por insistência de colegas professores, o Capítulo 10 foi substancialmente aumentado para permitir um tratamento mais extensivo da(s) estatística(s) não-paramétrica(s). Além disso, foram aumentados, esmiuçados ou acrescentados os seguintes tópicos: posto percentil, probabilidade, comparação múltipla de médias aritméticas, depois análise de variância, coeficiente gama e *r* de Pearson. Tendo em mente salientar a aplicação da Estatística em pesquisa, um novo capítulo (12) foi acrescentado; quanto às várias situações de pesquisa nele apresentadas, o aluno deve tentar selecionar o procedimento estatístico mais adequado a cada situação. O número de exercícios no fim de cada capítulo também foi aumentado. Finalmente, os apêndices foram ampliados com vistas a incluir uma revisão dos fundamentos de Matemática e uma lista de fórmulas.

Muitas pessoas contribuíram de diferentes formas para o desenvolvimento da segunda edição deste livro. A revisão aguçada de Kenneth Pollinger em *Contemporary Sociology* serviu de base para várias melhorias e adições. Agradeço a Richard Sprinthall e seus alunos do *American International College* (particularmente a Lynn Arnold, Cheryl Janes, Jim Lynch, Claire Nolen e Gary Zera) por terem-me alertado quanto à existência de alguns erros ou fontes de interpretação errônea na edição anterior. Meus agradecimentos especiais às seguintes pessoas pelas leituras críticas que fizeram das minhas revisões: George Bowlby, James Elliot, Roy Hansen, C. Lincoln Johnson, Carol Owen, Lawrence Rosen, Norman Roth, Ellen Bouchard Ryan e Larry Siegel. Também agradeço a Suzanne Johnson e Michael Weissbuch pelos comentários e sugestões espontâneas. Finalmente, uma palavra de agradecimento ao responsável pelo Patrimônio Literário do falecido Sir. Ronald A. Fisher, F.R.S., a Frank Yates, F.R.S e a Oliver & Boyd, Edinburgh, pela permissão para reimprimir as Tabelas III, IV, V e VI do seu livro *Statistical Tables for Biological, Agricultural and Medical Research*.

Jack Levin

Prefácio à Segunda Edição Brasileira

Seis anos após o lançamento de sua primeira edição, *Estatística Aplicada a Ciências Humanas* volta a ocupar as prateleiras: de roupagem nova e com conteúdo ampliado.

Para que o nível elementar da obra fosse mantido, nada se fez na primeira parte (Estatística Descritiva) exceto pequenas correções tipográficas; já na segunda parte, inúmeros foram os acréscimos, a saber:

- ao Capítulo 6 — Curva Normal — foram acrescentados alguns problemas e exercícios resolvidos e, ao fim do capítulo, dez exercícios suplementares, com respostas aos de números pares; acrescentou-se também um tópico sobre Distribuição Binomial com exercícios;
- ao Capítulo 10 — Testes Não-Paramétricos — foram acrescentados, além de exercícios e problemas — resolvidos e por resolver —, os seguintes tópicos:
    - Qui-quadrado de Aderência
    - Prova Exata de Fisher
    - Prova de Mann-Whitney (Prova *U*)
    - Prova de Subseqüências
    - Prova Binomial

Uma vez que o livro destina-se especificamente à área de ciências humanas, maior ênfase foi posta em assuntos de natureza não-paramétrica. E, para a redação desses assuntos*, a Editora convidou o Prof. Sérgio F. Costa, que, anos atrás, já se ocupou da tradução da primeira edição.

Foi mantido o esquema de respostas a apenas os problemas e exercícios pares, a fim de tornar possível a utilização dos ímpares em situações em que o conhecimento prévio das respostas burle a intenção pedagógica do professor.

Espera a Editora ter contribuído para o aprimoramento dos materiais disponíveis para o estudo da Estatística e ter facilitado a tarefa dos professores naqueles aspectos que, por muito teóricos, roubam tempo às aulas já tão curtas.

* Os tópicos acrescentados vêm assinalados com adaga (†).

Prefácio à Edição Brasileira

Estatística Aplicada a Ciências Humanas é um livro comparável a "coração de mãe": dá *muito* e exige *pouco*!

Mais do que lucidez e experiência — detectáveis não só a partir das explicações no texto como também da escolha bibliográfica criteriosa que embasou seu trabalho — o Prof. Jack Levin teve a *coragem* de escrever um livro *supersimplificado* — como se sua clientela consumidora tivesse parcos conhecimentos de Matemática. Sua coragem, repita-se, está muito mais em *fazer algo*, tomar a iniciativa de escrever este livro, do que na simples admissão dessa hipótese, que, quando mais não seja, deve dizer-lhe a longa prática, é *verdadeira* numa proporção alarmante!

Escrito numa linguagem amena e quase despojado do jargão matemático, *Estatística Aplicada a Ciências Humanas* destina-se a um público muito específico: estudantes de Ciências Humanas, refúgio errôneo dos que fogem das equações e dos cálculos, pois que, embora "humanas" — e talvez por isso mesmo — não podem prescindir das tão odiadas quantificações! Com a tradução que ora vai a público, procurou-se, sem perder de vista o caráter elementar que o livro deve manter, enriquecê-lo um pouco, à custa de notas de rodapé, talvez desnecessárias num primeiro contato com a Estatística, mas sem dúvida proveitosas em revisões posteriores.

No original, o Autor grifou muito menos palavras ou sentenças: a abundância de grifos, na tradução, foi uma tentativa *visual* de chamar a atenção do leitor para aspectos relevantes do texto.

O temor de que possa pairar a mais leve dúvida sobre o discernimento e saber do Prof. Levin obriga a que se diga o seguinte: transpira, evidente, em todo o livro, a intenção clara de levar o aluno ao raciocínio estatístico através de exemplos tirados do dia-a-dia, por isso, talvez, *menos refinados*, mas seguramente *mais próximos* de seu universo cognitivo.

Que o nome de Levin possa, através desta contribuição, ganhar lugar de destaque entre os estudiosos brasileiros.

São Paulo, julho de 1978

Sérgio Francisco Costa

1

Por que o Pesquisador usa Estatística?

Todos nós temos um pouco de cientista. Quase que diariamente, temos "palpites" com relação a acontecimentos futuros em nossas vidas, a fim de prever o que acontecerá em novas situações ou experiências. À medida que essas situações ocorrem, podemos, às vezes, confirmar ou sustentar nossas idéias; outras vezes, entretanto, não temos tanta sorte e, por isso, acabamos experimentando conseqüências desagradáveis.

Tomemos alguns exemplos familiares: poderíamos investir na bolsa de valores, votar em algum candidato que prometesse resolver os problemas nacionais, jogar nos cavalos, tomar um remédio para reduzir os incômodos de um resfriado, jogar dados num cassino, tentar "adivinhar" o que nossos examinadores irão perguntar nos exames ou acertar (às cegas) um encontro com uma garota desconhecida, marcado através de um amigo.

Às vezes, ganhamos; às vezes, perdemos. Desse modo, poderíamos fazer um belo investimento na bolsa, mas reconhecer que não fizemos a escolha do melhor candidato; poderíamos ganhar no cassino, mas descobrir que tomamos o remédio errado para nossa doença; conseguir aprovação nos exames, mas penar na companhia arranjada pelo amigo; e assim por diante. A verdade é que, infelizmente, nem todas as nossas previsões acabam-se tornando realidade.

A NATUREZA DA PESQUISA

De modo mais ou menos semelhante, o cientista tem idéias sobre a natureza da realidade (idéias que ele denomina *hipóteses*) e freqüentemente *testa* suas idéias através de pesquisa sistemática. Por exemplo, ele poderia levantar a hipótese de que crianças socialmente isoladas assistem mais televisão do que crianças bem integradas em seus grupos—e, a partir daí, fazer um levantamento em que fossem feitas perguntas tanto às crianças socialmente isoladas, quanto às crianças bem integradas com relação ao número de horas que elas passam diante do vídeo. Ou, ele poderia levantar a hipótese de que a família constituída por um só genitor (mãe ou pai ausente) gera maior delinqüência do que a família constituída por ambos os genitores (mãe e pai presentes); então, o pesquisador poderia entrevistar amostras de delinqüentes e não-delinqüentes que procedessem de famílias de ambos os tipos.

Desse modo, à semelhança do que ocorre nas ciências físicas, o cientista social freqüentemente faz pesquisas para aumentar seu cabedal dos problemas e suas conseqüências em seu campo de estudos. A pesquisa assume muitas formas e pode ser usada para investigar uma série enorme de problemas. O pesquisador pode, na qualidade de observador participante, trabalhar com grupos de delinqüentes; pode também trabalhar num levantamento amostral de suas tendências políticas, ou fazer uma análise de valores em folhetos políticos; ou, ainda, realizar um experimento a fim de determinar os efeitos do remanejamento de moradias em indivíduos atingidos por desapropriações.

POR QUE TESTAR HIPÓTESES?

É, de modo geral, desejável—para não dizer necessário—testar sistematicamente nossas hipóteses sobre a natureza da realidade, mesmo aquelas que parecem verdadeiras, lógicas ou evidentes. Os nossos "testes" de bom senso diário geralmente baseiam-se em idéias pré-concebidas muito estreitas, se não tendenciosas, e em experiências pessoais, o que pode levar-nos a aceitar conclusões não válidas a respeito da natureza dos fenômenos. Para demonstrar esse ponto, vamos examinar as seguintes hipóteses, que foram testadas com um grande número de soldados durante a Segunda Guerra Mundial. Você teria sido capaz de prever estes resultados com base em suas experiências diárias? Você acha que valeu a pena testá-las? Ou será que elas parecem tão óbvias e evidentes que dispensam pesquisas sistemáticas?

1. Homens de nível educacional mais alto apresentaram maior quantidade de sintomas psiconeuróticos do que aqueles que haviam recebido menos instrução.
2. Homens procedentes do meio rural mantiveram-se mais bem-humorados durante o período de guerra do que os soldados recrutados nos meios urbanos.
3. A capacidade dos soldados procedentes do Sul para suportar o clima quente das Ilhas dos Mares do Sul era maior que a dos soldados vindos do Norte.
4. À medida que a guerra prosseguia, os homens sentiam-se mais ansiosos para retornar aos Estados Unidos do que após a derrocada dos alemães.

Se você acredita que essas relações eram tão óbvias que qualquer teste sistemático teria sido dispensável, saiba que cada item "representa a conclusão oposta ao que realmente se verificou. Soldados parcamente educados neurotizaram-se mais do que os que receberam boa educação; os sulistas não demonstraram maior habilidade que os nortistas em adaptar-se a um clima tropical; ... e assim por diante."[1] Ficar apenas na dependência do bom senso ou da experiência do dia-a-dia tem seus limites!

OS ESTÁGIOS DA PESQUISA

Testar sistematicamente nossas idéias sobre a natureza da realidade muitas vezes requer uma pesquisa cuidadosamente planejada e executada, onde:

[1] Paul Lazarsfeld, "The American Soldier—An Expository Review", *Public Opinion Quarterly*, outono de 1949, pág. 380.

1. o problema a ser estudado é reduzido a uma hipótese testável* (por exemplo: "famílias de um genitor produzem mais delinqüência que famílias de dois genitores");
2. um conjunto de instrumentos adequado é desenvolvido (por exemplo, são elaborados um questionário ou esquema de uma entrevista);
3. dados são coletados (isto é, o pesquisador pode ir a campo e fazer uma contagem ou um inquérito);
4. os dados são analisados em cotejo com as hipóteses iniciais; e
5. os resultados da análise são interpretados e comunicados ao público, por exemplo, por meio de conferência ou publicação.

Como veremos nos capítulos subseqüentes, o material apresentado neste livro trata com maior profundidade daqueles aspectos da pesquisa relacionados com a análise de dados (ver 4, acima), os quais, depois de colecionados ou recolhidos pelo pesquisador, são analisados à luz das hipóteses iniciais. É nesse estágio da pesquisa que os dados brutos são tabulados, calculados, contados, resumidos, reclassificados, comparados ou, numa palavra, *organizados*, a fim de que a precisão ou a validade de nossas hipóteses possam ser testadas.

O USO DE SÉRIES NUMÉRICAS EM PESQUISA

Qualquer pessoa que já tenha feito pesquisa sabe que os problemas que surgem na análise dos dados devem ser seriamente confrontados na fase de *planejamento* (da pesquisa), uma vez que de seu significado depende a natureza de decisões em fases posteriores. Tais problemas amiúde afetam não só aspectos do delineamento da pesquisa** como também a escolha dos tipos de instrumentos utilizáveis na coleta de dados. Por esta razão, buscamos constantemente técnicas e métodos que permitam melhorar a qualidade da análise dos dados.

Muitos pesquisadores acham que é essencial fazer *mensuração* ou empregar uma série de números na análise dos dados. Por isso, os pesquisadores desenvolveram medidas*** para uma grande variedade de fenômenos, tais como: prestígio ocupacional, atitudes políticas, autoritarismo, alienação, anomia, delinqüência, classe social, preconceito, dogmatismo, realização, etnocentrismo, vizinhança, religiosidade, ajustamento conjugal, mobilidade ocupacional, urbanização, *status* sociométrico, fertilidade etc.

Os números têm pelo menos três funções importantes para o pesquisador, dependendo do *nível de mensuração* específico que ele empregue. Especificamente, uma série de valores pode ser usada

1. para categorizar ao nível de mensuração denominado *nominal*,

* N.T.: Reduzir um problema a uma *hipótese testável* significa definir o problema e a(s) hipótese(s) em *termos operacionais*. Uma definição operacional explicita comportamentos observáveis (direta ou indiretamente).

** N.T.: O ideal teria sido conservar a palavra "design". Há quem prefira "modelo" ou mesmo "desenho".

*** N.T.: Na verdade, desenvolveram "processos de mensuração".

2. para atribuir *postos* ou *ordem* ao nível de mensuração denominado *ordinal* e
3. para *avaliar* ao nível de mensuração denominado *intervalar.*

Antes de prosseguir com a discussão do papel da Estatística na pesquisa vamos parar aqui para examinar algumas características mais importantes desses níveis de mensuração, características que, mais adiante, assumirão grande importância ao tentarmos aplicar técnicas estatísticas a determinadas situações de pesquisa.

Nível Nominal

O nível *nominal* de mensuração envolve simplesmente o ato de nomear ou rotular; em outras palavras, consiste em colocar indivíduos* em categorias e contar a freqüência com que ocorrem. Para ilustrar, poderíamos usar uma medida de nível nominal para indicar se cada respondente tem ou não preconceito com relação a determinado grupo social ("Neutrália"). Como demonstra a Tabela 1.1, poderíamos submeter 10 alunos de determinada classe a um questionário e verificar que 5 podem ser considerados "preconceituosos" (= 1) enquanto que 5, "não-preconceituosos" (= 2).

Outras medidas de nível nominal em pesquisa são: sexo (masculino versus feminino); classe sócio-econômica (alta e baixa), partido político (A e B), caráter social (voltado "para dentro" e tradicional), modo de adaptação (conformismo, inovação, ritualismo, fuga, rebeldia), orientação no tempo (presente, passado e futuro), urbanização (urbano, rural e suburbano)—para mencionar algumas.

Ao trabalhar com dados nominais, devemos ter em mente que *cada sujeito deve ser colocado em uma—e somente uma—categoria.* Este requisito indica que as categorias não devem sobrepor-se, isto é, devem ser *mutuamente excludentes.*** Desse modo, a raça de um respondente não pode ser simultaneamente classificada em "branca" e "preta"; qualquer respondente que seja classificado como "macho" não pode ser também classificado como "fêmeo". O requisito implica também o fato de que as categorias devem ser *exaustivas*—ou seja, deve haver um lugar específico para cada sujeito experimental que apareça. A título de ilustração, imagine um estudo em que todos os respondentes sejam entrevistados e categorizados por raça, como brancos ou negros. Onde é que colocaríamos um respondente chinês na hipótese de surgir um? Neste caso, poderia tornar-se necessário ampliar as categorias originais e incluir "orientais" ou, então, admitindo-se que a maioria dos respondentes será de raça branca (ou de raça negra), incluir a categoria "outros", onde tais exceções possam ser colocadas.

O leitor deve observar que os dados nominais não são graduáveis, ordenáveis ou escalonáveis relativamente a qualidades tais como melhor ou pior, mais alto ou mais baixo, mais ou menos. Obviamente, então, uma medida nominal de sexo não significa que os machos sejam "superiores" ou "inferiores" às fêmeas. Os dados nominais são simplesmente rotulados, às vezes por nome (macho versus fêmea, preconceituosos versus não-preconceituosos), outras vezes por número (1 versus 2), mas sempre com o

---

* N.T.: É comum usarem-se em pesquisa os termos "sujeito", "indivíduo", "caso" sinonimamente. Além disso, não há necessariamente alusão a "pessoas" quando se usam os termos "sujeito" ou "indivíduo". A idéia básica é *sujeito experimental.*

** N.T.: É comum a expressão "mutuamente exclusivo".

propósito de agrupar os sujeitos em categorias separadas para indicar similitude ou diferença com referência a uma dada qualidade ou característica.

TABELA 1.1 Atitudes de Dez Estudantes Universitários com Relação a "Neutralianos": Dados Nominais

Atitude com Relação aos "Neutralianos"	*Freqüência*
1 = Preconceituosos	5
2 = Não-preconceituosos	5
Total	10

Nível Ordinal

Quando o pesquisador vai além desse nível de mensuração e procura *ordenar* seus sujeitos em função do grau que apresentam de determinada característica, ele está trabalhando no nível *ordinal* de mensuração. A natureza da relação entre categorias ordinais depende da característica que o pesquisador procura medir. Para citar um exemplo familiar, ele poderia classificar indivíduos com referência a *status* sócio-econômico e dispô-los em "classe baixa", "classe média" e "classe alta". Ou, em lugar de categorizar os alunos de uma dada classe em preconceituosos *ou* não-preconceituosos, ele poderia classificá-los de acordo com o grau de preconceito contra os "neutralianos", à semelhança do que vai indicado na Tabela 1.2.

O nível ordinal de mensuração fornece informações sobre a ordenação das categorias, mas não indica a *magnitude das diferenças* entre os valores. Por exemplo, o pesquisador que emprega uma medida de nível ordinal para estudar a variável preconceito *não sabe quanto um respondente é mais (ou menos) preconceituoso que outro.* No exemplo dado acima, não é possível determinar quanto Adriana é mais preconceituosa do que Maria ou quão menos preconceituosa é Roberta com relação a Patrícia ou Pedro. Isto ocorre porque os intervalos entre os pontos ou postos numa escala ordinal não são conhecidos ou, melhor dizendo, dotados de significado. Portanto, não é possível atribuir *valores* a sujeitos localizados em diferentes pontos da escala.

TABELA 1.2 Atitudes de Dez Estudantes Universitários com Relação a "Neutralianos": Dados Ordinais

Aluno	*Posto*
Adriana	1 – mais preconceituosa (primeiro posto)
Maria	2 – segundo
Benito	3 – terceiro
José	4 – quarto
Cátia	5 – quinto
Geraldo	6 – sexto
Filipe	7 – sétimo
Pedro	8 – oitavo
Patrícia	9 – nono
Roberta	10 – menos preconceituosa

Nível Intervalar

Em contraposição, o nível *intervalar* de mensuração orienta-nos relativamente à ordem das categorias, bem como indica-nos a *distância* exata entre elas. Escalas intervalares implicam *unidades constantes de medida* (por exemplo: cruzeiros ou centavos, graus Fahrenheit ou centígrados, metros ou centímetros, minutos ou segundos), as quais comportam *intervalos iguais* entre os vários pontos da escala.

Desse modo, uma medida intervalar de preconceito contra os "neutralianos"—por exemplo, as respostas a uma série de perguntas sobre esses sujeitos avaliadas numa escala de 0 a 100 (o máximo de preconceito correspondendo a 100)—poderia fornecer dados como os que constam da Tabela 1.3 sobre os dez estudantes numa dada sala de aula. Como bem demonstra a Tabela 1.3, agora podemos ordenar os alunos em termos de seus preconceitos e, além disso, indicar as distâncias que separam uns dos outros. Por exemplo, é possível dizer que Roberta é a menos preconceituosa da classe, uma vez que ela recebeu o escore* mais baixo. Podemos dizer também que Roberta é apenas ligeiramente menos preconceituosa que Patrícia ou Osvaldo, mas muito menos preconceituosa do que Adriana, Maria, Benito ou José, os quais receberam valores (escores) extremamente altos. Dependendo do propósito para o qual o estudo foi concebido, seria importante determinar tais informações, as quais, entretanto, não estariam disponíveis no nível ordinal de mensuração.

TABELA 1.3 **Atitudes de Dez Estudantes Universitários com Relação a "Neutralianos": Escala Intervalar**

Aluno	*Escore*[a]
Adriana	98
Maria	96
Benito	95
José	94
Cátia	22
Geraldo	21
Filipe	20
Osvaldo	15
Patrícia	11
Roberta	6

[a] Escores mais altos indicam maior preconceito.

FUNÇÕES DA ESTATÍSTICA

É quando o pesquisador usa números—quando ele *quantifica seus dados* no nível de mensuração nominal, ordinal ou intervalar—que ele muito provavelmente emprega a Estatística como um instrumento de (1) *descrição* ou de (2) *decisão*. Vamos agora examinar mais de perto essas importantes funções da Estatística.

* N.T.: Escore significa "nota", "valor", "medida", conforme o caso.

Descrição

Para chegar a conclusões ou obter resultados, o pesquisador geralmente estuda centenas, milhares ou mesmo quantidades ainda maiores de pessoas ou grupos. Como um caso extremo, consideremos o trabalho do "United States Bureau of the Census" (Serviço Norte-americano de Recenseamento) na enumeração exaustiva de toda a população dos Estados Unidos, para o que mais de 200 milhões de sujeitos foram entrevistados. Apesar da ajuda de numerosos procedimentos sofisticados idealizados para a tarefa, descrever e resumir a enorme quantidade de dados que se originam de projetos de pesquisa social é sempre um trabalho de proporções astronômicas.

Para dar um exemplo familiar, as notas de exame de uma classe de apenas 80 alunos foram catalogadas na Tabela 1.4. Percebe-se algum padrão nessas notas? Poder-se-ia descrevê-las em poucas palavras? Ou em poucas sentenças? De modo geral, seriam elas particularmente altas ou baixas?

TABELA 1.4 **Notas de Exame de 80 Alunos**

72	83	91	29
38	89	49	36
43	60	67	49
81	52	76	62
79	62	72	31
71	32	60	73
65	28	40	40
59	39	58	38
90	49	52	59
83	48	68	60
39	65	54	75
42	72	52	93
58	81	58	53
56	58	77	57
72	45	88	61
63	52	70	65
49	63	61	70
81	73	39	79
56	69	74	37
60	75	68	46

Ainda que se usem os mais rudimentares princípios de estatística descritiva—da maneira como são apresentados nos capítulos subseqüentes deste livro—é possível caracterizar a distribuição de notas de exame da Tabela 1.4 com bastante clareza e precisão, de sorte que as tendências gerais ou as características grupais possam ser mais rapidamente descobertas e mais facilmente comunicadas a quase todas as pessoas. Primeiro, poderíamos reagrupar as notas em ordem decrescente (da mais alta para a mais baixa), a fim de reuni-las num número bem menor de categorias. Essa *distribuição agrupada* (condensada) *de freqüências*, tal como figura na Tabela 1.5 (e que será discutida minuciosamente no Capítulo 2), apresentaria as notas "encaixadas" em categorias (classes) mais abrangentes, juntamente com o número ou a *freqüência* (*f*) de alunos cujas notas caíssem nessas categorias.

Pode-se prontamente observar, por exemplo, que 17 alunos receberam notas entre 60 e 69, enquanto que somente 2 alunos receberam notas entre 20 e 29.

Outro procedimento útil (explicado no Capítulo 3) seria reagrupar os dados graficamente. Tal como ilustra a Figura 1.1, poderíamos localizar as categorias (classes) de notas (de 20-29 a 90-99) ao longo de um dos eixos do gráfico (isto é, a *linha de base, horizontal*) e o número delas (categorias), ou seja, as freqüências, ao longo do outro eixo (vale dizer: o *eixo vertical*).

TABELA 1.5 Notas de Exame de 80 Alunos: Distribuição Agrupada de Freqüências

Notas	*f*
90-99	3
80-89	7
70-79	16
60-69	17
50-59	15
40-49	11
30-39	9
20-29	2

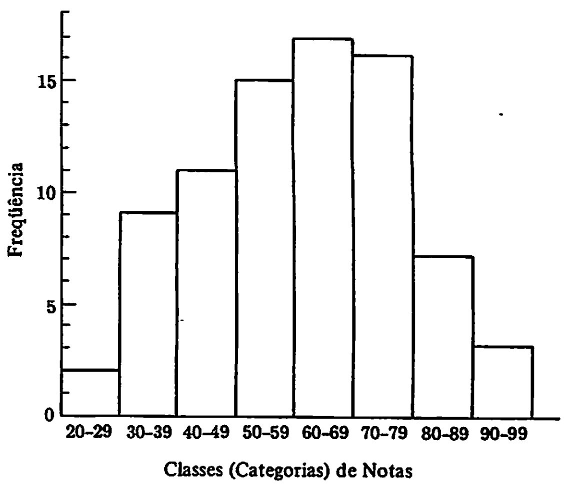


FIGURA 1.1 Notas de Exame de 80 Alunos Dispostas num Gráfico de Barras

Essa disposição tem como resultado uma representação gráfica facilmente visualizada (por exemplo, o gráfico de barras), onde podemos ver que a maioria das notas situa-se entre 50 e 80, enquanto que as notas muito mais altas ou baixas são relativamente poucas.

No Capítulo 4 será apresentado um método estatístico que é particularmente conveniente e útil. Trata-se de algo com que já estamos mais ou menos familiarizados e que consiste em perguntar: Qual a nota *média* desse grupo de 80 estudantes? A *média aritmética* (ou simplesmente *média*), que se obtém pela divisão da soma de todas as notas pelo número de alunos, dá-nos uma idéia mais clara da tendência geral do grupo. A média aritmética do exemplo em questão é de 60,5, nota bastante baixa se comparada com outras médias de classe com as quais a maioria dos alunos já está familiarizada. Aparentemente, esse grupo de 80 alunos saiu-se relativamente mal como um todo.

Assim, com o auxílio de recursos estatísticos, tais como distribuições condensadas de freqüências, gráficos e médias aritméticas, é possível detectar e descrever modelos ou tendências em distribuições de valores (por exemplo, as notas da Tabela 1.4), o que poderia ter passado despercebido pelo observador não preparado. Neste contexto, então, a Estatística pode ser definida como um *conjunto de técnicas para a redução de dados quantitativos (ou seja, uma série de números) a um número menor de termos descritivos que sejam mais convenientes, e facilmente comunicáveis.*

Tomada de Decisões

Quando o propósito é testar hipóteses, muitas vezes é necessário ir além da mera descrição. É também geralmente necessário fazer inferências, isto é, tomar decisões com base em dados colhidos de uma *amostra*, o que quer dizer: *estudar apenas uma pequena porção de um grupo maior*—que, por seu turno, representa o centro de interesse para estudo. Custo, tempo e necessidade de supervisão adequada são fatores que muitas vezes impedem uma enumeração completa ou rol do grupo todo (grupo que os pesquisadores chamam *população* ou *universo*).

Como veremos no Capítulo 7, todas as vezes que o pesquisador testa suas hipóteses numa amostra, ele precisa decidir se é de fato acertado generalizar suas descobertas para toda a população da qual a amostra foi tirada. Inevitavelmente surgem erros de amostragem, mesmo quando esta tenha sido convenientemente planejada e executada. Este é o problema da generalização ou *inferência* da amostra para a população.[2]

[2] *Observação para o leitor.* O conceito de "erro amostral" (de amostragem) vai ser discutido com maior profundidade no Capítulo 7. Todavia, a fim de entender melhor o problema da inevitabilidade de erro ao amostrar* de um grupo maior, o leitor pode querer desenvolver o seguinte experimento. Use a Tabela 1.4 (como fonte de referência) a qual contém as notas de uma população de 80 alunos. Aleatoriamente (por exemplo, feche os olhos e escolha com o dedo), selecione uma amostra de algumas notas (por exemplo, 5) da lista toda. Procure a nota média somando os valores sorteados e dividindo o resultado por 5, que equivale ao número total de médias. Já foi mencionado que a nota média para toda a classe (80 alunos) era 60,5. Em que medida a média da sua amostra difere da média da classe toda (60,5)? Faça este experimento usando várias outras amostras de notas selecionadas ao acaso—e sempre sorteadas do *mesmo* grupo (de 80 alunos). Com significativa perseverança você verificará que a sua média amostral vai diferir quase sempre, pelo menos um pouco, da média obtida para o grupo todo. É essa diferença, isto é, a diferença entre a média da população (80 alunos) e cada uma das médias amostrais encontradas que se chama "erro amostral".

* N.T.: "Amostrar" é um neologismo empregado em Estatística para significar "colher uma amostra".

A Estatística pode ser muito útil quando o propósito do pesquisador é generalizar suas descobertas a partir de amostras pequenas para populações maiores–com um alto grau de segurança. Para ficar mais bem entendido o propósito da Estatística como instrumento de tomada de decisão e ainda o conceito de generalização de amostras para populações, vamos examinar os resultados de um estudo hipotético que foi realizado para testar a seguinte hipótese:

Hipótese: A probabilidade de estudantes universitários do sexo masculino terem experimentado maconha é maior que a probabilidade de estudantes do sexo feminino terem feito o mesmo.

Os pesquisadores decidiram, neste estudo, testar sua hipótese numa universidade urbana na qual cerca de 20.000 estudantes estavam matriculados (10.000 do sexo masculino e 10.000 do sexo feminino). Em virtude do custo e de fatores ligados a tempo, eles não puderam entrevistar todos os estudantes da universidade, mas obtiveram da secretaria uma lista completa de todos os alunos. Dessa lista, cada centésimo aluno (metade do sexo masculino, metade do sexo feminino) foi selecionado para fazer parte da amostra, sendo subseqüentemente entrevistado por membros da equipe de pesquisa que tinham sido treinados para este propósito. Os entrevistadores perguntaram a cada um dos 200 sujeitos da amostra se ele ou ela tinha alguma vez experimentado maconha, anotando em seguida o sexo do respondente. Depois de terem sido completadas todas as entrevistas e remetidas para o escritório central, os resultados do estudo foram tabulados por sexo e apresentados conforme a Tabela 1.6.

Observe que os resultados obtidos nesta amostra de 200 estudantes, tal como figura na Tabela 1.6, têm a seguinte direção hipotética: de cada 100 alunos, 60 declararam que experimentaram maconha, enquanto que, com relação às alunas, esse número cai para 40 em cada 100. Obviamente, nessa pequena amostra, havia maior probabilidade de os sujeitos do sexo masculino terem experimentado maconha do que os de sexo feminino. Para os nossos propósitos, entretanto, a questão mais importante é saber se essas diferenças de sexo ligadas ao uso de maconha são grandes o suficiente para que possamos, com confiança, generalizá-las para uma população universitária muito maior de 20.000 alunos. Será que tais resultados representam diferenças populacionais verdadeiras? Ou será que obtivemos diferenças casuais entre alunos e alunas em virtude–estritamente–de erro amostral, ou seja, do erro que ocorre todas as vezes que retiramos de um grupo maior um grupo menor (amostra)?

TABELA 1.6 Uso de Maconha em Função do Sexo do Respondente. Caso I

	Sexo dos Respondentes	
Uso de Maconha	*Alunos*	*Alunas*
Número de estudantes que experimentaram	60	40
Número de estudantes que não experimentaram	40	60
Total	100	100

Para ilustrar o problema da generalização de resultados amostrais em função das respectivas populações, vamos imaginar que os pesquisadores tivessem obtido os resultados da Tabela 1.7. Observe que tais resultados mantêm-se na direção prevista: 55 alunos em oposição a apenas 45 alunas já experimentaram maconha. Ora, o nosso problema não é ainda o de generalizar esses resultados para o grupo maior, vale dizer, a população universitária? Não será possível que uma diferença dessa magnitude (10 alunos mais que alunas) tenha ocorrido simplesmente por acaso? Ou será que podemos dizer com confiança que tais diferenças "pequenas" refletem uma diferença real entre alunos e alunas daquela determinada universidade?

TABELA 1.7 Uso de Maconha em Função do Sexo do Respondente. Caso II

	Sexo dos Respondentes	
Uso de Maconha	*Alunos*	*Alunas*
Número de estudantes que experimentaram	55	45
Número de estudantes que não experimentaram	45	55
Total	100	100

Vamos levar este exemplo um passo mais adiante. Suponham que os pesquisadores tivessem obtido os dados constantes da Tabela 1.8. Aí as diferenças entre rapazes e moças não poderiam ser muito menores e prevaleceria ainda a direção estabelecida na hipótese: 51 rapazes em contraste com 49 moças experimentaram maconha–o que implica uma diferença de apenas 2 sujeitos a mais do sexo masculino. Quantos de nós estaríamos dispostos a chamar *esta* descoberta de diferença populacional verdadeira, em lugar de produto do acaso ou erro amostral?

TABELA 1.8 Uso de Maconha em Função do Sexo do Respondente. Caso III

	Sexo dos Respondentes	
Uso de Maconha	*Alunos*	*Alunas*
Número de estudantes que experimentaram	51	49
Número de estudantes que não experimentaram	49	51
Total	100	100

Então, onde se situa a linha divisória? Em que ponto (ou em que momento) uma diferença amostral torna-se suficientemente grande para ser tratada como significante ou real? Com o auxílio da Estatística podemos de pronto e com alto grau de confiança, tomar tais decisões sobre as relações existentes entre amostras e populações. Para ilustrar: se tivéssemos usado um dos testes de significância discutidos mais adiante (por exemplo, qui-quadrado–ver Capítulo 10), teríamos sabido que *somente aqueles resultados* que aparecem na Tabela 1.6 poderiam ser generalizados para a população de 20.000 estudantes universitários, o que equivale a dizer que se de cada 100 alunos 60 já usaram maconha, e se de cada 100 alunas 40 fizeram o mesmo, tal descoberta é substancial o suficiente para ser extrapolada para a população toda com um alto grau de confiança. O nosso teste estatístico diz-nos que há somente 5 possibilidades em 100 de estarmos errados! Em contrapartida, os resultados apresentados nas Tabelas 1.7 e 1.8 são estatisticamente *não-significantes*, de sorte que resultam muito mais de erro amostral do que de diferenças de sexo no uso dessa substância. Usando um critério estatístico mais uma vez, podemos concluir que esses resultados não refletem diferenças populacionais verdadeiras, e sim, meros erros amostrais.*

No presente contexto, portanto, *Estatística é um conjunto de técnicas para a tomada de decisões que auxiliam os pesquisadores na tarefa de fazerem inferências de amostras para populações e, a partir daí, nos testes das hipóteses levantadas sobre a natureza da realidade.*

RESUMO

Este capítulo "juntou" nossas previsões diárias sobre acontecimentos futuros com as experiências do pesquisador, que usa a Estatística como um recurso para testar suas hipóteses a respeito da natureza da realidade. A mensuração foi discutida em termos de dados nominais, ordinais e intervalares. Foram identificadas, com a análise de dados (e resumidamente discutidas e ilustradas), as duas seguintes funções básicas da Estatística:

1. descrição (isto é, redução de dados quantitativos a termos descritivos mais convenientes e em menor número) e
2. tomada de decisões (ou seja, inferências de amostras para populações).

* N.T.: Não confundir *ideal* com *real*. O *ideal* seria *não cometer erros*; na realidade, por maiores que sejam os controles, *erros casuais sempre ocorrem*. É praticamente impossível evitar erros amostrais.

PARTE I
Descrição

2

Organização de Dados

A coleta de dados envolve um sério esforço da parte do pesquisador interessado em aumentar seu conhecimento do comportamento humano. Para entrevistar ou, por outra, eliciar informações de um grupo de beneficiários da previdência, de estudantes universitários, de viciados em drogas, de moradores de pensões, de homossexuais, de norte-americanos da classe média ou de respondentes de outras categorias, tornam-se necessários certo grau de previsão, planejamento cuidadoso, controle, além do tempo despendido em campo.

Completar a coleta de dados, entretanto, é somente o começo daquilo que constitui a análise estatística. A coleta de dados é o passo inicial, a matéria-prima com que o pesquisador deve trabalhar se o seu objetivo é analisar os dados, obter resultados e testar hipóteses acerca da natureza da realidade.

DISTRIBUIÇÕES DE FREQÜÊNCIA DE DADOS NOMINAIS

O marceneiro transforma madeira bruta em mobiliário; o mestre-cuca dá aos alimentos básicos aquele toque de sabor que enriquece as mesas. Por processo análogo, o pesquisador–auxiliado por "receitas" denominadas fórmulas e técnicas–procura transformar seus dados brutos num conjunto de mensurações, organizadas e dotadas de sentido, que possam ser usadas para testar suas hipóteses iniciais.

TABELA 2.1 Estudantes de Ambos os Sexos Presentes a uma Exposição de Pintura

Sexo	*Freqüência (f)*
Masculino	80
Feminino	20
Total	100

O que o pesquisador pode fazer para organizar a avalanche de dados brutos que ele recolhe de seus respondentes? Que é que ele faz para transformar essa massa de dados brutos num conjunto-resumo fácil de entender? O primeiro passo poderia ser construir uma *distribuição de freqüências* sob a forma de tabela.

Examinaremos a distribuição de freqüências da Tabela 2.1. Observem, em primeiro lugar, que a tabela é encabeçada por um *número* (2.1) e por um *título* que dá ao leitor uma idéia da natureza dos dados apresentados—"Estudantes de ambos os sexos presentes a uma exposição de pintura". Esse é o procedimento padrão; todas as tabelas devem ser claramente tituladas e, quando forem apresentadas em seqüência, devem conter também um número.

As distribuições de freqüência de dados nominais consistem em duas colunas. Na Tabela 2.1, por exemplo, a coluna da esquerda indica a característica de interesse* (sexo) e contém as categorias de análise (homens e mulheres). Uma coluna adjacente titulada "Freqüência" ou "f" indica o número de sujeitos em cada categoria (80 e 20, respectivamente), bem como o número total de sujeitos ($N = 100$).

Uma rápida inspeção da distribuição de freqüência constante da Tabela 2.1 revela perfeitamente que muito mais alunos do que alunas compareceram à exposição: dos 100 alunos presentes 80 eram do sexo masculino.

COMPARAÇÃO DAS DISTRIBUIÇÕES

Suponham, entretanto, que desejemos comparar a presença de estudantes na ala de arte moderna com a presença de estudantes na ala de arte clássica. Fazer comparações *entre* distribuições de freqüências constitui um procedimento usado amiúde para tornar os resultados mais claros e a informação mais ampla. A comparação específica que o pesquisador faz é determinada pela questão a que ele procura responder.

Voltando ao nosso exemplo hipotético, poderíamos perguntar: será que os rapazes são mais propensos a comparecer a exposições de pintura moderna e clássica do que as moças, ou será que eles se distribuem casualmente? Para responder, poderíamos comparar a distribuição de 100 alunos** presentes à ala de pintura moderna com a distribuição de outros 100—da mesma universidade—presentes à ala de pintura clássica. Vamos imaginar que os resultados obtidos sejam os constantes da Tabela 2.2.

Tal como demonstra a tabela, de 100 estudantes presentes na ala de pintura moderna, 30 eram do sexo feminino; já na ala de pintura clássica, de 100 apenas 20 eram do sexo feminino. Isto já nos dá muito mais informações do que a simples distribuição de freqüência apresentada na Tabela 2.1. Assim, podemos agora dizer que os rapazes tiveram uma participação maior do que suas colegas de universidade. Podemos

* N.T.: A característica de interesse é habitualmente designada por *variável observacional, variável de observação, variável de análise.*

**N.T.: Obviamente, o exemplo é forçado, só se admitindo por motivos de ordem didática. Por outras palavras, ele chega a ser claro, ainda que não rigorosamente correto. No Capítulo 7 o leitor encontrará algumas noções de Técnica de Amostragem que muito o auxiliarão a entender as limitações do exemplo acima.

dizer também que nas ocasiões em que as moças compareceram, elas mostraram-se, de certa forma, mais inclinadas para a ala de pintura clássica que para a de moderna.

TABELA 2.2 Comparecimento de Alunos de Ambos os Sexos a uma Exposição de Pintura (Arte Clássica e Arte Moderna)

	Comparecimento às Exposições	
	Arte Moderna	*Arte Clássica*
Sexo	f	f
Masculino	80	70
Feminino	20	30
Total	100	100

Proporções e Porcentagens

Quando o pesquisador estuda distribuições de tamanho igual (isto é, amostras de mesmo tamanho), os dados freqüenciais podem ser usados no estabelecimento de comparações entre os grupos. Assim, o número de rapazes presentes à exposição de pintura clássica pode ser diretamente comparado com o número de rapazes presentes à de pintura moderna, uma vez que sabemos terem visitado cada ala da exposição exatamente 100 estudantes. Em geral não é possível, entretanto, estudar distribuições com exatamente o mesmo número de sujeitos. Por exemplo, como podemos estar certos ("a priori") de que precisamente 100 alunos comparecerão a ambas as exposições? *

A fim de tornar tais resultados mais claros (mais "legíveis"), necessitamos de um método que permita *padronizar distribuições de freqüências quanto ao tamanho*, ou seja, uma forma de comparar grupos a despeito de diferenças nas freqüências totais. Dois dos mais úteis e populares métodos de padronizar tamanho e comparar distribuições são o método da *proporção* e o método da *porcentagem.*

No método da proporção, compara-se o número de sujeitos de uma dada categoria com o total de sujeitos que compõe a distribuição (e na qual a dada categoria se inclui). Podemos transformar qualquer freqüência numa proporção P, através da divisão do número de sujeitos de uma dada categoria f pelo número total de sujeitos da distribuição (N). Em símbolos:

$$P = \frac{f}{N} \; {}^{**}$$

*N.T.: Veja-se, a propósito, o que ficou dito na nota de rodapé anterior sobre a arbitrariedade desse procedimento.

N.T.: Como se verá no Capítulo 6, intitulado "A Curva Normal", a ***probabilidade* de um evento é um ***quociente*** em que o numerador representa o número de casos *favoráveis* à ocorrência desse

Portanto, 10 rapazes num grupo de 40 estudantes presentes a uma exposição dão origem à proporção $P = \frac{10}{40} = 0,25$.

Apesar da utilidade do método das proporções, muitas pessoas preferem indicar o tamanho relativo de um conjunto de valores em termos de *porcentagem*, que, ao cabo, vem a ser a própria freqüência com que determinada categoria ocorre *em relação ao número 100.* Para calcular a porcentagem, simplesmente multiplicamos uma dada proporção por 100, donde a fórmula

$$\% = (100)\frac{f}{N}.$$

Portanto, a proporção $P = \frac{10}{40}$ do exemplo anterior transforma-se em $(100)\ \frac{10}{40} = 25\%$. Assim, 25% desse grupo de 40 estudantes compõe-se de rapazes. Para ilustrar a utilidade das porcentagens nas comparações entre distribuições, vamos examinar o comparecimento de estudantes a exposições de pintura realizadas num "campus" universitário em que a arte moderna é considerada de importância fundamental. Suponhamos, então, que a galeria de arte moderna atraiu um bom número de estudantes, digamos 1.352, enquanto que a galeria de pintura clássica conseguiu atrair um número bem menor, digamos 183.

A Tabela 2.3 indica, ao mesmo tempo, as freqüências e as porcentagens de visitação a ambas as galerias. Note o leitor a dificuldade que surge ao tentar-se estabelecer rapidamente diferenças ligadas a sexo em termos de comparecimento, se o quadro de referência ficar restrito apenas às freqüências. Em contrapartida, as porcentagens revelam de forma clara que em ambas as galerias o elemento feminino foi igualmente representado. Especificamente, 20% dos estudantes presentes a ambas as galerias eram do sexo feminino.

---

evento, e o denominador, o número de casos *possíveis* de realização desse mesmo evento. Por exemplo, se numa urna há 10 bolas amarelas (A) e 40 bolas vermelhas (V), a probabilidade de retirar-se ao acaso, isto é, por sorteio, 1 bola vermelha é:

$$P(V) = \frac{\text{Número de casos favoráveis à "bola vermelha"}}{\text{Número de casos possíveis de "retirar 1 bola"}} = \frac{40}{10+40} = \frac{40}{50} = 0{,}8$$

De modo análogo,

$$P(A) = \frac{10}{50} = 0{,}2$$

$P(V)$ e $P(A)$ representam, respectivamente, a "probabilidade de ocorrer bola vermelha" e a "probabilidade de ocorrer bola amarela".
Note-se, finalmente, a analogia existente entre a expressão

$$P = \frac{f}{N}$$

e o quociente que dá a probabilidade de ocorrência de um evento.

Razões

Um método menos usado na padronização de tamanho denomina-se *razão*. Tal método compara diretamente o número de sujeitos que se enquadram numa categoria (por exemplo, homens) com o número de sujeitos que se enquadram noutra categoria (por exempo, mulheres). Assim, pode-se obter uma razão mediante a fórmula

$$\text{razão} = \frac{f_1}{f_2},$$

onde f_1 = freqüência de sujeitos de uma categoria e f_2 = freqüência de sujeitos de outra categoria.

Se estivéssemos interessados, por exemplo, na determinação da razão existente entre torcedores de futebol, poderíamos comparar o número de simpatizantes do time A (f = 150) com o número de simpatizantes do time B (f = 100), o que daria $\frac{150}{100}$. Simplificando a fração, isto é, cancelando os fatores comuns do numerador e denominador, resulta: $\frac{150}{100} = \frac{3}{2}$. (Em termos concretos, isto significa que para cada 3 torcedores do time A existem 2 torcedores do time B.)

O pesquisador poderia aumentar a clareza dessa razão se desse ao denominador uma forma que fosse mais inteligível. Por exemplo, a chamada *razão de sexo* comumente empregada por demógrafos, relaciona o número de homens em função de grupos de 100 mulheres. Com isso, esses pesquisadores procuram comparar a quantidade de homens com a quantidade de mulheres existentes num núcleo populacional determinado.

Ilustrando: se a razão de homens para mulheres for $\frac{150}{50}$, haverá 150 homens para 50 mulheres,. ou, simplificando, 3 homens para cada mulher. Para obter a versão convencional da razão de sexo, basta multiplicar o resultado acima por 100, donde:

$$\text{razão de sexo} = (100)\frac{f \text{ de homens}}{f \text{ de mulheres}} = \frac{(100)\ 150}{50} = 300.$$

Esse número está a nos dizer que nessa população particular existem 300 homens para cada 100 mulheres.

TABELA 2.3 Comparecimento de Alunos de Ambos os Sexos a uma Exposição de Pintura (Arte Clássica e Arte Moderna)

	Comparecimento às Exposições			
	Arte Moderna		*Arte Clássica*	
Sexo	*f*	*%*	*f*	*%*
Masculino	1.082	(80)	146	(80)
Feminino	270	(20)	37	(20)
Total	1.352	(100)	183	(100)

As razões (como as indicadas acima) já não são exaustivamente usadas em pesquisas—talvez pelos seguintes motivos:

1. São necessárias muitas razões para descrever distribuições que possuem várias categorias de análise.
2. Pode tornar-se difícil a comparação de razões baseadas em números muito grandes.
3. Alguns pesquisadores preferem evitar as frações geradas pelas razões.*

Taxas

Outro tipo de razão—que tende a ser muito mais utilizado pelos pesquisadores—é a *taxa*. Os sociólogos habitualmente analisam populações quanto a taxas de reprodução, mortalidade, crime, divórcio, casamento e fatos semelhantes. Todavia, enquanto que a maioria das outras razões compara o número de sujeitos de um subgrupo qualquer (categoria) com o número de sujeitos de qualquer outro subgrupo (categoria), as taxas indicam comparações entre o número *efetivo* de sujeitos e o número *potencial*. Por exemplo, para determinar a taxa de nascimento de uma dada população, poderíamos tomar o número *efetivo* de crianças nascidas vivas e cotejá-lo com o número de mulheres cujas idades fossem compatíveis com gravidez (ou seja, aquelas mulheres da população que estão expostas ao risco de gravidez e que, por isso, representam casos** potenciais). De modo análogo, para determinar a taxa de divórcio, poderíamos comparar o número efetivo de divórcios com o número de casamentos durante certo período (por exemplo, um ano). As taxas são freqüentemente dadas em termos de uma base de 1.000 casos potenciais. Assim, taxas de nascimentos representam o número de nascimentos por 1.000 mulheres; taxas de divórcio, o número de divórcios por 1.000 casamentos. Então, se ocorrerem 500 nascimentos entre 4.000 mulheres (de idade compatível com gravidez), vem:

$$\text{Taxa de nascimento} = (1.000)\,\frac{f \text{ de casos efetivos}}{f \text{ de casos potenciais}} = \frac{(1.000)\,500}{4.000} = 125$$

Disso resulta que para cada 1.000 mulheres passíveis de gravidez ocorrem 125 nascimentos de crianças vivas.

Até aqui discutimos taxas que podem ser úteis no estabelecimento de comparações entre diferentes populações. Por exemplo, poderíamos querer comparar taxas de nascimento entre pretos e brancos, entre mulheres de classe média e de classe baixa, entre grupos religiosos ou sociedades inteiras e assim por diante. Outro tipo de taxa,

* N.T.: Dos três motivos apresentados, o mais fraco é, sem dúvida, o de n.º 3, pois, como se verá mais adiante, não se faz praticamente nada em Estatística sem contínua recorrência à Teoria das Probabilidades. Ora, a não ser nos casos (raros) de *certeza de que ocorre* e *certeza de que não ocorre*, cujas probabilidades são, respectivamente, 1 e 0, todas as demais probabilidades são fracionárias e menores que 1 (porém positivas). Não será, pois, evitando as razões que evitar-se-ão complicações de cálculo causadas pelo surgimento de números fracionários.

** N.T.: Comumente aparece a expressão *casos* com o sentido de *sujeitos*.

a *taxa de mudança*, pode ser usada para comparar uma população com si própria em dois momentos distintos. Ao computar a taxa de mudança, comparamos a mudança efetiva ocorrida entre um tempo 1 e um tempo 2, onde a extensão do tempo 1 serve de base. Desse modo, uma população que cresceu de 20.000 a 30.000 habitantes entre 1960 e 1970 apresenta a seguinte taxa de mudança

$$\frac{(100) f \text{ de tempo } 2 - f \text{ de tempo } 1}{f \text{ de tempo } 1} = \frac{(100)\ 30.000 - 20.000}{20.000} = 50\%$$

Em outras palavras, houve um aumento populacional de 50% no período de 1960/70.

Observe-se que a taxa de mudança pode ser *negativa*, caso em que indicará a diminuição de algo num dado período. Exemplo: se uma população mudar de 15.000 para 5.000 habitantes num dado período a taxa de mudança será

$$\frac{(100)\ 5.000 - 15.000}{15.000} = -67\%$$

DISTRIBUIÇÕES DE FREQÜÊNCIAS SIMPLES RELATIVAS A DADOS ORDINAIS E INTERVALARES

Posto que dados nominais são "rotulados" em vez de graduados ou mensurados, as categorias das distribuições de nível nominal não precisam ser construídas em nenhuma ordem específica. Assim, os dados relativos a preferências religiosas que aparecem na Tabela 2.4 podem ser apresentados de três modos diferentes, embora igualmente aceitáveis.

TABELA 2.4 Distribuição de Preferências Religiosas. (Três Apresentações Diferentes.)

Religião	*f*	*Religião*	*f*	*Religião*	*f*
Protestantes	30	Católicos	20	Judeus	10
Católicos	20	Judeus	10	Protestantes	30
Judeus	10	Protestantes	30	Católicos	20
Total	60	Total	60	Total	60

Em contraste, as categorias ou os escores das distribuições ordinais ou intervalares representam o grau em que uma característica particular está presente. O arrolamento de tais categorias ou escores nas distribuições de freqüências simples deve ser feito de tal modo que reflita essa ordem.

Por essa razão, as categorias intervalares e ordinais são sempre dispostas em ordem, isto é, do valor mais alto ao mais baixo. Por exemplo, poderíamos arrolar as categorias componentes da variável classe social da mais alta à mais baixa (alta, média e baixa), ou poderíamos atribuir postos aos resultados de um exame semestral de Biologia, onde a ordem desses postos fosse seqüencial; portanto, do mais alto ao mais baixo.

TABELA 2.5 Distribuição de Freqüências Relativas a Atitudes Quanto à Guerra. Apresentações Correta e Incorreta

Atitude Quanto à Guerra	*f*	*Atitude Quanto à Guerra*	*f*
Ligeiramente favorável	2	Muito favorável	0
Um pouco desfavorável	10	Um pouco favorável	1
Muito favorável	0	Ligeiramente favorável	2
Ligeiramente desfavorável	4	Ligeiramente desfavorável	4
Muito desfavorável	21	Um pouco desfavorável	10
Um pouco favorável	1	Muito desfavorável	21
Total	38	Total	38
Incorreto		Correto	

Perturbar a ordem de categorias intervalares e ordinais reduz a legibilidade das descobertas do pesquisador. Tal efeito pode ser observado na Tabela 2.5, onde as versões "correta" e "incorreta" de uma distribuição de "Atitudes quanto à Guerra" foram apresentadas. Qual versão você acha mais fácil de interpretar?

DISTRIBUIÇÕES DE FREQÜÊNCIAS COM DADOS AGRUPADOS

Os escores de nível intervalar espalham-se, às vezes, ao longo de uma extensa *amplitude* (valor maior menos valor menor), o que toma a distribuição de freqüências resultante não só longa, mas, também, difícil de ler. Quando ocorrerem tais situações, poucos sujeitos poderão estar em correspondência com cada um dos escores, donde se segue que o modelo do grupo torna-se indefinido. Ilustrando: a distribuição apresentada na Tabela 2.6 contém valores que variam de 50 a 99 e estende-se por quase quatro colunas.*

Para tomar mais clara nossa apresentação, poderíamos construir uma *distribuição de freqüências agrupadas,* condensando os escores separados num número menor de categorias ou grupos, cada um contendo mais de um escore. Cada categoria ou grupo numa distribuição assim condensada recebe o nome de *intervalo de classe*, cujo *tamanho* é determinado pela quantidade de escores nele contidos.

As notas de exame dos 71 alunos originalmente apresentados na Tabela 2.6 estão redispostas numa distribuição de freqüências agrupadas na Tabela 2.7. Aparecem, aí, 10 intervalos de classe, cada um de tamanho 5. Desse modo, o intervalo de classe mais alto (95-99) contém cinco valores, a saber: 95, 96, 97, 98 e 99. Da mesma forma, o intervalo 70-74, também de tamanho 5, contém os escores 70, 71, 72, 73 e 74.

Limites de Classe

De acordo com seu tamanho, cada intervalo de classe tem um *limite superior* e um *limite inferior*. À primeira vista, os limites superior e inferior de uma dada categoria *parecem* ser os *limites reais*. Assim, poderíamos com certa "lógica" imaginar que os limites superior e inferior do intervalo 60-64 fossem 64 e 60, respectivamente. Se assim procedêssemos, todavia, estaríamos *errados*, uma vez que 60 e 64 não são *de fato* os limites do intervalo 60-64.

Muitos leitores estarão perguntando: por que não? Para responder, vamos examinar um problema que poderia surgir se tivéssemos de definir limites de classe em termos do escore mais alto e mais baixo de um intervalo qualquer. Suponham que tentássemos localizar, num intervalo de classe, números dotados de uma parte fracionária (frações decimais);* tome-se por base a distribuição de freqüência apresentada na Tabela 2.7. Onde colocaríamos o escore 62,3? A maioria de nós concordaria que tal

* N.T.: Obviamente, se o autor reduzisse a altura das colunas de sua tabela, ele poderia "dar mais ênfase" ao que disse acima, pois, aí, os escores de 50 a 99 ocupariam, quem sabe (?!), oito colunas. O argumento é logicamente fraco.

TABELA 2.6 Notas de Exame Final de 71 Estudantes Dispostas numa Distribuição de Freqüências

Nota	*f*	*Nota*	*f*	*Nota*	*f*	*Nota*	*f*
99	0	85	2	71	4	57	0
98	1	84	1	70	9	56	1
97	0	83	0	69	3	55	0
96	1	82	3	68	5	54	1
95	1	81	1	67	1	53	0
94	0	80	2	66	3	52	1
93	0	79	8	65	0	51	1
92	1	78	1	64	1	50	1
91	1	77	0	63	2	Total	71
90	0	76	2	62	0		
89	1	75	1	61	0		
88	0	74	1	60	2		
87	1	73	1	59	3		
86	0	72	2	58	1		

TABELA 2.7 Notas de Exame Final de 71 Estudantes Redispostas numa Distribuição de Freqüências Agrupadas

Intervalo de Classe	*f*
95-99	3
90-94	2
85-89	4
80-84	7
75-79	12
70-74	17
65-69	12
60-64	5
55-59	5
50-54	4
Total	71

* N.T.: É curioso que o autor, ao referir-se a números dotados de uma parte fracionária, esclareça

valor pertence ao intervalo 60-64. Mas que dizer do escore 69,4? E que tal tentar situar 54,2 ou 94,6? O leitor poderia notar que os escores mais alto e mais baixo de um intervalo qualquer deixam lacunas (espaços vazios) entre grupos numéricos adjacentes, de sorte que alguns valores fracionários não podem ser incluídos em nenhum intervalo de classe da distribuição, donde decorre que eles devem simplesmente ser excluídos.

Contrariamente ao que sugerem o escore mais alto e mais baixo de um intervalo qualquer, os *limites de classe* estão localizados num ponto que se situa a meio caminho entre as classes intervalares adjacentes e, portanto, servem para preencher as lacunas existentes (ver Figura 2.1). Assim, o limite superior do intervalo 90-94 é 94,5 e o limite

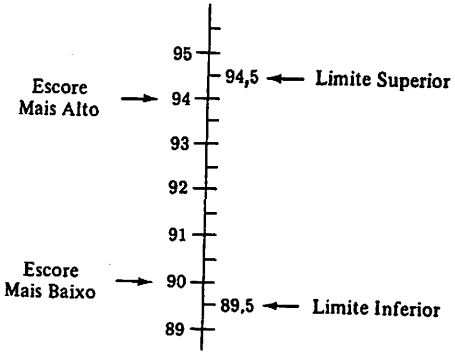


FIGURA 2.1 Escores Mais Alto e Mais Baixo Versus Limites Superior e Inferior da Classe Intervalar 90-94

inferior do intervalo 95-99 também é 94,5. Da mesma forma, 59,5 tanto serve de limite superior do intervalo 55-59 quanto do limite inferior do intervalo 60-64. O leitor poderia perguntar o que fazer com o valor 59,5, uma vez que ele cai *exatamente* no meio do espaço que separa duas classes intervalares adjacentes. Esse escore deveria ser incluído no intervalo 55-59 ou no intervalo 60-64? Este problema é geralmente contornado mediante *arredondamento para o inteiro par mais próximo*. Por exemplo: 59,5 seria colocado no intervalo 60-64; 84,5, no intervalo 80-84.* Como teremos oportuni-

tratarem-se de "frações decimais". Se, de um lado, o que ele diz não está errado, de outro, não chega a estar completo, uma vez que a *notação decimal* é uma forma de exprimir frações, mas não a *única*. Alternativamente, existe a notação *ordinária*, expressa por numerador-traço-denominador, capaz de, em muitos casos, atingir requintes de precisão que superam a notação decimal. Exemplificando: 2,25 (com notação decimal) é o mesmo que $2 + \frac{25}{100}$; entretanto 0,333 ... jamais será um dado tão preciso quanto 1/3.

* N.T.: Em Matemática, ocorrem com freqüência as seguintes notações de intervalos: —, ⊢⊣, ⊢, ⊣. Esses sinais significam, respectivamente: intervalo *aberto*; intervalo *fechado*; intervalo *fechado à esquerda* (ou *aberto à direita*); intervalo *fechado à direita* (ou *aberto à esquerda*). Por uma questão de rigor, sugere-se que os intervalos que aparecem no corpo da obra notados com "—" sejam substituídos por "⊢", dirimindo-se assim eventuais dúvidas. Com essa nova simbologia, os intervalos 95-99 e 90-94 passam a ser designados por 95 ⊢ 99 e 90 ⊢ 94, respectivamente.

dade de ver, a posição dos limites de classe deve ser estabelecida a fim de ser possível a aplicação de certos procedimentos estatísticos.

Ponto Médio

Outra característica de qualquer intervalo de classe é o seu *ponto médio*, que se define como o valor "mais central" desse mesmo intervalo. Um método simples e rápido de achar o ponto médio é procurar o valor do intervalo de classe que se situa exatamente no meio. Alguns exemplos: 50 é o ponto médio do intervalo 48-52; 3,5 é o ponto médio do intervalo 2-5. O ponto médio também pode ser calculado mediante uma fórmula* que relaciona os limites superior e inferior de um dado intervalo. Por exemplo, o ponto médio do intervalo 48-52 é

$$\frac{\text{limite inferior} + \text{limite superior}}{2} = \frac{48 + 52}{2} = 50$$

Determinação do Número de Intervalos

Para apresentar dados intervalares sob a forma de distribuição de freqüências agrupadas, o pesquisador deve levar em conta o número de categorias (classes) que ele deseja empregar. Os textos geralmente recomendam usar não menos que 5 nem mais que 20 intervalos. No tocante a este aspecto, cumpre lembrar que as distribuições de freqüências agrupadas são usadas com o intuito de revelar ou enfatizar um modelo grupal.** Intervalos de classe em excesso ou em escassez podem tornar esse modelo menos inteligível e, com isso, trabalhar contra o pesquisador—que procura aumentar a clareza de sua análise. Além disso, reduzir os escores individuais a um número desnecessariamente pequeno de intervalos, pode redundar em sacrifício de muita precisão—justamente a precisão que se conseguiu no início mediante o conhecimento da identidade dos escores individuais da distribuição. Em suma, portanto, o pesquisador geralmente toma sua

* N.T.: Tal como foi explicado nos tópicos anteriores, os números que aparecem nos vários intervalos de classe constituem seus *limites aparentes*. Então, obriga-nos o rigor matemático a respeitar—e *usar* nos cálculos—os limites reais de cada intervalo de classe. Assim, o ponto médio do intervalo 48-52 (que, segundo sugestão anterior, deveria grafar-se 48 ⊢ 52) não é 50, mas, sim, 49,5, a menos que os valores 48 e 52 sejam resultantes de uma *contagem*, pois, nesse caso, a variável *é discreta* (ou *descontínua*), e só admite *valores inteiros*. A fórmula que permite, com rigor matemático, calcular o ponto médio é: (limite inferior aparente − 0,5) + (metade da amplitude do intervalo). Em símbolos:

l = limite inferior aparente
L = limite superior aparente
h = (L − l) = amplitude do intervalo

$$\text{ponto médio} = (l - 0{,}5) + \frac{h}{2}$$

Exemplo: achar o ponto médio do intervalo 48 ⊢ 52.

$$\text{Ponto médio} = (48 - 0{,}5) + \frac{4}{2} = 47{,}5 + 2 = 49{,}5.$$

** N.T.: Entenda-se por "modelo", aqui, "tendência".

decisão quanto ao número de intervalos baseado em seu próprio conjunto de dados e em seus objetivos pessoais, fatores que podem variar consideravelmente de uma situação de pesquisa para outra.

DISTRIBUIÇÕES CUMULATIVAS

É desejável, algumas vezes, apresentar freqüências de forma cumulativa, especialmente quando procuramos situar a posição de um sujeito em função do desempenho total do grupo. Definem-se *freqüências cumulativas* como o número total de sujeitos portadores de um dado escore ou como o número total de sujeitos portadores de *um escore menor que outro previamente especificado.* Assim, a freqüência acumulada, ou cumulativa (*fa*), de qualquer categoria (ou classe intervalar) obtém-se através da soma da freqüência daquela categoria com a freqüência total de todas as categorias que estão abaixo dela. No caso dos pontos obtidos pelos universitários (Tabela 2.8), observamos que a freqüência *f* associada ao intervalo de classe 301-350 é 12. Tal freqüência (12) corresponde, também, à freqüência acumulada relativa ao intervalo 301-350, uma vez que nenhum membro do grupo fez menos pontos do que 301. A freqüência* da classe intervalar

TABELA 2.8 **Distribuição de Freqüências Acumuladas Variável: Escores de Desempenho Acadêmico Obtidos por 336 Universitários**

Intervalo de Classe	*f*	*fa*
751-800	6	336
701-750	25	330
651-700	31	305
601-650	30	274
551-600	35	244
501-550	55	209
451-500	61	154
401-450	48	93
351-400	33	45
301-350	12	12
Total	336	

seguinte (351-400) é 33, enquanto que a freqüência acumulada para o mesmo intervalo é 45 (isto é, 33 + 12). Essas informações permitem-nos tomar conhecimento, portanto, de que 33 universitários situaram-se na faixa 351-400, e de que 45 receberam escores iguais ou inferiores a 400. Poderíamos continuar com este procedimento e obter freqüências acumuladas para todos os intervalos de classe, até chegarmos ao intervalo de maiores limites, 751-800, cuja freqüência acumulada (336) é igual à totalidade de sujeitos–uma vez que nenhum membro do grupo conseguiu mais que 800 pontos.

* N.T.: É comum falar-se em freqüência *simples* para reduzir a possibilidade de confusão com freqüências *acumuladas.*

Além da freqüência acumulada, podemos também construir uma distribuição que indica a *porcentagem acumulada* (*c%*), ou seja, a porcentagem de sujeitos que obtiveram determinado número de pontos ou, dito de outra forma, a porcentagem de sujeitos que obtiveram, no máximo, determinado número de pontos. Calcula-se a porcentagem acumulada através da modificação da fórmula de porcentagem (%) apresentada páginas atrás neste mesmo capítulo:

$$c\,\% = (100)\,\frac{fa}{N}\ \text{, onde}$$

fa = freqüência acumulada para qualquer categoria (classe);
N = total de sujeitos (casos) da distribuição.

Se aplicarmos a fórmula acima aos dados da Tabela 2.8, verificaremos que a porcentagem de alunos que obtiveram 350 ou menos foi

$$c\,\% = (100)\,\frac{12}{336} = (100)(0{,}0357) = 3{,}57\%$$

A porcentagem dos que conseguiram 400 ou menos foi

$$c\,\% = (100)\,\frac{45}{336} = (100)(0{,}1339) = 13{,}39\%$$

A porcentagem dos que obtiveram 450 ou menos foi

$$c\,\% = (100)\,\frac{93}{336} = (100)(0{,}2768) = 27{,}68\%$$

A distribuição de porcentagens acumuladas baseada nos dados da Tabela 2.8 figura na Tabela 2.9.

POSTO PERCENTIL

Suponhamos que você tenha obtido 80 num exame de Estatística. A fim de determinar com exatidão o seu grau de desempenho, seria provavelmente útil saber como a sua nota 80 compara-se com as notas dos demais colegas de classe que passaram pela mesma prova. A maioria de seus colegas ficou em torno de 80 ou 90? Se assim foi, a sua classificação (nota) pode não ter sido "terrivelmente alta"! Por outro lado, se a maioria de seus colegas não tirou mais que 60 ou 70, a nota 80 pode muito bem situar-se entre as mais altas do grupo.

Com o auxílio da distribuição de porcentagens acumuladas, podemos fazer comparações precisas entre qualquer sujeito e o grupo ao qual ele pertence. Especificamente, podemos calcular o *posto percentil* de um escore, que nada mais é que um simples número indicativo da porcentagem (total) de sujeitos de uma distribuição que conseguiram menos que um dado valor. Por exemplo, se ao escore 80 corresponde o posto percentil 95, então 95% dos alunos receberam, nesse exame de Estatística, escores inferiores a 80 (apenas 5% obtiveram valores acima de 80). Entretanto, se a um escore de 80 corresponder um posto percentil de 45, então somente 45% receberam notas de exame abaixo de 80 (donde decorre que 55% conseguiu notas superiores). Em símbolos:

$$\text{Posto percentil} = \begin{pmatrix} c\% \text{ abaixo} \\ \text{do limite} \\ \text{inferior do} \\ \text{intervalo} \\ \text{crítico} \end{pmatrix} + \left[\left(\frac{\text{escore} - \begin{pmatrix}\text{limite inferior} \\ \text{do intervalo} \\ \text{crítico}\end{pmatrix}}{\text{tamanho do intervalo crítico}} \right) \cdot \begin{pmatrix} \% \text{ no} \\ \text{intervalo} \\ \text{crítico} \end{pmatrix} \right]$$

Para ilustrar o procedimento, vamos calcular, relativamente aos dados que aparecem na Tabela 2.8, o posto percentil do escore 620. Antes de aplicar a fórmula,

TABELA 2.9 Distribuição de Porcentagens Acumuladas. Variável: Escores de Desempenho Acadêmico Obtidos por 336 Estudantes Universitários. (Esta Tabela Baseia-se na de n.º 2.8)

Intervalo de Classe	*fa*	*c%*
751-800	336	100%
701-750	330	98,21
651-700	305	90,77
601-650	274	81,55
551-600	244	72,62
501-550	209	62,20
451-500	154	45,83
401-450	93	27,68
351-400	45	13,39
301-350	12	3,57

devemos primeiro localizar o *intervalo crítico*, que corresponde ao intervalo de classe onde o escore 620 está contido. Tal como demonstrado abaixo, o intervalo crítico deste problema é 601-650:

Intervalo de Classe	
751-800	
701-750	
651-700	
601-650	← Intervalo de classe que contém o escore 620
551-600	
501-550	
451-500	
401-450	
351-400	
301-350	

Há várias características relativas ao intervalo crítico que devem ser determinadas *antes* da aplicação da fórmula do posto percentil:

1. O limite inferior do intervalo crítico—que corresponde ao ponto situado *entre* o intervalo crítico 601-650 e o intervalo imediatamente abaixo dele, vale dizer: 551-600. Neste caso, o limite inferior (*real*) do intervalo 601-650 é 600,5.
2. O tamanho do intervalo* crítico—que se determina a partir do número de escores contidos nesse intervalo. No caso do intervalo 601-650, a *amplitude* é 50, uma vez que entre 601 e 650 há 50 valores.**
3. A porcentagem dentro do intervalo crítico—que se obtém dividindo o número de sujeitos contidos nesse intervalo particular (f), pelo número total de sujeitos da distribuição (N), e multiplicando a resposta por 100. Em símbolos:

$$\% = (100)\frac{f}{N} = (100)\frac{30}{336} = (100)(0{,}089) = 8{,}93\%$$

Vemos, portanto, que 8,93% desses escores obtidos pelos universitários caem dentro do intervalo de classe 601-650.

4. A porcentagem acumulada abaixo do limite inferior do intervalo crítico. O valor de $c\%$ pode ser lido diretamente na Tabela 2.9, que fornece a distribuição das porcentagens acumuladas. Ao percorrermos a coluna da $c\%$ de baixo para cima, observamos que 72,62% dos escores ficam *abaixo* do intervalo crítico; esta é, pois, a porcentagem acumulada associada ao intervalo de classe que se situa imediatamente abaixo do intervalo crítico.

Estamos agora preparados para aplicar a fórmula do posto percentil:

$$\text{Posto Percentil} = 72{,}62 + \left[\left(\frac{620 - 600{,}5}{50}\right)(8{,}93)\right] = 72{,}62 + \left[\left(\frac{19{,}50}{50}\right)(8{,}93)\right]$$

$$= 72{,}62 + (0{,}39)(8{,}93) = 72{,}62 + 3{,}48 = 76{,}10$$

Resulta, então, que pouco mais do que 76% recebeu escores inferiores a 620. Somente 23,90% obteve valores superiores.

* N.T.: O *tamanho do intervalo* é também chamado *amplitude* (amplitude da classe, amplitude do intervalo). À diferença entre o *maior valor* e o *menor valor* de uma distribuição dá-se o nome de *amplitude total*.

** N.T.: Observe-se que a amplitude 50 decorre da aplicação da seguinte fórmula:

$(L - \ell) + 1 = (650 - 601) + 1 = 49 + 1 = 50.$

Na verdade, trata-se, aqui, da aplicação da formula que permite calcular o número de termos de uma progressão aritmética, ou seja:

$$n = \frac{a_n - a_1}{r} + 1, \text{ onde } a_n = L, a_1 = \ell, r = 1$$

Como exemplo adicional, vamos calcular o posto percentil relativo ao escore 92 da seguinte distribuição:

Intervalo de Classe	*f*	*fa*	*c%*
90-99	6	49	100%
80-89	8	43	87,76
70-79	12	35	71,43
60-69	10	23	46,94
50-59	7	13	26,53
40-49	6	6	12,24
	N = 49		

Conforme a ilustração abaixo, o intervalo crítico para um escore de 92 é 90-99:

Intervalo de Classe	
90-99	← Intervalo de classe ao qual pertence o valor 92
80-89	
70-79	
60-69	
50-59	
40-49	

As características do intervalo crítico que devemos determinar são as seguintes:

1. O limite inferior do intervalo crítico é 89,5.
2. O tamanho (amplitude) do intervalo crítico é 10, já que há 10 escores de 90 a 99 (90, 91, 92, 93, 94, 95, 96, 97, 98, 99).*
3. A porcentagem dentro do intervalo crítico é 12,24%. Em símbolos:

$$\% = (100)\,\frac{f}{N},\ \text{donde}$$

$$\% = (100)\,\frac{6}{49} = (100)\,0{,}1224 = 12{,}24\%$$

4. A porcentagem acumulada abaixo do limite inferior do intervalo crítico pode ser encontrada na coluna *c%*, em correspondência ao intervalo de classe imediatamente abaixo do já citado intervalo crítico. A porcentagem acumulada que corresponde à classe intervalar 80-89 é 87,76%.

*N.T.: Veja-se, mais uma vez, a nota de rodapé anterior. Mediante sua aplicação, resulta, de imediato, o valor 10, pois: $(L - \ell) + 1 = (99 - 90) + 1 = 10$.

Estamos, agora, em condições de substituir esses dados na fórmula do posto percentil:

$$\text{Posto percentil} = 87{,}76 + \left[\left(\frac{92 - 89{,}5}{10}\right)(12{,}24)\right] = 87{,}76 + \left[\left(\frac{2{,}50}{10}\right)(12{,}24)\right]$$

$$= 87{,}76 + (0{,}25)(12{,}24) = 87{,}76 + 3{,}06 = 90{,}82\%$$

Quase 91% dos alunos, receberam notas menores que 92. Somente 9,18% receberam notas superiores a 92.*

A escala de postos percentis consiste em 100 unidades. Certos postos percentis situados ao longo dessa escala têm nomes específicos. Os *decis* dividem a escala em porções de dez. Assim, se um escore estiver situado no primeiro decil (posto percentil = 10), sabemos que 10% dos casos situam-se abaixo desse escore; se um escore estiver no segundo decil (posto percentil = 20), então 20% dos casos estarão abaixo dele e assim por diante. Os pontos que dividem a escala de postos percentis em grupos de 25 (isto é, 25%) são conhecidos pelo nome de *quartis.*** Se um escore estiver localizado no primeiro quartil (posto percentil = 25), sabemos que 25% dos casos caem abaixo dele; se um escore estiver no segundo quartil (posto percentil = 50), então 50% dos

Posto Percentil	*Decil*	*Quartil*
90 =	9.º	
85		
80 =	8.º	
75 =		3.º
70 =	7.º	
65		
60 =	6.º	
55		
50 =	5.º	2.º
45		
40 =	4.º	
35		
30 =	3.º	
25 =		1.º
20 =	2.º	
15		
10 =	1.º	

FIGURA 2.2. **Escala de Postos Percentis Dividida em Decis e Quartis**

* N.T.: Observe-se que 91%+ 9,18% > 100%. Isto porque o autor arredondou para mais a porcentagem 90,82–anteriormente encontrada.

** N.T.: Os sujeitos de uma distribuição que se localizam entre o primeiro e o terceiro quartis pertencem ao que se convencionou chamar de *intervalo de normalidade.* Apesar de tratar-se de escores estarão situados abaixo dele; por igual raciocínio, para um escore que esteja no terceiro quartil (posto percentil = 75), haverá 75% de escores abaixo dele (ver Figura 2.2).

RESUMO

Foram introduzidas, neste capítulo, algumas das técnicas básicas usadas pelo pesquisador na organização do emaranhado de dados originais (brutos) que ele coleta de seus respondentes. Distribuições de freqüência e métodos para comparar essas distribuições–quando compostas de dados nominais (proporções, porcentagens, razões e taxas)–foram discutidos e ilustrados. Examinaram-se, também, as características de distribuições de freqüências simples, agrupadas e acumuladas com relação a dados ordinais e intervalares. Finalmente, apresentou-se o procedimento que permite obter o posto percentil de um escore bruto.

PROBLEMAS

1. Usando os dados do quadro abaixo, onde aparecem telespectadores e não-telespectadores, reclassificados quanto ao seu grau de realização, calcular: (a) a porcentagem dos que não são telespectadores, porém grandes realizadores; (b) a porcentagem dos telespectadores que são grandes realizadores; (c) a proporção de não-telespectadores que são grandes realizadores; (d) a proporção de telespectadores que são grandes realizadores.

Realização versus Assistência a Programas de Televisão

	Situação	
Realização	*Não-telespectadores* *f*	*Telespectadores* *f*
Alta realização	93	46
Baixa realização	90	127
Total	183	173

2. O quadro seguinte representa (para uma dada região) a estrutura familiar de crianças pretas e crianças brancas.

"convenção", há uma lógica defensável por trás disso: hão de ser *normais* os sujeitos que, relativamente a uma dada variável, ocupem, pelo menos, 50% da distribuição. Esses 50% devem ocupar o "centro" da distribuição, ficando os 50% restantes divididos em duas porções de 25%, uma à esquerda, outra à direita. Os 25% da esquerda são chamados *subnormais* e os 25% da direita, *supernormais*. Finalmente, chamando o 1.º quartil de Q_1, e o 3.º quartil de Q_3, define-se a *amplitude do intervalo de normalidade* pela diferença $(Q_3 - Q_1)$.

Estrutura Familiar de Crianças Pretas e Crianças Brancas

	Raça da Criança	
Estrutura Familiar	*Preta* *f*	*Branca* *f*
Um genitor	53	59
Dois genitores	130	167
Total	183	226

Calcular: (a) a porcentagem de crianças pretas que possuem dois genitores; (b) a porcentagem de crianças brancas que possuem dois genitores; (c) a proporção de crianças negras que possuem dois genitores; (d) a proporção de crianças brancas que têm dois genitores.

3. Num grupo de 4 sujeitos de alta realização e 24 de baixa realização, qual é a razão dos primeiros para os segundos?
4. Num grupo de 125 machos e 80 fêmeas, qual é a razão macho/fêmea?
5. Num grupo de 15 crianças negras e 20 brancas, qual é a razão negros/brancos?
6. Se 300 crianças nascem vivas entre 3.500 mulheres (em idade compatível com gravidez), qual é a taxa de nascimento?
7. Qual é a taxa de mudança de uma população que cresce de 15.000 habitantes em 1950 a 25.000 habitantes em 1970?
8. Transforme a seguinte distribuição de escores numa distribuição de freqüências agrupadas com quatro intervalos de classe e (a) determine o tamanho dos intervalos de classe; (b) indique os limites superior e inferior de cada classe; (c) identifique o ponto médio de cada intervalo de classe; (d) calcule a freqüência acumulada para cada intervalo de classe; (e) calcule a porcentagem acumulada para cada intervalo de classe.

Escores	*f*
12	3
11	4
10	4
9	5
8	6
7	5
6	4
5	3
4	2
3	1
2	1
1	2
	$N =$ 40

9. Na seguinte distribuição de escores, calcule o (posto)* percentil relativo a (a) 75; (b) 52.

Intervalo de Classe	*f*	*fa*
90-99	6	48
80-89	9	42
70-79	10	33
60-69	10	23
50-59	8	13
40-49	5	5
	$N =$ 48	

10. Na seguinte distribuição de escores, calcular o (posto) percentil para (a) 36; (b) 18.

Intervalo de Classe	*f*
40-44	5
35-39	5
30-34	8
25-29	9
20-24	10
15-19	8
10-14	6
5-9	5
	$N =$ 56

* N.T.: São sinônimas as expressões *percentil* e *posto percentil*.

3

Representações Gráficas

Colunas numéricas já conseguiram despertar medo, aborrecimento, apatia e mal-entendidos. Enquanto que algumas pessoas parecem "desligar-se" ao serem expostas a informações estatísticas em forma tabular, numérica, elas podem prestar bastante atenção às mesmas informações apresentadas em forma gráfica ou pictórica. Daí resulta que muitos pesquisadores e autores preferem usar gráficos em substituição a tabelas numéricas. Razões semelhantes levam pesquisadores a utilizar, comumente, gráficos tais como os *setoriais*, gráficos de *barras*, *polígonos de freqüência* etc., num evidente esforço de tornar mais legíveis as suas investigações.

GRÁFICOS SETORIAIS

O gráfico *setorial* (ou de setor) é um dos mais simples recursos gráficos, posto que consiste em um círculo cujos setores (isto é, partes do mesmo círculo) somam 100%. Os gráficos setoriais são particulármente úteis quando se trata de visualizar diferenças freqüenciais entre algumas categorias de nível nominal. Ilustrando: A Figura 3.1 representa a origem urbana, suburbana ou rural de uma população de 2.000 universitários. Note-se que 70% desses alunos vêm de áreas suburbanas, enquanto que apenas 12% de áreas rurais.*

* N.T.: Não se tome ao pé da letra essa repartição porcentual, uma vez que (a) a sociedade-referência é a norte-americana; (b) essas porcentagens referem-se a um *particular universo* de universitários. Nada garante, portanto, que mesmo nos Estados Unidos todas as universidades apresentem igual repartição de clientela.

Origem dos Alunos	*f*	*%*
Urbana	240	12
Suburbana	1.400	70
Rural	360	18
Total	2.000	100

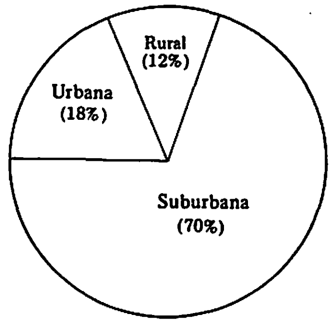


FIGURA 3.1 Origem Urbana, Suburbana e Rural de uma População de 2.000 Universitários

GRÁFICOS DE BARRAS

O gráfico de setores ilustra, de modo fácil e rápido, dados que podem ser subdivididos em algumas categorias. Comparativamente, o *gráfico de barras* (ou *histograma*) pode "acomodar" qualquer quantidade de categorias de qualquer nível de mensuração e, por isso, é muito mais usado em pesquisas.

Examinemos o gráfico de barras, Figura 3.2, que ilustra a distribuição de freqüência de classes sociais. Esse gráfico é construído de acordo com o seguinte procedimento:

Classe Social	*f*
A	5
B	14
C	23
D	45
E	38
F	25
Total	150

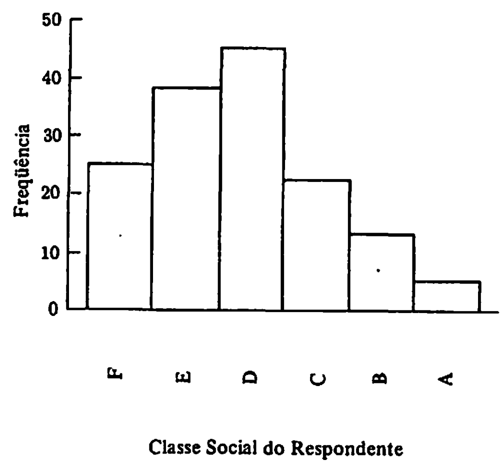


FIGURA 3.2 Gráfico de Barras Ilustrativo de uma Distribuição de Classes Sociais*

* N.T.: A classificação mais usual no Brasil é a da ABA-ABIPEME. Essa classificação distribui os indivíduos em cinco classes sócio-econômicas – A, B, C, D, E –, onde A é a classe mais alta e E, a mais baixa. Ver os itens e a pontuação indicados no Apêndice D.

Ocupação (= Profissão)	*f*
A = Artesanato	52
B = Trabalho não qualificado	65
C = Gerencial	29
D = Serviços burocráticos	34
Total	180

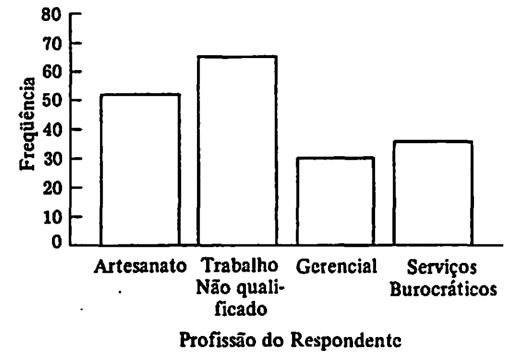


FIGURA 3.3 Gráfico de Barras Ilustrativo de uma Distribuição Ocupacional

na linha de base horizontal (ou eixo dos *X*) marcam-se os valores ou as categorias (no presente exemplo são classes sociais); na linha vertical (ou eixo dos *Y*) localizam-se as freqüências relativas a cada categoria ou a cada valor. (No caso de dados agrupados, são os pontos médios dos intervalos de classe que se colocam no eixo dos *X*.) Note-se que as freqüências de cada valor (da variável) são indicadas por barras retangulares, e de conformidade com o seguinte critério: quanto mais alta a barra, maior a freqüência de ocorrência.

Na Figura 3.2 as barras retangulares foram unidas a fim de enfatizarem-se as diferenças entre os vários graus de *status* representados por diferentes classes sociais. Acrescente-se a isso o seguinte: as classes sociais foram localizadas no eixo horizontal em *ordem crescente*, isto é, da mais baixa para a mais alta. Esse, aliás, é o critério convencional adotado na construção de gráficos de barras em que a variável de observação é de nível ordinal ou de nível intervalar.

Nos gráficos de barras em que a variável observacional é *categórica*, isto é, de nível nominal, as barras devem apresentar-se *separadas*, pois, do contrário, sua união poderá implicar continuidade. Além disso, as categorias resultantes de variáveis nominais podem ser dispostas em qualquer ordem ao longo do eixo horizontal. O gráfico de barras da Figura 3.3 ilustra bem essa característica.

POLÍGONOS DE FREQÜÊNCIAS

Outra representação gráfica comumente usada é o *polígono de freqüências*. Muito embora o polígono de freqüências tenha flexibilidade bastante para ajustar-se a uma ampla variedade de situações, ele tende a sugerir muito mais *continuidade* do que *discriminação*; daí sua particular utilidade na representação de dados ordinais ou intervalares. Isto acontece porque as freqüências são indicadas por uma série de pontos colocados exatamente acima (numa vertical) dos valores constantes do eixo dos *X* ou, quando for o caso, exatamente acima dos valores que representam os pontos médios dos intervalos de classe. Esses pontos são, a seguir, unidos por segmentos de retas (donde o nome "polígonos"), sendo que, em ambas as extremidades, a linha poligonal, graças a um prolongamento convencional, toca o eixo dos X.* Como bem demonstra a Figura 3.4, a altura de cada ponto indica a freqüência com que cada valor particular ocorre.

Intervalo de Classe	*f*
136-145	11
126-135	16
116-125	29
106-115	40
96-105	44
86-95	25
76-85	13
Total	178

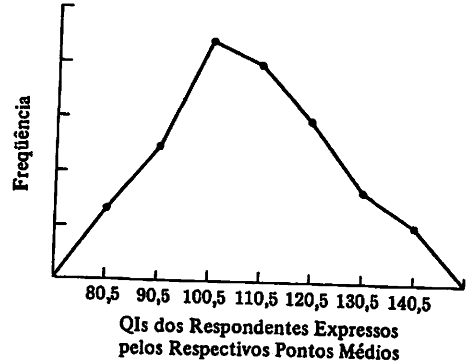
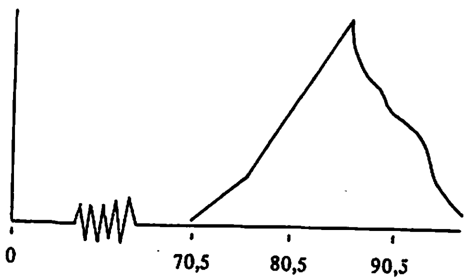


FIGURA 3.4 Polígono de Freqüências de uma Distribuição de QIs**

Quando o problema é representar graficamente freqüências acumuladas (ou porcentagens acumuladas), é possível construir o chamado *polígono de freqüências acumuladas****. Na Figura 3.5, as freqüências acumuladas estão marcadas no eixo vertical e são indicadas pela altura, isto é, pela distância de cada ponto até a linha de base. Contrariamente ao que ocorre com um polígono de freqüências (ver Fig. 3.4),

* N.T.: Prolongar a linha poligonal em ambas as extremidades é um recurso que permite "aterrizar" o desenho, que, de outra forma, ficaria como que "solto" no quadrante determinado pelas coordenadas cartesianas ortogonais. O critério geralmente aceito para essa "aterrizagem" é o seguinte: unir a extremidade esquerda da linha poligonal ao valor teórico que antecede o primeiro valor real da tabela, e a extremidade direita ao valor teórico que sucede o último valor real da mesma tabela.

**N.T.: É prática não só comum, mas também bastante aconselhável, não unir a extremidade esquerda da linha poligonal à origem zero do sistema de coordenadas. Isto porque, se houvesse um valor anterior a 80,5, com certeza ele seria 70,5 e não 0 (zero). Então, contorna-se o problema "sanfonando" uma parte do eixo das abscissas, conforme indicado abaixo:



***N.T.: O polígono de freqüências acumuladas também é conhecido por *ogiva de Galton.*

Intervalo de Classe	*f*	*fa*
751-800	6	336
701-750	25	330
651-700	31	305
601-650	30	274
551-600	35	244
501-550	55	209
451-500	61	154
401-450	48	93
351-400	33	45
301-350	12	12
	N = 336	

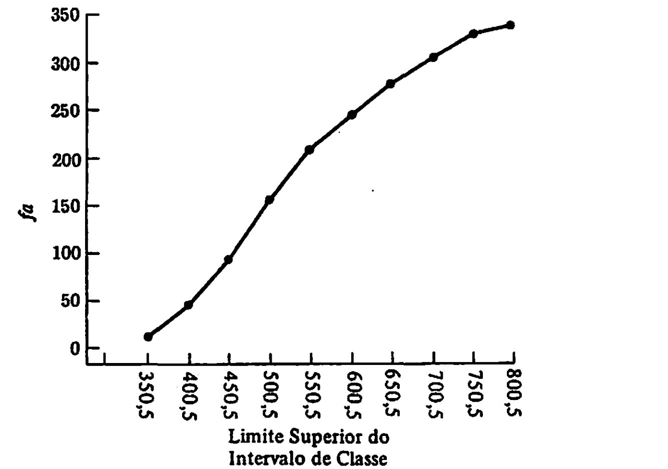


FIGURA 3.5 Polígono de Freqüências Acumuladas. Dados da Tabela 2.8

a linha poligonal que une os pontos num gráfico de freqüências acumuladas não pode ser "amarrada" ao eixo das abcissas. A razão desse fato é que estas freqüências representam um conjunto de adições sucessivas. Uma freqüência acumulada qualquer nunca é menor (e geralmente é maior) do que a freqüência acumulada precedente. Acrescente-se, ainda, que os pontos num gráfico cumulativo são marcados sobre perpendiculares levantadas a partir dos limites superiores de cada intervalo de classe, coisa que não ocorre com o polígono de freqüências comum, onde as marcações têm como referência os pontos médios das classes. Isto explica-se pelo fato de que qualquer freqüência acumulada representa o número total de sujeitos compreendidos, ao mesmo tempo, *dentro* e *abaixo* de um intervalo de classe particular.

CONSTRUÇÃO DE GRÁFICOS DE BARRAS E DE POLÍGONOS DE FREQÜÊNCIAS

As regras e procedimentos apresentados a seguir podem ser aplicados tanto à construção de diagramas de barras quanto a de polígonos de freqüências:

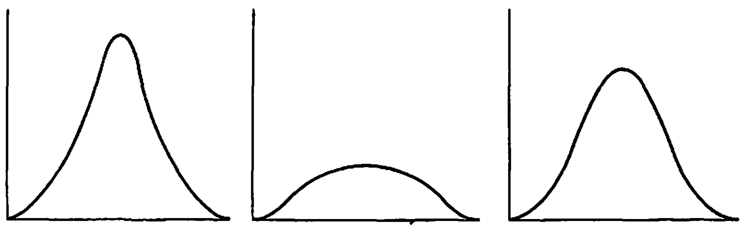
1. Por questão de tradição e para evitarem-se confusões, o pesquisador sempre dispõe os valores mensurados (escores) ao longo do eixo horizontal, e as freqüências (ou as porcentagens) no eixo vertical.*
2. Todos os componentes de um gráfico devem ter nome. A linha de base horizontal costuma representar uma característica** (por exemplo, a idade do

* N.T.: Isso nem sempre é assim. Talvez fique mais fácil (ou prático) estabelecer o seguinte: no eixo das abscissas escrevem-se os valores da *variável independente* (designada por V.I.) e, no eixo das ordenadas, os valores correspondentes da *variável dependente*, ou *função* (designada por V.D.).

**N.T.: Essa característica, na verdade, é o próprio nome da variável que constitui objeto de investigação. É conveniente identificar seu nível de mensuração e, além disso, observar o que foi dito na nota de rodapé anterior sobre variáveis independentes e dependentes.

respondente); a linha vertical costuma ser rotulada de acordo com o que se pretende representar (freqüências ou porcentagens), sendo que os valores numéricos correspondem a pontos situados ao longo dos eixos. Além disso, todo gráfico deve ter um título capaz de indicar a natureza dos dados ilustrados.

3. Na construção de um gráfico, o comprimento do eixo vertical deve ser 75% do comprimento do eixo horizontal. Tal procedimento introduz uma forma relativamente padronizada de construir gráficos e reduz ao mínimo fontes de confusão potencial.*
4. O primeiro ponto do eixo vertical–o ponto onde o eixo vertical cruza com o horizontal–deve ser sempre zero, uma vez que qualquer outro procedimento poderia fornecer uma imagem distorcida dos dados.



(a) Leptocúrtica (b) Platicúrtica (c) Mesocúrtica

FIGURA 3.6 Variações em Curtose entre Distribuições Isométricas

FORMA DE UMA DISTRIBUIÇÃO DE FREQÜÊNCIAS

Os processos gráficos podem ajudar-nos a visualizar a imensa variedade de figuras e formatos que as distribuições de freqüência assumem. Algumas distribuições são *simétricas:* "dobrando-se"a curva ao meio as duas metades coincidem. Portanto, tais distribuições contêm o mesmo número de escores extremos em ambos os lados (escores altos do lado direito e escores baixos do lado esquerdo).** Outras distribuições, denominadas *assimétricas*, apresentam maior quantidade de dados extremos numa das caudas.

As chamadas distribuições simétricas apresentam considerável variação. Por exemplo, elas podem diferir acentuadamente em termos de *alongamento* (ou *curtose*). As distribuições simétricas, tais como as da Figura 3.6, podem ser: (a) *leptocúrticas* (bastante pontiagudas ou altas); (b) *platicúrticas* (quando são achatadas); (c) *mesocúrticas* (quando não são nem muito pontiagudas nem muito achatadas). Um tipo de distribuição simétrica mesocúrtica (ver Figura 3.6(c)) é a *curva normal*, que tem especial importância na pesquisa. (A curva normal será estudada pormenorizadamente no Capítulo 6.)

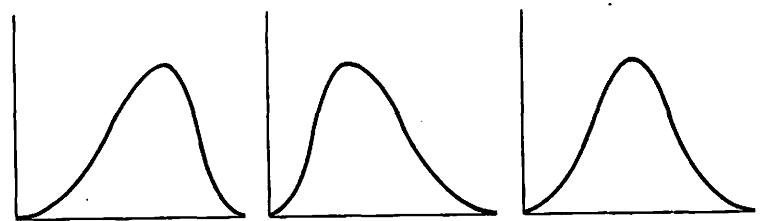
---

* N.T.: O procedimento indicado na nota 3, acima, é irrelevante uma vez que a interpretação do fenômeno posto em gráfico depende *apenas* das *escalas de medidas* usadas e não do comprimento dos eixos.

**N.T.: As extremidades esquerda e direita de uma curva são comumente chamadas *caudas*.

Existem vários tipos de distribuições assimétricas ou enviesadas. Quando os escores se amontoam num dos lados da curva assimétrica, a distribuição apresenta uma cauda acentuada. A posição dessa causa indica o *sentido* da assimetria, sentido esse que aponta para o lado onde se situam os escores mais extremos.

A distribuição (a) da Figura 3.7 é *assimétrica negativa* (enviesada para a esquerda), uma vez que sua cauda é muito maior do lado esquerdo do que do lado direito. Essa distribuição mostra que a maioria dos respondentes recebeu notas altas, e que somente alguns receberam notas baixas. Se essa fosse a distribuição dos resultados de um exame final, poderíamos dizer que a maioria dos estudantes saiu-se muito bem, enquanto que apenas alguns "rodaram".



(a) (b) (c)

FIGURA 3.7 Três Distribuições Representativas do Sentido da Assimetria

Observe-se em seguida a distribuição (b), cuja cauda está à direita. Uma vez que a assimetria é uma função do sentido da cauda, podemos dizer que essa distribuição é *assimétrica positiva* (ou enviesada à direita). Os resultados do exame final dos alunos dessa nossa classe hipotética seriam bem baixos!

Finalmente, vamos examinar a distribuição (c), que contém duas caudas idênticas. Em tal situação há, em ambas as caudas, a mesma quantidade de escores extremos. A simetria dessa distribuição é absoluta. Se esta fosse a distribuição dos resultados de nosso exame final, teríamos um número grande de estudantes cujas notas estariam próximas da média, e poucos estudantes com notas muito altas ou muito baixas.

RESUMO

As representações gráficas de dados podem ser usadas para aumentar a legibilidade de resultados de pesquisas. Nosso estudo de representações gráficas abrange gráficos setoriais, gráficos de barras e polígonos de freqüências. Os gráficos setoriais servem para ilustrações simples de dados que podem ser classificados em algumas categorias. Os gráficos de barras são de uso mais generalizado, já que têm a propriedade de acomodar qualquer quantidade de categorias. Os polígonos de freqüências também gozam dessa elasticidade, mas são particularmente úteis quando se trata de representar dados de natureza ordinal ou intervalar, pois o "efeito de continuidade" torna-se mais evidente.

As variações na forma das distribuições podem ser caracterizadas em termos de simetria ou, conforme haja um acúmulo de dados extremos na cauda esquerda ou na cauda direita da curva, em termos de assimetria positiva ou negativa.

4

Medidas de Tendência Central

Pesquisadores em muitos campos têm usado o termo "média" em questões tais como: qual a renda *média* de colegiais e universitários já graduados? Quantos cigarros fuma, em *média*, o adolescente? Qual a nota *média* de uma universitária? Em *média*, quantos são os acidentes automobilísticos que resultam diretamente da ingestão de álcool ou drogas?

Uma forma útil de descrever um grupo como um todo consiste em encontrar um único número que represente o que é "médio" ou "típico" naquele conjunto particular de dados. Em pesquisa, tal valor é conhecido por *medida de tendência central*, uma vez que ela geralmente se localiza em torno do meio ou centro de uma distribuição—onde a maior parte dos dados tende a concentrar-se.

A idéia que o leigo faz do termo "média" é quase sempre vaga e mesmo confusa. A concepção do pesquisador é muito mais precisa do que a do público em geral; ela é expressa numericamente como uma das várias espécies de "medidas médias"—ou de tendência central—que, para um mesmo conjunto de dados, pode assumir diferentes valores. Somente as três medidas de tendência central mais conhecidas são aqui discutidas: a *moda*, a *mediana* e a *média aritmética*.

MODA

Para obter a moda (Mo), procure simplesmente o escore ou categoria que, numa distribuição, ocorre com maior freqüência. A moda pode ser localizada com muito mais facilidade por exame do que por cálculo. Por exemplo, no conjunto de escores 1, 2, 3, 1, 1, 6, 5, 4, 1, 4, 4, 3, a moda é 1, uma vez que 1 é o número que ocorre com maior freqüência no conjunto (quatro vezes).

No caso de uma distribuição de freqüências com dados isolados, onde os escores e as freqüências figuram em colunas separadas, a moda é o valor (escore) que aparece mais amiúde na tabela. Portanto, na distribuição de freqüências da Tabela 4.1, Mo = 4.

Algumas distribuições de freqüências apresentam duas ou mais modas. No seguinte conjunto de dados, por exemplo, os escores 2 e 6 ocorrem *ambos* com igual freqüência alta: 6, 6, 7, 2, 6, 1, 2, 3, 2, 4. Graficamente, tais distribuições têm dois pontos de

freqüências máximas, que sugerem o costado de um camelo. Essas distribuições são chamadas de *bimodais*, em contraste com a variedade mais comum, *unimodal*, que tem apenas uma corcova—ou ponto de freqüência máxima (ver Figura 4.1).

TABELA 4.1 Procedimento para Encontrar a *Moda* numa Distribuição de Freqüências com Dados Isolados

	Valores	*f*
	7	2
	6	3
	5	4
Mo →	4	5
	3	4
	2	3
	1	2
	Total	23

MEDIANA

Quando dados ordinais ou intervalares são dispostos por ordem de tamanho, torna-se possível localizar a mediana (Md), que corresponde ao ponto *central* da distribuição.

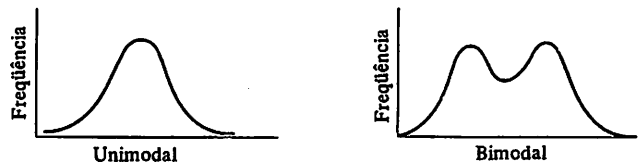


FIGURA 4.1 Representações Gráficas de uma Distribuição Unimodal e de uma Bimodal

Portanto, a mediana é considerada a medida de tendência central que corta a distribuição em duas partes iguais.

Se estivermos diante de uma distribuição com número *ímpar* de dados, a mediana será o dado que cai exatamente no meio da distribuição. A posição do valor mediano pode ser determinada pelo exame dos dados ou pela fórmula:

$$\text{Posição da mediana} = \frac{N+1}{2}$$

Assim, 16 é o valor mediano na distribuição 11, 12, 13, 16, 17, 20, 25; em outras palavras, 16 é o valor que divide a distribuição de tal forma que, de ambos os lados, há igual quantidade de dados. De acordo com a fórmula, (7 + 1)/2, vemos que a mediana, 16, é o quarto valor da distribuição independentemente do lado por onde se inicie a contagem.

Se o número de dados for *par*, a mediana será sempre aquele *ponto* da distribuição que é antecedido e precedido por igual número de dados.* Para uma distribuição com número par de dados, sempre há dois valores considerados "centrais". Ilustrando: os números 16 e 17 representam os dados centrais na seguinte distribuição: 11, 12, 13, 16, 17, 20, 25, 26. Pela fórmula, $(8 + 1)/2 = 4,5$, o que significa que a mediana vai cair entre o quarto e o quinto valores; o ponto central dessa distribuição vai, então, ser 16,5, uma vez que ele fica precisamente no meio de 16 (4.º valor) e 17 (5.º valor). Por igual raciocínio, a mediana é 9 na seqüência 2, 5, 8, 10, 11, 12, porque está situado exatamente entre os dois dados centrais, ou seja, $(6 + 1)/2 = 3,5$.

Outra circunstância que precisa ser explanada e ilustrada, caracteriza-se pelo cálculo da mediana de uma distribuição em que figuram, na região central, vários dados numericamente idênticos. A solução é simples: aquele valor particular constitui a mediana. Assim, na seqüência 11, 12, 13, 16, 16, 16, 25, 26, 27, a mediana é 16, apesar de esse valor ocorrer mais de uma vez.

Obtenção da Mediana numa Distribuição de Freqüências com Dados Isolados

Para encontrar a mediana de dados dispostos numa tabela de freqüências simples, iniciamos com o procedimento apresentado acima. No caso da Tabela 4.1,

$$\text{Posição da mediana} = \frac{23 + 1}{2} = \frac{24}{2} = 12$$

A mediana vem a ser o décimo segundo valor nessa distribuição de freqüências. Para facilitar a localização desse décimo segundo escore, poderíamos construir uma distribuição de freqüências acumuladas, tal como ilustra a terceira coluna da Tabela 4.2 (para um pequeno número de escores isso pode ser feito de cabeça). A partir do escore

TABELA 4.2 Localização da Mediana numa Distribuição de Freqüências de Dados Isolados

	Valores dos Escores	*f*	*fa*
	7	2	23
	6	3	21
	5	4	18
Md →	4	5	14
	3	4	9
	2	3	5
	1	2	2
	Total	23	

* N.T.: De acordo com o texto original, a tradução deveria ser: aquele ponto acima e abaixo do qual ocorrem 50% dos valores da distribuição. Esta afirmação foi evitada no corpo do texto por tratar-se de uma *incoerência*. De fato, se antes da Md há 50% dos escores, e depois dela também 50% de escores, é evidente que 50% + Md + 50% > 100%, o que é um absurdo!

mais baixo, vamos somando as freqüências até chegar a um valor que contenha o décimo segundo escore da distribuição. No exemplo em foco, o valor do escore mediano é 4.

MÉDIA ARITMÉTICA

Sem sombra de dúvida, a medida de tendência central mais comumente usada é a média aritmética, $\bar{X}$, cujo cálculo consiste em somar um conjunto de escores e dividir o total pelo número de parcelas. Portanto, podemos definir média aritmética, mais formalmente, como *a soma de um conjunto de escores dividida pelo número de escores desse conjunto*. Em símbolos:

$$\bar{X} = \frac{\Sigma X}{N}, \text{ onde}$$

$\bar{X}$ = média (leia-se "xis-barra")
Σ = soma (expressa pela letra grega, maiúscula, "sigma")[1]
X = qualquer escore bruto do conjunto (isto é, a própria variável)
N = total de escores do conjunto.

Usando a fórmula acima, verificamos que o QI médio dos oito respondentes arrolados na Tabela 4.3 é 108.

TABELA 4.3 Cálculo da Média Aritmética: Uma Ilustração

Respondentes	*X(QI)*	
Benedito	125	
Francisco	92	$\bar{X} = \frac{\Sigma X}{N}$
Sara	72	
Miguel	126	$= \frac{864}{8}$
Jesuíno	120	
Helena	99	
Benjamim	130	$= 108$
Paulo	100	
	$\Sigma X = 864$	

Em oposição ao que acontece com a moda, a média nem sempre é o escore que ocorre com maior freqüência. Também, ao contrário do que se passa com a mediana, a média não é necessariamente o ponto central de uma distribuição. O que, então, significa *média*? De que maneira ela pode ser interpretada? Veremos adiante que a média pode ser tomada como o "centro de gravidade", isto é, o ponto de qualquer distribuição

[1] A letra maiúscula grega sigma, (Σ), será observada várias vezes ao longo do texto. Ela indica simplesmente que devemos *somar* ou adicionar os valores da variável em foco. No exemplo presente, ΣX indica a soma de todos os escores brutos.

em torno do qual se equilibram as discrepâncias positivas e negativas. A fim de que seja possível perceber essa característica da média, devemos primeiro entender o conceito de *discrepância*, que nada mais é senão a *distância da média a qualquer escore bruto*. Para calcular uma determinada discrepância, basta subtrair a média aritmética de qualquer escore bruto dado. Em símbolos:

$$x = X - \bar{X}\text{ , onde}$$

x = escore-discrepância ou simplesmente discrepância (sempre simbolizada por x, isto é, "xis minúsculo")
X = qualquer escore bruto da distribuição
$\bar{X}$ = média aritmética.

TABELA 4.4 Discrepâncias de um Conjunto de Escores Brutos com Relação a $\bar{X}$

X	x		
9	+3	+5	
8	+2		
6	0		$\bar{X} = 6$
4	−2	−5	
3	−3		

Uma vez que $\bar{X} = 6$ para o conjunto de escores brutos 9, 8, 6, 4 e 3, o valor $X = 9$ fica exatamente três unidades *acima* da média 6 (ou $X - \bar{X} = 9 - 6 = +3$). De modo análogo, o escore bruto 4 fica duas unidades *abaixo* da média (ou $X - \bar{X} = 4 - 6 = -2$). Conclusão: quanto maior a discrepância x, tanto maior a distância entre um determinado escore bruto e a média da distribuição.

Considerando-se a média como o ponto de equilíbrio da distribuição, pode-se dizer, agora, que a soma das discrepâncias situadas acima da média é igual, em valor absoluto (descartados os sinais "menos"), à soma das discrepâncias que ficam abaixo dela. Vamos voltar a um exemplo anterior–o conjunto de escores 9, 8, 6, 4, 3, onde $\bar{X} = 6$. Se a média dessa distribuição for seu "centro de gravidade", então, desprezando os sinais "menos" e somando as discrepâncias positivas (discrepâncias relativas aos escores 8 e 9), o resultado deve ser igual ao obtido mediante a soma das discrepâncias negativas (isto é, relativas aos escores 4 e 3). Como evidencia a Tabela 4.4, é precisamente esse caso–já que a soma das discrepâncias abaixo de $\bar{X}$ (− 5) é igual à soma das discrepâncias acima de $\bar{X}$ (+ 5).

Vamos a mais um exemplo: 4 é a média de 1,2, 3, 4, 5, 6 e 7. Observe-se que a soma das discrepâncias abaixo desse escore é −6, enquanto que a soma das discrepâncias acima dele é +6. Voltaremos a examinar o conceito de discrepância nos Capítulos 5 e 6.

Cálculo da Média de uma Distribuição de Freqüências (com Dados Isolados)

Para obter-se a média de um pequeno número de escores, a fórmula básica $\bar{X} = \Sigma X/N$ serve perfeitamente. Entretanto, quando estamos diante de um número maior de dados, pode ser mais prático e menos demorado calcular a média de uma distribuição de freqüências (dados isolados) pela fórmula

$$\bar{X} = \frac{\Sigma fX}{N}\text{ , onde}$$

$\bar{X}$ = média
X = um escore bruto qualquer da distribuição (isto é, a própria variável)
fX = um escore qualquer multiplicado pela respectiva freqüência de ocorrência
ΣfX = soma dos fX's
N = total de escores.

A Tabela 4.5 ilustra o cálculo da média de um conjunto de dados isolados (porém afetados de freqüências).

TABELA 4.5 Cálculo de $\bar{X}$ de uma Distribuição de Freqüências Simples (Dados Isolados)

X	f	fX	
8	2	16	
7	3	21	
6	5	30	
5	6	30	$\bar{X} = \frac{\Sigma fX}{N} = \frac{132}{28} = 4{,}71$
4	4	16	
3	4	12	
2	3	6	
1	1	1	
	$N = 28$	$\Sigma fX = 132$	

COMPARAÇÃO ENTRE MODA, MEDIANA E MÉDIA

Há um momento em que o pesquisador procura uma medida de tendência central para a sua situação particular de pesquisa. Empregará ele a moda, a mediana ou a média? Sua decisão envolve vários fatores tais como:

1. o nível de mensuração;
2. o aspecto ou forma de sua distribuição de dados e
3. o objetivo da pesquisa.

Nível de Mensuração

Uma vez que a moda requer apenas o conhecimento das freqüências, ela pode ser calculada para qualquer conjunto de dados, sejam eles do nível nominal, ordinal ou intervalar. Por exemplo, poderíamos estabelecer que a categoria modal do nível nominal da variável "filiação religiosa" (protestante, católico ou judeu) é protestante, se a maioria dos nossos respondentes se identificassem como tais. Similarmente, poderíamos dizer que a maioria dos estudantes freqüentadores de uma dada universidade apresenta uma nota média 2,5 (Mo = 2,5).

A mediana exige uma ordenação de categorias, da mais alta à mais baixa. Por essa razão, ela só pode ser obtida a partir de dados ordinais ou intervalares, mas *nunca* de dados nominais. Ilustrando: poderíamos descobrir que a renda anual mediana de dentistas de uma pequena cidade é da ordem de Cr$ 289.000,00. Esse resultado fornece-nos um critério sólido para o exame da tendência central de nossos dados. Contrariamente, faria pouco sentido calcular a mediana de escalas de filiação religiosa (protestante, católico ou judeu), sexo (masculino ou feminino), ou país de origem (Inglaterra, Polônia, França ou Alemanha), onde não foi feita uma atribuição de postos (escalonamento).

O uso da média está restrito exclusivamente a dados intervalares. Calcular a média de dados ordinais ou nominais dá origem a um resultado sem sentido, que em geral não indica a tendência central. Que sentido faria o cálculo da média de uma distribuição onde a variável fosse filiação religiosa ou sexo? Embora seja menos óbvio, é igualmente inapropriado calcular a média de dados que podem ser *graduados* porém não *medidos*.

Forma da Distribuição

O aspecto ou forma de uma distribuição é outro fator que pode influenciar o pesquisador na escolha de uma medida de tendência central. Numa distribuição unimodal e perfeitamente simétrica, a moda, a mediana e a média serão idênticas, uma vez que o ponto de freqüência máxima (Mo) é também o valor "mais" central (Md) e o "centro de gravidade" dos dados ($\bar{X}$). Como demonstra a Figura 4.2, as medidas de tendência central coincidirão no ponto central, exatamente no "pico" da distribuição simétrica.

Quando o pesquisador trabalha com uma distribuição simétrica, a escolha de uma medida de tendência central baseia-se principalmente nos objetivos particulares de sua investigação e no nível em que seus dados foram colhidos. Entretanto, quando ele trabalha com uma distribuição assimétrica, sua decisão é muito influenciada pelo formato dos dados de que dispõe.

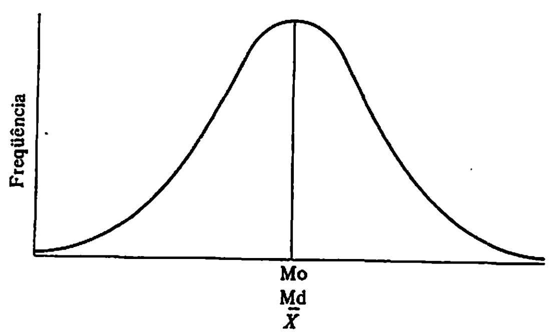


FIGURA 4.2 Distribuição Unimodal, Simétrica, com a Localização da Moda, da Mediana e da Média (onde Mo = Md = $\bar{X}$)

A Figura 4.3 mostra que a moda, a mediana e a média não coincidem em distribuições assimétricas, *muito embora suas posições relativas permaneçam constantes;* dito de outra forma, no sentido "pico"/"cauda", a ordem é sempre da moda para a média, com a mediana entre esses dois pontos. A moda é a que cai mais perto do "pico" da curva, já que ela corresponde ao ponto da escala cujo valor é mais freqüente. A média, ao contrário, é a que está mais próxima da "cauda", onde se situam os escores extremos (e poucos). Por essa razão, o escore médio numa distribuição positivamente assimétrica (Figura 4.3(a)) localiza-se entre os escores mais altos; a média numa distribuição negativamente assimétrica (Figura 4.3(b)) cai perto dos escores mais baixos.

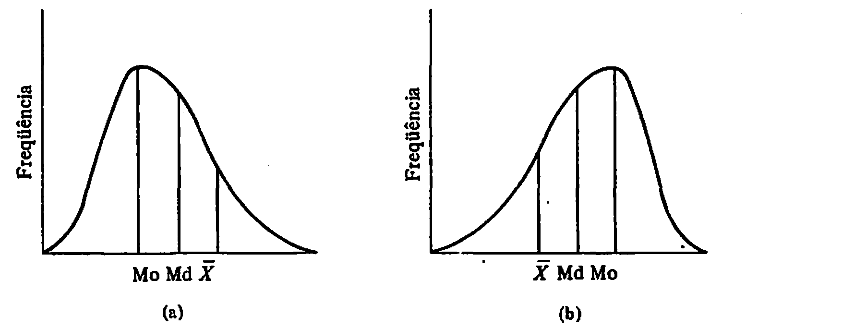


FIGURA 4.3 Posições Relativas de Medidas de Tendência Central: (a)–Distribuição Positivamente Assimétrica e (b)–Distribuição Negativamente Assimétrica

Enquanto que a média é muitíssimo influenciada pelos escores extremos (em ambos os sentidos), a mediana sofre pouca ou nenhuma influência de alterações nos valores extremos. Isto porque a média considera todos os escores de qualquer distribuição, enquanto que a mediana (por definição) só se preocupa com o valor numérico do escore que ocupa a posição "mais" central. Como ilustrado abaixo, a mudança do valor de um escore extremo de 10 (distribuição a) para 95 (distribuição b), não modifica de maneira nenhuma o valor da mediana (Md = 7,5), enquanto que a média pula de 7,63 para 18,25.

distribuição A:	5 6 6 7 8 9 10 10	Md = 7,5	$\bar{X}$ = 7,63
distribuição B:	5 6 6 7 8 9 10 95	Md = 7,5	$\bar{X}$ = 18,25

Numa distribuição assimétrica, a mediana sempre cai em algum lugar entre a média e a moda. É essa característica que a torna a medida de tendência central mais desejável para descrever uma distribuição assimétrica. Para ilustrar essa vantagem da mediana, voltemo-nos para a Tabela 4.6 e examinemos o salário anual "médio" dos empregados que trabalham numa pequena corporação. Se trabalhássemos no departamento de relações públicas dessa corporação e desejássemos angariar-lhe uma imagem pública favorável, calcularíamos, talvez, a média, a fim de demonstrar que o empregado "típico" ganha Cr$ 306.000,00 (por ano), e é, portanto, relativamente bem pago. Por outro lado, se nós fôssemos representantes sindicais e estivéssemos procurando melhorar os níveis salariais, iríamos provavelmente empregar a moda para demonstrar que o salário anual "médio" é de apenas Cr$ 17.000,00, o que representa uma quantia bastante reduzida. Para finalizar, se fôssemos pesquisadores sociais desejosos de dar uma informação acurada do salário "médio" dos empregados dessa corporação, empregaríamos, inteligentemente, a mediana (Cr$ 51.000,00), uma vez que esse valor cai entre as outras medidas de tendência central e, portanto, oferece um quadro mais equilibrado da estrutura salarial. O método mais aceitável seria indicar as três medidas de tendência central e permitir que os interessados interpretassem os resultados. Infelizmente é verdade que poucos pesquisadores sociais–não contando os relações públicas e os líderes sindicais–indicam mais do que uma única medida de tendência central.

TABELA 4.6 **Medidas de Tendência Central numa Distribuição Assimétrica de Salários Anuais**

Salário Medido em Cruzeiros (Cr$)	
1.700.000,00	
425.000,00	
170.000,00	$\bar{X}$ = Cr$ 306.000,00
85.000,00	
17.000,00	Md = Cr$ 51.000,00
17.000,00	
17.000,00	Mo = Cr$ 17.000,00

Ainda mais grave é o fato de que alguns relatórios de pesquisa não especificam de forma exata a medida de tendência central–moda, mediana ou média–que foi usada para calcular a quantia ou posição "média" de um grupo de escores. Como demonstrou a ilustração anterior, uma interpretação razoável de certos dados pode tornar-se literalmente impossível sem tal informação.

Salientou-se páginas atrás que algumas distribuições de freqüências podem ser caracterizadas como bimodais, uma vez que elas contêm dois pontos de freqüência máxima. Para descrever adequadamente distribuições bimodais, é, em geral, útil identificar *ambas* as modas; aspectos importantes de tais distribuições podem ser encobertos se a mediana ou a média forem usadas.

Considerem a situação em que um pesquisador realizou entrevistas pessoais com 26 respondentes de renda baixa, a fim de estabelecer seus conceitos ideais de tamanho de família. A cada um deles foi dito o seguinte: "Suponha que você pudesse decidir exatamente o tamanho de sua família. Incluindo todas as crianças e os adultos, quantas pessoas você gostaria que ela tivesse?" Como demonstra a Tabela 4.7, os resultados desse estudo indicaram amplas preferências por tamanho de família, preferências que variaram desde *viver sozinho* (1) até *viver com muitas pessoas* (10). Usando ou a média ou a mediana, poderíamos concluir que a família ideal de um respondente "médio" consistiria de seis pessoas ($\bar{X}$ = 5,58; Md = 6). Entretanto, sabendo que a distribuição é bimodal, percebemos que, de fato, houve *duas* concepções ideais de tamanho de família

nesse grupo de respondentes: uma com um número bastante grande de pessoas (Mo = 8) e outra com apenas algumas pessoas (Mo = 3).

TABELA 4.7 **Concepções Ideais de Tamanho de Família numa Amostra de 26 Respondentes de Renda Baixa. (Distribuição Bimodal)**

Tamanho Ideal de Família	*f*
10	1
9	2
8	6
7	3
6	2
5	1
4	2
3	6
2	2
1	1
	N = 26

Objetivo da Pesquisa

Até aqui, discutimos a escolha de uma medida de tendência central em termos do nível de mensuração e da forma de uma distribuição de escores. Perguntamos agora: que o pesquisador espera fazer com a medida de tendência central escolhida? Se ele procura uma medida descritiva que seja rápida, simples, ainda que grosseira, ou se ele estiver trabalhando com uma distribuição bimodal, ele geralmente empregará a moda. Na maioria das situações com que um pesquisador se defronta, entretanto, a moda só é útil como um indicador preliminar de tendência central, que pode ser obtida com rapidez a partir de um exame cuidadoso dos dados. Se ele procura uma medida de tendência central que seja exata,* a decisão fica geralmente entre a mediana e a média.

* N.T.: Muito embora o uso popular tenha tornado sinônimas as palavras *precisão* e *exatidão*, do ponto de vista formal estamos diante de coisas diferentes. Um instrumento é *preciso* quando, em sucessivas mensurações da mesma magnitude, ele apresenta um *erro constante*. Um instrumento preciso caracteriza-se pela *precisão*.

Um instrumento é *exato* quando, em sucessivas mensurações da mesma magnitude, o *erro* é *variável* (para mais ou para menos), porém a média aritmética de todas as mensurações representa o valor mais provável da dita magnitude. Por isso, a média aritmética representa o valor *exato* buscado. O instrumento que produz mensurações exatas caracteriza-se pela *exatidão*.

Exemplificando: Se um objeto pesa *realmente* 2 kg, pesado com uma balança exata repetidas vezes, as mensurações resultantes poderão ser: 2,1 kg; 2,2 kg; 1,9 kg; 1,8 kg; etc., onde a média aritmética produz o valor real buscado. Assim:

$$\frac{2{,}1\text{ kg} + 2{,}2\text{ kg} + 1{,}9\text{ g} + 1{,}8\text{ kg}}{4} = 2\text{ kg}$$

Seja agora outra balança com a seguinte característica: em cada quilo ela erra, para mais, 10 gramas. Então, esse objeto, cujo peso real é 2 kg, pesará *sempre* 2 kg + 20 g. Se a pesagem for repetida n vezes ($n > 1$), o erro será consistente e constante: 20 g.

Na descrição de uma distribuição assimétrica, o pesquisador normalmente escolhe a mediana (como, aliás, já foi mencionado em outro local), uma vez que ela tende a dar um quadro balanceado* dos escores extremos. Além disso, a mediana é algumas vezes usada como um ponto da distribuição que divide os escores em duas categorias—*ambas* com o *mesmo número* de sujeitos. Por exemplo, poderíamos dividir os respondentes (Tabela 4.7) em dois grupos: aqueles que preferem uma família grande (valores superiores à mediana) e os que preferem família pequena (valores menores que a mediana).

Sempre que se deseja uma medida exata de distribuições simétricas, a média tem preferência sobre a mediana pelo fato de poder ser facilmente usada em análises estatísticas mais avançadas, tais como as que aparecem em capítulos subseqüentes deste livro. Acrescente-se a isso o fato de ser a média mais estável que a mediana, ou seja, se de uma dada população forem extraídas diferentes amostras, haverá menor variação nas médias do que nas medianas. Essa vantagem da média—muito embora talvez ainda não compreendida ou avaliada pelo leitor—tornar-se-á mais evidente em estudos posteriores do tópico que trata do problema das *decisões em Estatística* (ver Capítulo 7).

CÁLCULO DA MODA, DA MEDIANA E DA MÉDIA NUMA DISTRIBUIÇÃO DE FREQÜÊNCIAS COM DADOS AGRUPADOS

Numa distribuição de freqüências com dados agrupados, a moda é o ponto médio do intervalo de classe que apresenta a freqüência mais alta. De acordo com essa definição, a moda da distribuição constante da Tabela 4.8 é 72, uma vez que esse valor corresponde ao ponto médio do intervalo que aparece mais vezes (ele ocorre 17 vezes)**

TABELA 4.8 Localização da Moda numa Distribuição de Freqüências com Dados Agrupados

Intervalo de Classe	*Ponto Médio*	*f*
95-99	97	3
90-94	92	2
85-89	87	4
80-84	82	7
75-79	77	12
70-74	72	17
65-69	67	12
60-64	62	5
55-59	57	5
50-54	52	4
		$N = 71$

*N.T.: Balanceado significa aqui "que não sofre a influência de valores extremos".

**N.T.: Rigorosamente, o ponto médio da 6ª classe (de cima para baixo) é 71,5, se a variável for contínua. Para que esse ponto médio seja 72 é preciso que, nas condições da tabela, sejam desprezados os limites do intervalo, ou seja, 70 e 74. Isso porque a classe foi indicada pela notação 70–74, que é a maneira tradicional de indicar um intervalo aberto (dos dois lados).

Para localizar a mediana de dados agrupados numa distribuição de freqüências, devemos (1) encontrar o intervalo de classe que contém a mediana e (2) interpolar.

Passo 1—Para localizar o intervalo mediano, construímos primeiro uma distribuição de freqüências acumuladas, tal como figura na terceira coluna da Tabela 4.9. A partir do intervalo que contém os valores mais baixos (as idades menores—20/29), vamos somando as freqüências até chegar àquele intervalo que contém o ponto (o valor) que divide a distribuição em duas partes iguais, isto é, o ponto central. No presente exemplo, $N = 100$. Portanto, devemos procurar o qüinquagésimo valor, que é: $N/_2 = 100/2 = 50$. Consultando a coluna das freqüências acumuladas (de baixo para cima), observamos que 26 sujeitos têm idades de 39 anos ou menos. Vemos também que o qüinquagésimo sujeito situa-se no intervalo 40/49, já que esse é o intervalo de classe cujas freqüências acumuladas contém 53—portanto, mais do que a metade dos sujeitos. Em outras palavras: com relação às freqüências acumuladas, do vigésimo sétimo ao qüinquagésimo terceiro sujeitos encontramo-nos no intervalo 40/49. Esse é o intervalo mediano.*

TABELA 4.9 Distribuição de Freqüências de Dados Agrupados (idades)

Intervalo	*f*	*fa*
60-69	15	100
50-59	32	85
40-49	27	53
30-39	16	26
20-29	10	10
	$N = 100$	

Passo 2—Para encontrar o valor exato da mediana, aplicamos a fórmula

$$\text{Mediana} = \left(\begin{array}{c}\text{Limite inferior real}\\ \text{do intervalo}\\ \text{mediano}\end{array}\right) + \left(\frac{\dfrac{N}{2} - \left(\begin{array}{c}\text{Freqüência acumulada}\\ \text{correspondente ao}\\ \text{intervalo anterior ao}\\ \text{que contém a mediana}\end{array}\right)}{\begin{array}{c}\text{Freqüência absoluta do}\\ \text{intervalo mediano}\end{array}}\right) \times$$

$$\times \left(\begin{array}{c}\text{Amplitude}\\ \text{do intervalo}\\ \text{mediano}\end{array}\right)$$

* N.T.: Melhor seria dizer apenas o seguinte: o intervalo mediano é o intervalo que corresponde, na coluna de freqüências acumuladas, ao valor que contém (pela primeira vez) metade do total de sujeitos da distribuição. Sendo fa_i uma freqüência acumulada de ordem i, o intervalo mediano será obtido por correspondência com a relação $fa_i \geqslant N/2$. O intervalo mediano é também chamado de *lugar mediano*.

Para os dados da Tabela 4.9*, a mediana é calculada como segue:

$$\text{Mediana} = 39{,}5 + \left(\frac{50-26}{27}\right)10 = 39{,}5 + 8{,}89 = 48{,}39 \text{ anos.}$$

Para calcular a média de uma distribuição de freqüências com dados agrupados, podemos usar uma versão modificada da fórmula vista anteriormente (ver Tabela 4.5). Conforme exemplificado abaixo, o símbolo X não será mais usado para designar um escore, mas, sim, o *ponto médio de um intervalo de classe.* Portanto,

$$\bar{X} = \frac{\Sigma fX}{N}\text{, onde}$$

$\bar{X}$ = média
X = ponto médio de um intervalo de classe
fX = ponto médio de um intervalo de classe multiplicado pela respectiva freqüência absoluta
N = número total de escores.

Podemos ilustrar o cálculo da média de dados agrupados usando como referência a seguinte distribuição.**

Intervalo	*f*
17 ⊢20	1
14 ⊢17	2
11 ⊢14	3
8 ⊢11	5
5 ⊢8	4
2 ⊢5	2
	$N = 17$

PASSO 1: Calcular o ponto médio de cada intervalo de classe.

Intervalo	*X = Ponto Médio*
17 ⊢20	18
14 ⊢17	15
11 ⊢14	12
8 ⊢11	9
5 ⊢8	6
2 ⊢5	3

*N.T.: A variável "idade" é uma variável *contínua*. Por isso, qualquer *limite inferior real* é obtido a partir do *limite inferior aparente menos 0,5*. Para ajudar a "enxergar" este fato, o ideal seria adotar a seguinte notação nos intervalos de classe: 40 ⊢50, 50 ⊢60, etc. Assim fica claro que os limites inferiores reais são, respectivamente, 39,5 e 49,5, que os superiores são, na mesma ordem, 49,4 e 59,4.

**N.T.: Veja-se a propósito, o que ficou dito na nota do tradutor, página 24, Cap. 2, sobre a notação de intervalo.

PASSO 2: Multiplicar cada ponto médio pela respectiva freqüência absoluta, a fim de obter ΣfX.

Intervalo	*X = Ponto Médio*	*f*	*fX*
17 –20	18	1	18
14 –17	15	2	30
11 –14	12	3	36
8 –11	9	5	45
5 –8	6	4	24
2 –5	3	2	6
		$N = 17$	$\Sigma fX = 159$

PASSO 3: "Entrar" com o resultado do Passo 2 na fórmula de $\bar{X}$.

$$\bar{X} = \frac{\Sigma fX}{N} = \frac{159}{17} = 9{,}35$$

RESUMO

Este capítulo introduziu as três medidas de tendência central mais conhecidas, vale dizer, medidas do que é "médio" ou "típico" num conjunto de dados. A moda foi definida como sendo a categoria ou escore que ocorre com maior freqüência; a mediana, como o ponto que divide a distribuição ao meio; a média, como a soma de todos os valores dividida pelo número de parcelas (isto é, de valores do conjunto). Essas medidas de tendência central foram comparadas relativamente a: nível de mensuração (da variável), forma da distribuição (dos dados) e objetivo da pesquisa. Podemos resumir as condições que regem a escolha de uma dessas medidas de tendência central da seguinte maneira:

Moda:
1. nível de mensuração: nominal, ordinal ou intervalar;
2. forma da distribuição: mais apropriada para distribuições bimodais;*
3. objetivo: permite obter uma medida de tendência central rápida, simples, embora grosseira.

Mediana:
1. nível de mensuração: ordinal ou intervalar;
2. forma da distribuição: mais adequada para distribuições muito assimétricas;

* N.T.: Melhor seria dizer *multimodais*, já que a "bimodalidade" está implícita na "multimodalidade". Observe-se, ainda, que nada impede que a *moda* seja a medida "preferível" mesmo numa distribuição *unimodal*. O que rege isso é o *objetivo* da pesquisa.

3. objetivo: é uma medida de tendência central "confiável"; pode, às vezes, ser usada em operações estatísticas mais avançadas ou para "quebrar" uma distribuição em duas categorias distintas (por exemplo: alto X baixo).

Média:

1. nível de mensuração: intervalar (no mínimo);
2. forma da distribuição: mais apropriada para distribuições unimodais simétricas;
3. objetivo: medida de tendência central exata; pode freqüentemente ser usada em operações estatísticas mais avançadas, tais como os testes para tomada de decisões que aparecerão mais adiante neste livro.

PROBLEMAS

1. Sete empregados horistas numa companhia de porte médio ganham Cr$ 153,00; Cr$ 136,00; Cr$ 153,00; Cr$ 68,00; Cr$ 17,00; Cr$ 102,00 e Cr$ 51,00. Calcule (a) o salário-hora modal; (b) o salário-hora mediano e (c) o salário-hora médio.
2. Suponha que a companhia do problema 1 tenha recrutado mais um empregado pelo salário-hora de Cr$ 17,00, resultando, assim, a seguinte distribuição: Cr$ 153,00; Cr$ 136,00; Cr$ 153,00; Cr$ 68,00; Cr$ 17,00; Cr$ 102,00; Cr$ 51,00; e Cr$ 17,00. Calcule (a) o salário-hora modal; (b) o salário-hora mediano; (c) o salário-hora médio.
3. Para os escores 205, 6, 5, 5, 5, 2, e 1 calcule (a) a moda; (b) a mediana e (c) a média. Que medida de tendência central *não* deveria ser usada para descrever (isto é, para analisar) esse conjunto de escores? Por quê?
4. Seis estudantes num seminário de sociologia foram testados por meio de um instrumento que produz mensurações de nível intervalar. O objetivo da pesquisa era mensurar suas atitudes com relação a um grupo minoritário. Suas respostas, numa escala de 1 a 10 (quanto mais alto o escore, mais favorável a atitude), foram: 5, 2, 6, 3, 1 e 1.

   Relativamente aos escores acima, calcule (a) a moda; (b) a mediana e (c) a média. No geral, até que ponto esses estudantes são favoráveis a esse grupo minoritário?
5. Dados os escores 10, 12, 14, 8, 6, 7, 10, 10, encontre (a) a moda; (b) a mediana e (c) a média.
6. Dados os escores 3, 3, 4, 3, 1, 6, 5, 6, 6, 4, encontre (a) a moda; (b) a mediana e (c) a média..
7. Dados os escores 8, 8, 7, 9, 10, 5, 6, 8, 8, encontre (a) a moda, (b) a mediana e (c) a média.
8. Dados os escores 5, 4, 6, 6, 1, 3, encontre (a) a moda, (b) a mediana e (c) a média.
9. Dados os escores 8, 6, 10, 12, 1, 3, 4, 4, encontre (a) a moda, (b) a mediana e (c) a média.
10. Dados os escores 12, 12, 1, 12, 5, 6, 7, encontre (a) a moda, (b) a mediana e (c) a média.
11. Qual é a discrepância de cada um dos seguintes escores com relação a $\bar{X} = 20{,}5$? (a) $X = 20{,}5$; (b) $X = 33{,}0$; (c) $X = 15{,}0$; (d) $X = 21{,}0$.
12. Qual é a discrepância de cada um dos seguintes escores com relação a $\bar{X} = 3{,}0$? (a) $X = 4{,}0$; (b) $X = 2{,}5$; (c) $X = 6{,}3$; (d) $X = 3{,}0$.

13. Qual é a discrepância de cada um dos seguintes escores com relação a $\bar{X} = 15$? (a) $X = 22{,}5$; (b) $X = 3$; (c) $X = 15$; (d) $X = 10{,}5$.
14. Os escores relativos a atitudes de 31 estudantes frente a um grupo minoritário, foram dispostos na seguinte distribuição de freqüências (quanto maior o escore mais favorável a atitude:

Escores Relativos e Atitudes	*f*
7	3
6	4
5	6
4	7
3	5
2	4
1	2
	$N = 31$

Calcule (a) a moda; (b) a mediana; (c) a média.

15. Trinta e uma crianças matriculadas numa terceira série do primeiro grau de uma zona urbana indicaram, numa pesquisa, o número de irmãos e/ou irmãs que viviam com cada um deles em casa. Os dados resultantes foram dispostos na seguinte tabela de distribuição de freqüências:

Número de Irmãos/irmãs	*f*
5	6
4	7
3	9
2	5
1	4
	$N = 31$

Para essa classe de 31 crianças, calcule
(a) o número modal de irmãos/irmãs;
(b) o número mediano de irmãos/irmãs;
(c) a média de irmãos/irmãs.

16. Dada a seguinte distribuição de freqüências, calcule (a) a moda; (b) a mediana e (c) a média:

Escores	*f*
10	3
9	4
8	6
7	8
6	9
5	7
4	5
3	2
2	1
1	1
	$N = 46$

17. Dada a seguinte distribuição de freqüências, calcule (a) a moda; (b) a mediana e (c) a média:

Intervalo de Classe	f
20-24	2
15-19	4
10-14	8
5-9	5
	$N = 19$

18. Dada a seguinte distribuição de freqüências, calcule (a) a moda; (b) a mediana e (c) a média:

Intervalo de Classe	f
90-99	16
80-89	17
70-79	15
60-69	3
50-59	2
40-49	3
	$N = 56$

19. Dada a seguinte distribuição de freqüências, calcule (a) a moda; (b) a mediana; (c) a média:

Intervalo de Classe	f
17-19	2
14-16	3
11-13	6
8-10	5
5-7	1
	$N = 17$

5

Medidas de Variabilidade

No Capítulo 4, vimos que a moda, a mediana e a média podiam ser usadas para resumir, num único número, aquilo que é "médio" ou "típico" numa distribuição. Quando empregada sozinha, entretanto, qualquer medida de tendência central fornece apenas uma visão incompleta de um conjunto de dados e, portanto, pode confundir ou distorcer tanto quanto esclarecer.

Com vistas a ilustrar essa situação, admitam que Honolulu (Havaí) e Houston (Texas) tenham quase a mesma temperatura média diária de 75°.* Será que, por isso, podemos admitir que a temperatura é basicamente a mesma em ambas as localidades? Ou não será possível que enquanto uma cidade é melhor para natação a outra o seja para atividades externas?

Como bem ilustra a Figura 5.1, a temperatura em Honolulu varia muito pouco ao longo do ano, oscilando, em geral, entre 70°F e 80°F. Por outro lado, a temperatura em Houston pode diferir estacionalmente, isto é, apresentar-se baixa em janeiro—cerca de 40°F—e alta em julho e agosto—bem perto de 100°F. Desnecessário dizer, pois, que as praias de Houston não estão apinhadas de gente durante todo o ano!

Vamos a um outro exemplo. Suponham que, numa particular cidade, tanto ladrões quanto professores secundários tenham uma renda média anual de Cr$ 136.000,00. Será que essa informação indica que as duas distribuições de renda são *necessariamente* semelhantes? Muito ao contrário, poder-se-ia descobrir que elas diferem—e *muito*—num outro aspecto importante, qual seja, o fato de as rendas dos professores *concentrarem-se* ao redor de Cr$ 136.000,00, enquanto que as dos ladrões espalham-se mais, o que reflete, portanto, maiores oportunidades para prisões, desemprego, pobreza e, em *alguns casos*, fortunas excepcionais.

* N.T.: A falta de uma unidade de medida pode, às vezes, confundir o leitor. Este livro, de autor norte-americano, traduz, em muitos pontos, aspectos da cultura vigente em países anglo-saxões; entre esses aspectos, destaca-se a *escala de temperatura Fahrenheit*, cuja característica é a seguinte: 32°F correspondem ao gelo fundente e 212°F, à agua em ebulição. Em graus *Celsius* (centígrados), 75°F correspondem a 23,9°C (aprox.). A fórmula de conversão usada foi a seguinte:

$$°C = \frac{5}{9}(°F - 32°).$$

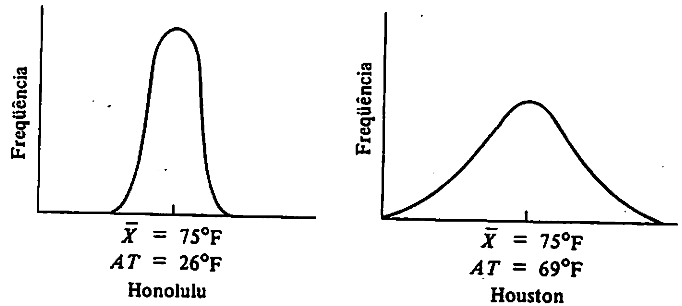


FIGURA 5.1 **Diferenças de Variabilidade: Distribuições de Temperaturas em Locais Distintos: Honolulu e Houston (Valores Aproximados)**

Tal fato demonstra que necessitamos, além de uma medida de tendência central, de um índice que indique o grau de dispersão dos escores em torno do centro da distribuição (isto é, em torno da média). Numa palavra, precisamos de uma medida indicativa do que costumeiramente se chama *variabilidade* (também designada como *variação* ou *dispersão*). Voltando a um exemplo anterior, poderíamos dizer que a distribuição de temperaturas em Houston (Texas) tem *maior variabilidade* do que a distribuição de temperaturas em Honolulu (Havaí). Da mesma forma, podemos dizer que a distribuição de rendas entre professores apresenta *menos variabilidade* do que a distribuição de rendas entre ladrões. O presente capítulo discute apenas as medidas de variabilidade mais conhecidas: a *amplitude total*, o *desvio médio* e o *desvio padrão*.

AMPLITUDE TOTAL

Pode-se obter uma medida de variabilidade rápida (cômoda), embora não muito exata, pelo cálculo da chamada *amplitude total* (*AT*), que nada mais é senão a diferença entre o maior e o menor escore da distribuição. Por exemplo, se a temperatura anual mais alta em Honolulu foi de 88°F e a mais baixa, 62°F, a amplitude total da temperatura anual em Honolulu foi de 26°F (isto é, 88°F – 62°F = 26°F). Se o dia mais quente em Houston apresentou 102°F e o mais frio, 33°F, a amplitude total da temperatura anual em Houston foi de 69°F (isto é, 102°F – 33°F = 69°F).

A vantagem da amplitude total—cálculo rápido e fácil—constitui também sua mais importante desvantagem. Por outras palavras, a amplitude total é inteiramente dependente de *apenas* dois escores: o maior e o menor num dado conjunto de valores. Como resultado, a amplitude total fornece, via de regra, um mero índice grosseiro da variabilidade de uma distribuição. Por exemplo, $AT = 98$ no seguinte conjunto de dados: 2, 6, 7, 7, 10, 12, 13, 100 ($AT = 100 - 2 = 98$); entretanto, $AT = 12$ neste outro conjunto: 2, 6, 7, 7, 10, 12, 13, 14 ($AT = 14 - 2 = 12$). Portanto, pela simples troca de um *único* valor (14 em lugar de 100), fizemos com que a amplitude total flutuasse bruscamente de 98 para 12. Qualquer medida que seja tão afetada pelo escore de um único respondente (ou por um único valor da variável) não pode, por certo, fornecer uma idéia precisa quanto à variabilidade; quando muito, é possível considerá-la como um índice preliminar ou, até, pouco exato.

DESVIO MÉDIO

No capítulo anterior, *desvio** foi definido como sendo a distância entre qualquer escore bruto e a média da distribuição. Aprendemos, assim, que, para calcular discrepâncias (desvios), bastava subtrair a média (aritmética) de qualquer escore bruto: $x = (X - \bar{X})$. Se desejarmos agora obter uma medida de variabilidade que leve em conta *todos* os escores da distribuição (e não *apenas dois*), poderemos tomar o *valor absoluto* de cada discrepância (isto é, das distâncias com relação à média), somar esses valores e, então, dividir o total pelo número de escores. O resultado será o *desvio médio*. Em símbolos:

$$\text{DM} = \frac{\Sigma|x|}{N}\text{, onde}$$

DM = desvio médio
$\Sigma|x|$ = soma dos valores absolutos das discrepâncias (ou seja, soma das discrepâncias *sem* considerar-se a presença dos sinais + e –)
N = total de escores (ou de sujeitos, respondentes ou dados).

Nota importante—No cálculo de $\Sigma|x|$ *devemos* ignorar os sinais "mais" (+) e "menos" (–) e somar apenas os valores absolutos. Este procedimento é *correto*, porque a soma das discrepâncias reais (Σx)—onde os sinais indicam o sentido do desvio com relação à média**—é sempre igual a zero.*** As discrepâncias positivas e as negativas tendem a cancelar-se (compensar-se) e não podem, por isso, ser usadas para descrever ou comparar a variabilidade de distribuições. Por outro lado, a soma dos valores absolutos das discrepâncias tende a tornar-se maior à medida que a variabilidade da distribuição aumenta.

Podemos agora ilustrar, passo a passo, o procedimento para o cálculo do desvio médio. Para tanto, consideremos o seguinte conjunto de escores: 9, 8, 6, 4, 2 e 1.

PASSO 1: Calcular a média (aritmética) da distribuição.

X	
9	
8	$\bar{X} = \dfrac{\Sigma X}{N}$
6	
4	$= \dfrac{30}{6}$
2	
1	$= 5$
$\Sigma X = 30$	

* N.T.: Insista-se no seguinte: *desvio, afastamento* e *discrepância* são expressões sinônimas. Entretanto, é prática comum usar o termo *discrepância* para designar qualquer x_i que resulte de $(X_i - \bar{X})$.

** N.T.: Os valores afetados de + (mais) estão situados à *direita* do zero; os afetados de – (menos), à *esquerda*. Nos termos do texto, a *média* (aritmética), à semelhança do zero, funciona como uma *origem*. Então: à direita da média, discrepâncias positivas; à esquerda, discrepâncias negativas.

*** N.T.: Dizer que a soma das discrepâncias é sempre igual a zero é o mesmo que dizer, em símbolos, o seguinte: $\Sigma x = 0$. Ora, $\Sigma x = \Sigma(X - \bar{X})$. Então: $\Sigma(X - \bar{X}) = 0$.

PASSO 2: Subtrair de cada escore bruto a média da distribuição e somar as discrepâncias sem levar em conta os sinais (isto é, "fazer de conta" que *todas* as discrepâncias têm o sinal +).

X	x
9	+4
8	+3
6	+1
4	−1
2	−3
1	−4
$\Sigma X = 30$	$\Sigma\|x\| = 16$

PASSO 3: Dividir $\Sigma|x|$ por N, a fim de garantir uma eqüidistribuição que leve em conta o número de dados envolvido.

$$DM = \frac{\Sigma|x|}{N} = \frac{16}{6} = 2{,}67$$

De acordo com o procedimento precedente, vemos que o desvio médio para o conjunto 9, 8, 6, 4, 2 e 1 é 2,67. Isso indica que, em média, os escores dessa distribuição oscilam 2,67 unidades em torno da média.

Com vistas a entender melhor a utilidade do desvio médio, vamos voltar ao problema das distribuições (a), (b) e (c) de rendas diárias, tal como figuram na Tabela 5.1.

TABELA 5.1 Variabilidade em Distribuições de Rendas Diárias, onde a Média é Fixa (Cr$340,00)

Distribuição (a)		*Distribuição* (b)		*Distribuição* (c)	
X Cr$	$\|x\|$	X Cr$	$\|x\|$	X Cr$	$\|x\|$
340	0	391	+51	595	+255
340	0	374	+34	510	+170
340	0	357	+17	425	+ 85
340	0	340	0	340	0
340	0	323	−17	255	− 85
340	0	306	−34	170	−170
340	0	289	−51	85	−255
	$\Sigma\|x\| = 0$		$\Sigma\|x\| = 204$		$\Sigma\|x\| = 1.020$
$\bar{X}$ = Cr$340		$\bar{X}$ = Cr$340		$\bar{X}$ = Cr$340	
AT = Cr$0		AT = Cr$102		AT = Cr$510	
DM = Cr$0		DM = Cr$29,14		DM = Cr$145,71	
Nenhuma Variabilidade		Pouca Variabilidade		Muita Variabilidade	

Observem, antes de mais nada, que a média de cada distribuição é Cr$340,00. Notem também que parece haver diferenças marcantes em termos de variabilidade entre essas distribuições—diferenças que podem ser detectadas com o auxílio da amplitude total e do desvio médio.

Vamos examinar, primeiro, a distribuição de rendas (a), onde todos os valores são idênticos. Uma vez que todos os escores dessa distribuição são expressos pelo mesmo valor numérico (Cr$340,00), podemos dizer que ela não apresenta variabilidade alguma. Todas as pessoas (isto é, as sete da amostra (a)) ganharam a mesma importância naquele determinado dia. Como resultado, a amplitude total é 0 (zero) e não há absolutamente nenhum desvio com relação à média (DM = 0). As distribuições (b) e (c) apresentam variabilidade. De forma específica, a distribuição (b) tem uma amplitude total de Cr$102,00 e um desvio médio de Cr$29,14; a distribuição (c) apresenta uma amplitude total de Cr$510,00 e um desvio médio de Cr$145,71. Podemos, pois, dizer que a distribuição (b) contém menos variabilidade que a distribuição (c), ou seja, as rendas que fazem parte da distribuição (b) são, entre si, mais semelhantes que as rendas da distribuição (c).

DESVIO PADRÃO

Por motivos que logo se tornarão aparentes, o desvio médio já não é muito usado pelos pesquisadores; na maioria dos casos, ele foi abandonado como "medida de variabilidade" em favor de um "parente próximo" mais eficiente: o *desvio padrão*. Como teremos a oportunidade de verificar, entretanto, o desvio médio não pode ser considerado uma perda de tempo, pois, na pior das hipóteses, ele oferece-nos uma boa base para a compreensão da natureza do desvio padrão.

Em discussões anteriores, vimos que o desvio médio evita o problema de números negativos que cancelam os positivos; isso foi conseguido graças à convenção de *ignorarem-se* os sinais "mais" (+) e "menos" (−), somando, em seguida, os valores absolutos dos desvios com relação à média. Tal procedimento para a criação de uma medida de variabilidade tem a nítida desvantagem de impedir que esses valores absolutos sejam sempre úteis em análises estatísticas mais avançadas (pois eles não comportam facilmente manipulações algébricas).

Para superar esse problema e obter uma medida de variabilidade que seja mais conveniente (isto é, ajustada, "usável") em procedimentos estatísticos avançados, podemos *elevar ao quadrado as discrepâncias reais* (com os respectivos sinais) *e somá-las a seguir* (Σx^2). Como ilustrado na Tabela 5.2, esse procedimento permite "fugir" à influência dos sinais, já que o quadrado de qualquer número é sempre *positivo*.

Após termos somado os quadrados das discrepâncias, podemos *dividir o total por N, a fim de garantir uma eqüidistribuição desse resultado (total) relativamente a todos os escores envolvidos.* O valor que se obtém dessa operação é conhecido pelo nome de *média quadrática** (ou média das discrepâncias ao quadrado). (*Nota*—Recorde-se que procedimento semelhante foi adotado no cálculo do desvio médio, quando $\Sigma|x|$ foi dividido por N.) Continuando com a ilustração da Tabela 5.2, vemos que

$$\frac{\Sigma x^2}{N} = \frac{52}{6} = 8{,}67$$

TABELA 5.2 Elevação de Discrepâncias ao Quadrado para a Eliminação de Números Negativos. (Dados da Tabela 5.1)

X	x	x^2
9	+4	16
8	+3	9
6	+1	1
4	−1	1
2	−3	9
1	−4	16
	$\Sigma x = 0$	$\Sigma x^2 = 52$

Surge, agora, um problema adicional. Como conseqüência direta do fato de termos elevado as discrepâncias ao quadrado, a unidade de medida sofreu uma alteração, o que dificulta a interpretação do resultado 8,67. De fato, temos 8,67, mas 8,67 unidades de quê? Com vistas, então, a voltar à unidade de medida original, *extraímos a raiz quadrada da média quadrática*, donde resulta:

$$\sqrt{\frac{\Sigma x^2}{N}} = \sqrt{8{,}67} = 2{,}94$$

Podemos agora definir o desvio padrão—resultado das operações atrás realizadas—como sendo a *raiz quadrada da média das discrepâncias ao quadrado.* Indicando o desvio padrão por DP ou pela letra grega minúscula σ** (sigma), vem:

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}} \quad \text{onde}$$

* N.T.: Na verdade, a *média quadrática* é muito mais conhecida pelo nome de *variância*. Em símbolos, $S^2(X) = \frac{\Sigma x^2}{N}$, onde $S^2(X)$ representa a variância. Às vezes, em lugar de $S^2(X)$ aparece $\sigma^2(X)$; quando tal ocorrer, interprete-se assim: $S^2(X)$ é a variância da *amostra* e $\sigma^2(X)$, da *população.*

** N.T.: Releia-se a nota de rodapé anterior. O que ficou dito relativamente à variância também vale para o desvio padrão. Observe-se, ainda, que o Autor não faz, a esta altura, nenhuma distinção na maneira de notar o desvio padrão, quer seja ele calculado para uma população, quer para uma amostra. Essa distinção será feita, entretanto, no Capítulo 7. Mais uma observação: o *desvio padrão* pode ser definido como sendo a *raiz quadrada positiva da variância.* Assim:

$$+\sqrt{\text{Variância}} = \text{Desvio padrão.}$$

σ = desvio padrão
Σx^2 = soma das discrepâncias ao quadrado
N = total de escores da distribuição.

Para resumir: o procedimento para o cálculo do desvio padrão não difere muito do método que vimos anteriormente para o cálculo do desvio médio. Com referência ao exemplo presente, os seguintes passos devem ser seguidos:

PASSO 1: Calcular a média da distribuição.

X
9
8
6
4
2
1
$\Sigma X = 30$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{30}{6} = 5$$

PASSO 2: Subtrair de cada escore bruto o valor da média da distribuição para, dessa forma, obterem-se as discrepâncias.

X	x
9	+4
8	+3
6	+1
4	−1
2	−3
1	−4

PASSO 3: Elevar ao quadrado cada discrepância e somá-las.

X	x	x^2
9	+4	16
8	+3	9
6	+1	1
4	−1	1
2	−3	9
1	−4	16
		$\Sigma x^2 = 52$

PASSO 4: Dividir por N e extrair a raiz quadrada do resultado.

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{\frac{52}{6}} = \sqrt{8{,}67} = 2{,}94$$

Podemos agora dizer que o desvio padrão do conjunto de escores 9, 8, 6, 4, 2 e 1 é 2,94.

Fórmula para o Cálculo do Desvio Padrão a Partir de Dados Brutos

Até este ponto, a única fórmula usada para o cálculo do desvio padrão tem sido $\sqrt{\Sigma x^2/N}$. No caso particular de estar disponível uma máquina de calcular, há um método mais fácil de obter-se o DP—um método que prescinde das discrepâncias e que trabalha diretamente com os dados brutos.

A fórmula desse novo método é

$$\sigma = \sqrt{\frac{\Sigma X^2}{N} - \bar{X}^2}$$

onde

σ = desvio padrão
ΣX^2 = soma dos quadrados dos escores brutos (*Importante: primeiro,* elevar cada escore bruto ao quadrado; *depois*, efetuar a soma.)
N = número total de escores
$\bar{X}^2$ = média aritmética ao quadrado.

O procedimento passo a passo para o cálculo do DP mediante a aplicação da fórmula anterior pode ser ilustrado a partir dos dados da Tabela 5.2.

PASSO 1: Quadrar (isto é, elevar ao quadrado) cada escore bruto separadamente e somar.

X	X^2
9	81
8	64
6	36
4	16
2	4
1	1
	$\Sigma X^2 = 202$

PASSO 2: Calcular a média aritmética e elevá-la ao quadrado.

X	
9	$\bar{X} = \frac{\Sigma X}{N} = \frac{30}{6} = 5$
8	
6	
4	$\bar{X}^2 = 25$
2	
1	
$\Sigma X = 30$	

PASSO 3: "Entrar" com os resultados dos Passos 1 e 2 na fórmula.

$$\sigma = \sqrt{\frac{\Sigma X^2}{N} - \bar{X}^2} = \sqrt{\frac{202}{6} - 25} = \sqrt{33{,}67 - 25{,}00} = \sqrt{8{,}67} =$$

$$= 2{,}94$$

Como podemos observar acima, a aplicação da nova fórmula aos dados da Tabela 5.2 produz uma resposta idêntica à obtida com o método original.

Cálculo do Desvio Padrão (DP) de uma Distribuição de Freqüências com Dados Isolados

Para obter-se o desvio padrão de uma distribuição de freqüências com dados isolados, aplique-se a fórmula

$$\sigma = \sqrt{\frac{\Sigma f X^2}{N} - \bar{X}^2}$$

Vamos, a título de ilustração, calcular o desvio padrão da seguinte distribuição:

Escores	f
7	1
6	2
5	3
4	5
3	2
2	2
1	1
	$N = 16$

PASSO 1: Multiplicar cada escore (X) pela respectiva freqüência (f), a fim de obter fX.

X	f	fX
7	1	7
6	2	12
5	3	15
4	5	20
3	2	6
2	2	4
1	1	1

PASSO 2: Multiplicar cada fX pelo respectivo X, a fim de obter fX^2. *Em seguida*, somar tudo para obter ΣfX^2.

X	fX	fX^2
7	7	49
6	12	72
5	15	75
4	20	80
3	6	18
2	4	8
1	1	1
		$\Sigma fX^2 = 303$

PASSO 3: Calcular a média aritmética e elevá-la ao quadrado.

fX
7
12
15
20
6
4
1
$\Sigma fX = 65$

$$\bar{X} = \frac{\Sigma fX}{N} = \frac{65}{16} = 4{,}06 \qquad \bar{X}^2 = 16{,}48$$

PASSO 4: "Entrar" com os resultados dos Passos 1, 2 e 3 na fórmula.

$$\sigma = \sqrt{\frac{\Sigma fX^2}{N} - \bar{X}^2} = \sqrt{\frac{303}{16} - 16{,}48} = \sqrt{18{,}94 - 16{,}48} =$$

$$= \sqrt{2{,}46} = 1{,}57$$

Significado do Desvio Padrão

Essa série de passos para o cálculo do desvio padrão pode deixar o aluno com uma desagradável sensação relacionada com o *significado* de seu resultado. Por exemplo, suponham que, numa particular distribuição de escores, o desvio padrão seja 4 ($\sigma = 4$). Que indica esse número? O que, exatamente, podemos dizer *agora* a respeito dessa distribuição que não poderíamos ter dito *antes*?

O capítulo seguinte tentará elucidar o significado concreto do desvio padrão. Por ora, estabeleçamos sumariamente que o desvio padrão (como o desvio médio) representa a "variabilidade média" de uma distribuição, já que ele mede a média de discrepâncias (desvios) com relação a $\bar{X}$. Os procedimentos de quadrar e extrair a raiz quadrada também entram no problema, mas basicamente no sentido de eliminar os sinais de "menos" e voltar a uma unidade de medida mais conveniente, vale dizer, a unidade de medida dos dados brutos.

Observamos também que quanto maior a variabilidade em torno da média de uma distribuição, maior o desvio padrão. Assim, σ = 4,5 indica maior variabilidade que σ = 2,5. Por exemplo, a distribuição da temperatura diária em Houston (Texas) tem um desvio padrão maior do que o da distribuição da temperatura diária, no mesmo período, em Honolulu (Havaí).

Se desejássemos estudar a distância que separa uma mesa da parede da sala de estar, poderíamos pensar em termos de centímetros ou metros como unidade de medida. (Por exemplo, "a mesa da sala de estar dista 5 metros desta parede".) Mas como medir a extensão da linha de base de um polígono de freqüências, onde estejam registrados os escores de um grupo de respondentes–escores classificados em ordem crescente? Ainda com relação ao mesmo assunto, que método usar para encontrar a distância entre qualquer escore bruto e a média de distribuição–método padronizado que nos possibilitasse fazer comparações entre escores brutos da mesma distribuição ou entre escores de distribuições diferentes? Se estivéssemos falando de mesas, poderíamos dizer que uma dista 5 metros da parede da sala de estar, enquanto que outra está a 10 metros da parede da cozinha. Do ponto de vista metros, dispomos de uma unidade padronizada de medida e, portanto, podemos fazer tais comparações de modo objetivo. Mas como proceder na comparação de escores brutos? Por exemplo, será que é sempre possível comparar a nota 85–resultante de um exame de inglês–com a nota 80–resultante de um de alemão? Qual dessas notas é realmente mais alta? Um pouco de atenção vai demonstrar que isso depende fundamentalmente da maneira como se saíram os *outros alunos* em cada uma dessas disciplinas.

Uma forma de obter-se um indicador grosseiro da extensão (largura) de uma linha de base é calcular a amplitude total, uma vez que ela fornece a distância entre o escore mais alto e o mais baixo. Tal medida porém não serve para localizar, com eficácia, um escore particular com relação à média, pois–deixando de lado outras fragilidades dessa medida–ela cobre *toda a extensão* da linha de base. Em contraposição, o "tamanho" do desvio padrão é menor do que o da amplitude total e geralmente cobre muito menos do que toda a extensão da linha de base.

Da mesma forma que medimos o tamanho de um tapete em metros ou centímetros, também podemos medir a linha de base em unidades de desvio padrão (isto é, em

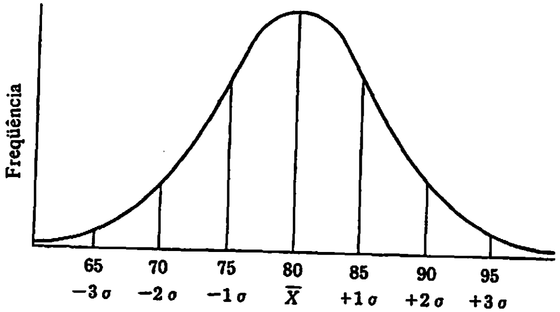


FIGURA 5.2 Mensuração da Linha de Base em Unidades de Desvio Padrão. (Neste exemplo, o desvio padrão (σ) é 5 e a média ($\bar{X}$) é 80)

unidades "sigma"). Por exemplo, poderíamos somar um desvio padrão ao valor da média, a fim de determinar qual escore bruto localiza-se a exatamente um desvio padrão dela (isto é, um escore bruto distante da média *um sigma*). Como ficou demonstrado na Figura 5.2, se $\bar{X} = 80$ e DP = 5, então o escore bruto 85 fica a exatamente um desvio padrão *acima* da média (80 + 5 = 85); isso equivale a dizer que o escore 85 dista dela $+1\sigma$. Essa direção é "mais" (+) porque todas as discrepâncias *acima* da média são positivas; todas as discrepâncias *abaixo* da média são "menos" (−) ou negativas.

Podemos continuar "graduando" a linha de base mediante a adição de mais um desvio padrão ao escore bruto 85. Tal procedimento dá-nos o escore bruto 90, que se situa a exatamente dois desvios padrões acima da média (85 + 5 = 90). Por igual procedimento, adiciona-se um desvio padrão ao escore bruto 90 e obtemos 95—que representa o escore bruto localizado a exatamente três desvios padrões acima da média. Para continuar esse processo do "lado negativo" (abaixo da média), vamos subtraindo, a partir de $\bar{X}$, um desvio padrão, dois desvios padrões etc., donde resulta: 75 (localizado a -1σ), 70 (localizado a -2σ), e assim por diante.

Como bem ilustra a Figura 5.3, o processo de medir a linha de base em unidades de desvio padrão é, em muitos aspectos, semelhante à mensuração da distância em metros entre uma mesa e uma dada parede. No entanto, a analogia falha em pelo menos um

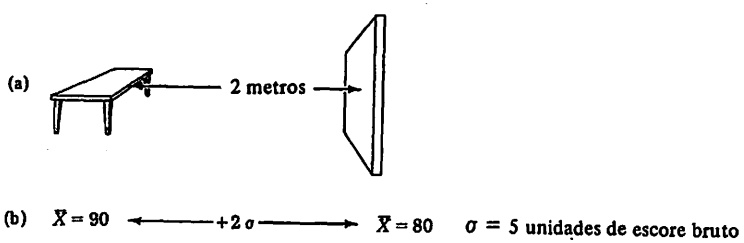


FIGURA 5.3 Mensuração de Distâncias: (a) entre uma mesa e uma parede, onde a unidade de medida usada é o metro; (b) entre um escore bruto e a média da distribuição, onde a unidade de medida vem expressa em unidades de desvio padrão

ponto importante: enquanto que metros, centímetros sempre têm *tamanho constante*, o valor do desvio padrão varia de distribuição para distribuição. Aliás, se não fosse assim, não poderíamos usar o desvio padrão—como ilustramos páginas atrás—na comparação de distribuições quanto à sua variabilidade (por exemplo, DP = Cr$ 85.000,00 para a distribuição da renda de professores secundários; DP = Cr$ 255.000,00 para a distribuição da renda de larápios). Por essa razão, precisamos calcular o "tamanho" do desvio padrão de *todas* as distribuições com que estejamos trabalhando. Também em conseqüência deste fato, é geralmente mais difícil compreender o significado do desvio padrão, em contraste com o significado de metros ou centímetros como unidades de mensuração. No capítulo seguinte voltaremos a examinar o conceito de desvio padrão.

COMPARAÇÃO ENTRE AMPLITUDE TOTAL, DESVIO MÉDIO E DESVIO PADRÃO

A amplitude total pode ser considerada um mero índice preliminar ou grosseiro da variabilidade de uma distribuição. Sua obtenção é rápida e fácil; todavia, ela não é muito "confiável". Pode ser aplicada a dados intervalares ou ordinais.

A amplitude total tem utilidade no que diz respeito ao cálculo do desvio padrão. A Figura 5.2 ilustra que seis desvios padrões cobrem quase que inteiramente a distância que vai desde o menor escore até o maior numa distribuição (de -3σ a $+3\sigma$). Esse fato isolado já nos dá um método conveniente para estimar (mas *não* calcular) o desvio padrão. De modo geral, o tamanho do desvio padrão é cerca de um sexto ($^1/_6$) do tamanho da amplitude total. Por exemplo, se a amplitude total for 36, espera-se que o DP caia bem perto de 6; se a amplitude total for 6, o DP deve de estar próximo de 1.

Essa regra pode assumir considerável importância aos olhos do aluno que deseja "testar" se seu resultado tem grande probabilidade de estar correto. Tomemos um caso extremo: se $AT = 10$ e o DP (calculado) = 12, devemos ter cometido algum erro, pois o DP não pode ser maior que a amplitude total. Uma nota de alerta: a regra do "um sexto" só se aplica a distribuições compostas de grande número de escores. Quando o número de escores for pequeno, via de regra, haverá necessidade de menos desvios padrões para cobrir toda a amplitude da distribuição.

Enquanto que a amplitude total é calculada a partir de *apenas dois escores*, tanto o desvio padrão quanto o desvio médio levam em conta *todos os escores* da distribuição. Apesar da sua relativa estabilidade, o desvio médio não é muito usado em pesquisa, uma vez que ele não se presta a muitas análises estatísticas avançadas. Por outro lado, o cálculo do desvio padrão implica a utilização de um procedimento aceitável do ponto de vista matemático, com vistas a contornar o problema dos sinais (+; −): quadrar as discrepâncias em vez de simplesmente ignorar os sinais. Como resultado, o desvio padrão acabou por tornar-se o passo inicial na obtenção de certas medidas estatísticas, em particular no terreno das inferências. Essa característica do desvio padrão será explorada em detalhes em capítulos subseqüentes, especialmente nos de números 6 e 7.

Apesar da sua utilidade como medida de dispersão "confiável", o desvio padrão também apresenta algumas desvantagens. Comparado com outras medidas de variabilidade, seu cálculo é não só *difícil* como *demorado*. Todavia, essas desvantagens estão dia a dia sendo superadas pelo crescente uso ou de máquinas de calcular de alta velocidade ou de computadores programados para a execução de análises estatísticas. O desvio padrão (da mesma forma que o desvio médio) também tem a característica de ser uma medida intervalar, donde decorre que sua utilização torna-se *impossível* com dados *nominais* ou *intervalares*–via de regra utilizados no trabalho de muitos pesquisadores.

CÁLCULO DA AMPLITUDE TOTAL, DO DESVIO MÉDIO E DO DESVIO PADRÃO DE DADOS AGRUPADOS

Estejam os dados agrupados ou não, a amplitude total é sempre a diferença entre o maior e o menor escore. Seu cálculo não requer o uso de nenhum método especial ou fórmula.

Com o propósito de ilustrar passo a passo o procedimento para a obtenção do desvio médio numa distribuição de freqüências com dados agrupados, considere-se a seguinte tabela*:

Intervalo de Classe	f
17–19	1
14–16	2
11–13	3
8–10	5
5–7	4
2–4	2
	$N = 17$

PASSO 1: Calcular o ponto médio de cada intervalo de classe.

Intervalo	*X = Ponto Médio*
17–19	18
14–16	15
11–13	12
8–10	9
5–7	6
2–4	3

PASSO 2: Determinar a média da distribuição.

X = ponto médio	f	fX	
18	1	18	$\bar{X} = \frac{\Sigma fX}{N}$
15	2	30	
12	3	36	
9	5	45	$= \frac{159}{17}$
6	4	24	
3	2	6	$= 9{,}35$
		$\Sigma fX = 159$	

* N.T.: É sempre interessante especificar, na coluna onde figuram os intervalos de classe, a *unidade* em que a variável foi mensurada.

PASSO 3: Calcular o desvio (= a discrepância) de cada ponto médio com relação à média.

X = Ponto Médio	$X - \bar{X} = \lvert x\rvert$
18	8,65
15	5,65
12	2,65
9	0,35
6	3,35
3	6,35

PASSO 4: Multiplicar cada discrepância pela freqüência correspondente ao intervalo de classe e, em seguida, somar esses produtos.

Intervalo	f	x	$f\,x$
17–19	1	8,65	8,65
14–16	2	5,65	11,30
11–13	3	2,65	7,95
8–10	5	0,35	1,75
5–7	4	3,35	13,40
2–4	2	6,35	12,70
	$N = 17$		$\Sigma f\,x = 55{,}75$

PASSO 5: Dividir por N.

$$\text{DM} = \frac{\Sigma f\lvert x\rvert}{N} = \frac{55{,}75}{17} = 3{,}28$$

Chegamos, assim, a um desvio médio de 3,28.

Para o cálculo do desvio padrão de dados agrupados, podemos usar uma fórmula que se utiliza apenas dos dados brutos da tabela inicial. Em símbolos, temos:

$$\sigma = \sqrt{\frac{\Sigma fX^2}{N} - \bar{X}^2}$$

onde

σ = desvio padrão
f = freqüências dos intervalos de classe
X = ponto médio de cada intervalo de classe
N = número total de escores
$\bar{X}^2$ = quadrado da média (aritmética).

O procedimento passo a passo para o cálculo do desvio padrão pode ser ilustrado, por exemplo, com relação à seguinte tabela de dados agrupados:

Intervalo de Classe	*f*
17–19	1
14–16	2
11–13	3
8–10	5
5–7	4
2–4	2

PASSO 1: Multiplicar cada ponto médio pela freqüência do intervalo de classe correspondente e somar esses produtos.

Intervalo de Classe	*f*	*Ponto Médio (X)*	*fX*
17–19	1	18	18
14–16	2	15	30
11–13	3	12	36
8–10	5	9	45
5–7	4	6	24
2–4	2	3	6
	$N = 17$		$\Sigma fX = 159$

PASSO 2: Calcular a média aritmética e elevá-la ao quadrado.

$$\bar{X} = \frac{\Sigma fX}{N} = \frac{159}{17} = 9{,}35$$

$$\bar{X}^2 = 87{,}42$$

PASSO 3: Multiplicar cada ponto médio por fX e somar esses produtos.

Intervalo de Classe	*f*	*Ponto médio* (X)	*fX*	*fX*²
17–19	1	18	18	324
14–16	2	15	30	450
11–13	3	12	36	432
8–10	5	9	45	405
5–7	4	6	24	144
2–4	2	3	6	18
				$\Sigma fX^2 = 1773$

PASSO 4: "Entrar" com os resultados dos Passos 2 e 3 na fórmula.

$$\sigma = \sqrt{\frac{\Sigma fX^2}{N} - \bar{X}^2} = \sqrt{\frac{1773}{17} - 87{,}42} = \sqrt{104{,}29 - 87{,}42} =$$

$$= \sqrt{16{,}87} = 4{,}11$$

O desvio padrão, conforme sugere a própria fórmula, é 4,11.

RESUMO

Introduzimos, neste capítulo, o conceito de *amplitude total*, de *desvio médio* e de *desvio padrão*—que são três medidas de variabilidade ou, para dizer de outra forma, medidas que "mostram" a maneira como os escores agrupam-se em torno do centro da distribuição. A amplitude total foi definida como um indicador de variabilidade rápido e grosseiro, facilmente calculada a partir da diferença entre o maior e o menor escore da distribuição. O desvio médio—ou seja, a soma dos valores absolutos das discrepâncias (desvios), dividida por N—foi, desde o início, tratado como uma medida inadequada de variação, do ponto de vista matemático, apesar de constituir sólida base para a compreensão do desvio padrão; este, por sua vez, foi definido como sendo a raiz quadrada da variância da distribuição. O desvio padrão é uma medida de variabilidade *confiável*, de nível intervalar, que pode ser utilizada em operações estatísticas avançadas, descritivas ou inferenciais. O significado amplo do desvio padrão será objeto de estudos posteriores: nas discussões relacionadas com a curva normal e nas generalizações de amostras para populações.

PROBLEMAS

1. Um grupo de 5 estudantes obteve as seguintes notas de exame, numa escala de 10 pontos: 7, 5, 3, 2 e 1. Relativamente a esse conjunto de escores, calcule (a) a amplitude total; (b) o desvio médio; (c) o desvio padrão.
2. Numa escala destinada a mensurar atitudes para com o problema da segregação racial, duas classes de universitários obtiveram os seguintes pontos:

Classe A	*Classe B*
4	3
6	3
2	2
1	1
1	4
1	2

Compare a variabilidade de atitudes relativamente à segregação racial entre os membros das classes A e B mediante a determinação do seguinte: (a) amplitude total para cada classe; (b) o desvio médio dos escores de cada classe; (c) o desvio padrão de cada classe. Qual das classes apresenta maior variabilidade de atitudes?

3. Para o seguinte grupo de escores: 3, 5, 5, 4, 1, calcule (a) a amplitude total; (b) o desvio médio; (c) o desvio padrão.

4. Dado o conjunto de escores: 1, 6, 6, 3, 7, 4, 10, calcule o desvio padrão.

5. Dado o conjunto de escores: 12, 12, 10, 9, 8, calcule o desvio padrão.

6. Calcule o desvio padrão relativo à seguinte distribuição freqüencial de escores:

X	f
5	3
4	5
3	6
2	2
1	2
	$N = 18$

7. Calcule o desvio padrão relativo à seguinte distribuição freqüencial de escores:

X	f
7	2
6	3
5	5
4	7
3	4
2	3
1	1
	$N = 25$

8. Calcule o desvio padrão relativo à seguinte distribuição freqüencial de escores:

X	f
10	2
9	5
8	8
7	7
6	4
5	3
	$N = 29$

9. Calcule, relativamente à seguinte tabela de dados agrupados por freqüências, (a) a amplitude total; (b) o desvio médio; (c) o desvio padrão.

Intervalo de Classe	f
90–99	6
80–89	8
70–79	4
60–69	3
50–59	2
	$N = 23$

10. Relativamente à seguinte tabela de dados agrupados por freqüências, calcule: (a) a amplitude total; (b) o desvio médio; (c) o desvio padrão.

Intervalo de Classe	f
17–19	2
14–16	3
11–13	6
8–10	5
5–7	1

11. Relativamente à seguinte tabela de dados agrupados por freqüências, calcule: (a) a amplitude total; (b) o desvio médio; (c) o desvio padrão.

Intervalo de Classe	f
20–24	2
15–19	4
10–14	8
5–9	5
	$N = 19$

PARTE II
Da Descrição à Tomada de Decisões

6

A Curva Normal

Nos capítulos anteriores, vimos que as distribuições de freqüências podem assumir uma grande variedade de formas. Algumas são *simétricas* ou livres de alongamento; outras são negativa ou positivamente alongadas; outras, ainda, apresentam-se com mais de uma "corcova", e assim por diante. Apesar dessa grande diversidade, há uma distribuição de freqüências com a qual muitos de nós já nos familiarizamos, se não por outra razão, pelo fato de termos sido classificados de acordo com ela por nossos professores. Essa distribuição, comumente chamada *curva normal*, é um modelo teórico ou ideal que resulta muito mais de uma equação matemática do que de um real delineamento de pesquisa com posterior coleta de dados.[1] Entretanto, a utilidade da curva normal para o pesquisador pode ser evidenciada através de suas aplicações a efetivas situações de pesquisa.

Como veremos neste capítulo, por exemplo, a curva normal pode ser usada na descrição de distribuições de escores, na interpretação do desvio padrão e em afirmações relacionadas com a noção de probabilidade. Em capítulos subseqüentes, veremos que a curva normal constitui um ingrediente essencial para a tomada de decisões estatísticas, a partir da qual o pesquisador pode generalizar para populações as conclusões a que tenha chegado ao lidar com amostras. Antes de prosseguirmos com a discussão de técnicas para tomada de decisões, faz-se necessário que conheçamos alguns dados relacionados com as propriedades da curva normal.

---

[1] A curva normal pode ser construída a partir da seguinte fórmula:

$$Y = \frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\bar{X}}{\sigma}\right)^2}, \text{ onde}$$

Y = ordenada correspondente a um dado valor de X (isto é, a ordenada representa a freqüência com que X ocorre na distribuição)
π = 3,1416
e = 2,7183

CARACTERÍSTICAS DA CURVA NORMAL

Como pode ser caracterizada a curva normal? Quais as propriedades que a distinguem de outras distribuições? Tal como indica a Figura 6.1, a curva normal é um tipo de curva simétrica, suave, cuja forma lembra um sino e, por isso, é amplamente conhecida por "curva em sino". É possível que o aspecto mais marcante dessa curva seja a sua *simetria*; se "dobrássemos" a curva em seu ponto central (que corresponde à freqüência máxima), daríamos origem a duas metades, sendo que cada uma delas seria a imagem espelhada da outra.*

Além disso, a curva normal é *unimodal*, isto é, possui um só (pico ou) ponto de freqüência máxima; esse ponto, por sua vez, é aquele situado no meio da distribuição (curva), em que a média, a mediana e a moda coincidem. (Recorde-se que, conforme o Capítulo 3, a média, a mediana e a moda ocorrem em pontos diferentes (isto é, não coincidentes) numa distribuição alongada (assimétrica).) A partir do topo (central, arredondado), a curva normal "cai" gradualmente até formar as caudas (duas, uma de cada lado), que se estendem de forma indefinida, aproximando-se cada vez mais da linha de base (eixo das abscissas) sem, entretanto, jamais tocá-la.**

CURVAS NORMAIS: O MODELO E O MUNDO REAL

Poderíamos perguntar: até que ponto as distribuições de dados reais (ou seja, de dados coletados por pesquisadores ao longo de suas pesquisas) ajustam-se ou aproximam-se à figura (ao formato) da curva normal? A título de ilustração, vamos imaginar que todos os fenômenos sociais, psicológicos e físicos sejam normalmente distribuídos. Como seria esse mundo hipotético?

Se atentássemos para as características físicas dos seres humanos, estatura, por

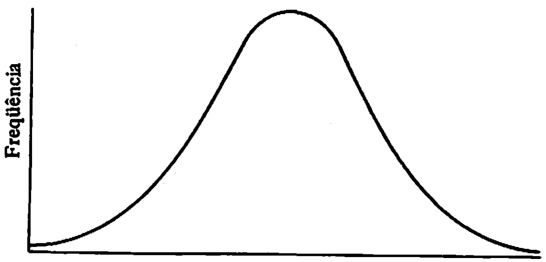


FIGURA 6.1 Formato da Curva Normal

* N.T.: Essa "dobra" representa, de maneira mais formal, um *eixo de simetria*. Tomando-se esse eixo como referência, observa-se que a curva normal é *bicaudal*. Então, para retomar a idéia do autor, dizer que uma metade é a imagem espelhada da outra é o mesmo que dizer o seguinte: "dobrando-se" a curva em seu eixo de simetria, a cauda da *esquerda* ajusta-se (sobrepõe-se) perfeitamente à cauda da *direita*.

** N.T.: Diz-se, por essa razão, que o *campo de variação* de uma distribuição normal estende-se de $-\infty$ a $+\infty$

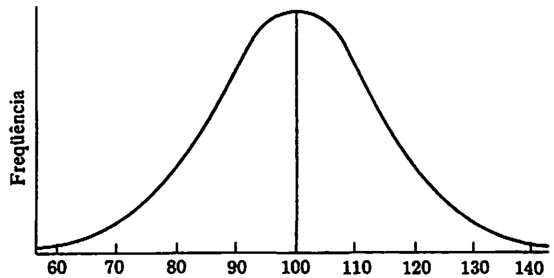


FIGURA 6.2 Distribuição Hipotética de Quocientes de Inteligência (QIs)

exemplo, veríamos que a maioria dos adultos estaria na faixa que vai de 152 cm (aprox.) até 183 cm (aprox.), com muito pouca gente menor que 152 cm ou maior que 183 cm. A Figura 6.2 demonstra que o QI também seria previsível–a maioria dos QIs situando-se entre 90 e 110; como podemos observar, há uma "descida" gradual dos escores para ambas as caudas, com pouquíssimos "gênios" que têm QI superior a 140 e, da mesma forma, pouquíssimas criaturas menos privilegiadas, cujos QIs estão abaixo de 60. Por igual raciocínio, relativamente poucos sujeitos poderiam ser considerados políticos extremistas–de direita ou de esquerda–enquanto que a tendência política da maioria seria considerada moderada.

Finalmente, mesmo o desgaste dos pisos, resultante do fluxo de transeuntes, lembra a distribuição normal: a maior parte do desgaste ocorre no centro dos pisos (degraus etc.), enquanto que nos lados, à medida que nos afastamos do centro, o desgaste vai-se tornando cada vez menor.

Os leitores terão, a esta altura, observado que o mundo hipotético da curva normal não difere de forma radical do mundo "real" em que vivemos no momento. Fenômenos tais como estatura, QI, orientação política, desgaste dos pisos etc. aproximam-se, na prática, até que muito bem da distribuição normal teórica. Pelo fato de tantos fenômenos terem essa característica–isto é, pelo fato de ela ocorrer tão freqüentemente na natureza (e por outras razões que logo se tornarão aparentes)–pesquisadores de diferentes campos têm feito uso extensivo da curva normal, aplicando-a aos dados que eles coletam e analisam.

Observe-se, porém, que alguns fenômenos no campo social–como em qualquer outro–simplesmente não se ajustam à noção teórica da distribuição normal. Muitas distribuições são assimétricas; outras têm mais de uma moda; outras são simétricas, mas não têm a forma de "sino". Como exemplo concreto, consideremos a distribuição de riqueza no mundo. É fato bem conhecido que "os que têm" superam de longe "os que não têm". Assim, como ilustra a Figura 6.3, a distribuição de riquezas (indicada pela renda *per capita*) é de extrema assimetria (pelo menos na aparência), de sorte que apenas uma pequena proporção da população mundial recebe porção significativa da renda total. De forma análoga, especialistas em demografia dizem-nos que os Estados Unidos da América do Norte tornaram-se, nos últimos tempos, uma terra de jovens e velhos. Do ponto de vista econômico, essa distribuição de idades representa um fardo pesado

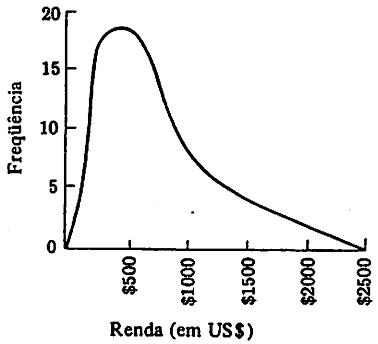


FIGURA 6.3 Distribuição da Renda "Per Capita" (Nações do Mundo)

para um grupo relativamente pequeno de trabalhadores, isto porque, sendo todos cidadãos de meia-idade, têm a seu encargo um número assustador tanto de velhos (aposentados) quanto de jovens (ainda em período escolar).

Naquelas circunstâncias em que temos boas razões para esperar grandes divergências da normalidade–como, por exemplo, no caso da idade e da renda–a curva normal não pode ser usada como "modelo" para os dados coletados. Vemos, assim, que não é possível aplicá-la com liberdade a todas as distribuições que o pesquisador obtém, e deve, ao contrário, ser usada com uma boa dose de bom senso. Felizmente os estatísticos sabem que grande quantidade de fenômenos de interesse segue o modelo normal.

A ÁREA SOB A CURVA NORMAL

A fim de podermos empregar a curva normal na solução de problemas, precisamos, antes, aprender o significado da expressão "área sob a curva normal": *é aquela porção do plano, compreendida entre a curva e a linha de base, que corresponde, em qualquer distribuição normal, a 100% dos dados considerados.* A Figura 6.4 ilustra essa característica.

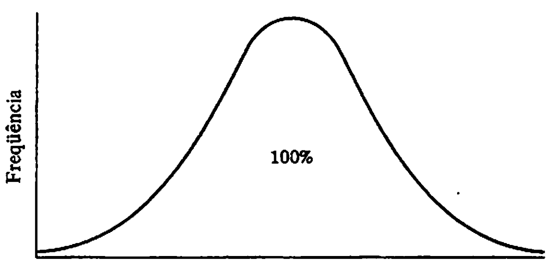


FIGURA 6.4 A Área sob a Curva Normal

Poderíamos limitar uma porção dessa área total se traçássemos, a partir de dois pontos quaisquer tomados na linha de base, segmentos perpendiculares até a própria curva. Por exemplo, usando a média como ponto de partida, poderíamos traçar um segmento em $\bar{X}$ e um outro no ponto que coincide com 1 DP (1 distância sigma) acima (isto é, à direita) de $\bar{X}$. Esta porção sombreada da curva normal ilustrada na Figura 6.5 abrange 34,13% da freqüência total.

Da mesma forma, podemos dizer que 47,72% dos sujeitos sob a curva normal caem entre $\bar{X}$ e 2 DP acima (à direita) da média; igualmente, que 49,87% caem entre $\bar{X}$ e 3 DP acima da média (ver Figura 6.6).

Como veremos mais adiante, *uma proporção constante da área total sob a curva normal cairá entre a média e qualquer distância dada a contar de $\bar{X}$–desde que a mensuração seja feita em unidades de desvio padrão.* Esta afirmação é verdadeira, independentemente da média e do desvio padrão da distribuição particular que estejamos estudando e aplica-se a todos e quaisquer dados que tenham distribuição normal. Assim, a área sob a curva normal compreendida entre $\bar{X}$ e 1 DP acima da média vai *sempre* incluir 34,13% da totalidade dos dados, quer estejamos discutindo a distribuição de estaturas, de inteligência, de filiação partidária ou de desgaste nos degraus das portas à entrada de nossas casas. O requisito básico, em cada situação, é que estejamos trabalhando com escores que tenham *realmente* distribuição *normal.*

A natureza simétrica da curva normal leva-nos a tirar outra conclusão importante: *qualquer distância medida em "sigmas", acima ou abaixo da média, contém a mesma porção da área sob a curva.* Então, se 34,13% da área total situam-se entre a média e 1 DP *acima* de $\bar{X}$, também 34,13% da área total situam-se entre a média e 1 DP *abaixo* de $\bar{X}$; se 47,72% situam-se entre a média e 2 DP *acima* de $\bar{X}$, também 47,72% situam-se entre a média e 2 DP *abaixo* de $\bar{X}$; finalmente, se 49,87% situam-se entre a média e 3 DP *acima* de $\bar{X}$, também 49,87% situam-se entre a média e 3 DP *abaixo* de $\bar{X}$. Em outras palavras, como ilustra a Figura 6.7, 68,26% da área total sob a curva normal (34,13% + 34,13% = 68,26%) caem entre -1σ e $+1\sigma$, sendo a média (aritmética), $\bar{X}$, o ponto de referência; 95,44% da área total (47,72% + 47,72%) caem entre -2σ e $+2\sigma$ a partir de $\bar{X}$; 99,74% da área total–que, aliás, é praticamente toda a área sob a curva–caem entre -3σ e $+3\sigma$ (sempre $\bar{X}$ como ponto de partida). Podemos

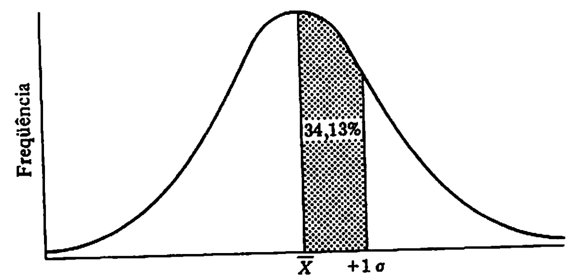


FIGURA 6.5 Porcentagem da Área Total sob a Curva Normal Limitada por $\bar{X}$ e por um Desvio Padrão Acima de $\bar{X}$

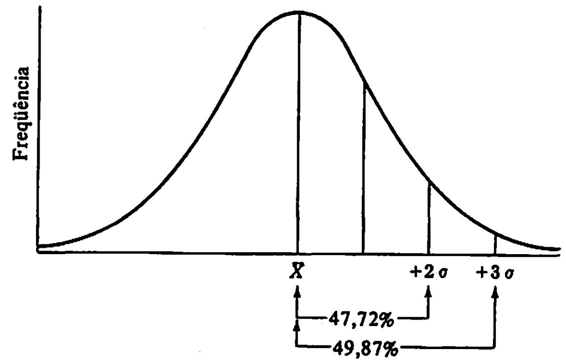


FIGURA 6.6 **Porcentagem da Área Total sob a Curva Normal Limitada por $\bar{X}$ e por Dois Desvios Padrões Acima de $\bar{X}$.**

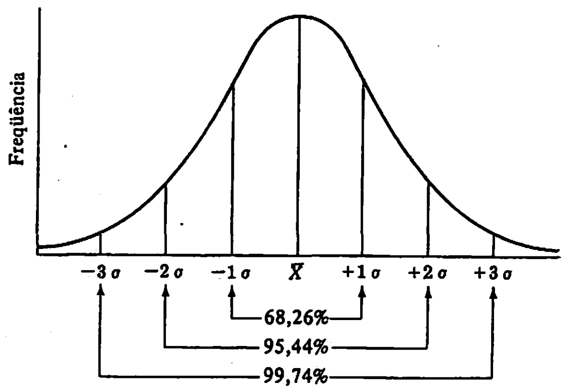


FIGURA 6.7 **Porcentagens da Área Total sob a Curva Normal Compreendidas Entre $\pm 1\sigma$, $\pm 2\sigma$ e $\pm 3\sigma$**

dizer, então, que 6 DPs abarcam, para efeitos práticos, a totalidade dos dados sob qualquer curva normal (mais de 99%).

TORNANDO A NOÇÃO DE DESVIO PADRÃO MAIS CLARA: UMA ILUSTRAÇÃO

Uma importante característica da curva normal é auxiliar na interpretação e compreensão do desvio padrão. Para entendermos melhor essa característica, vamos examinar o que os antropólogos nos dizem a respeito de diferenças de QIs ligadas a sexo. Muito embora isso contrarie a opinião dos chauvinistas, há provas de que homens e mulheres têm um QI médio aproximadamente igual a 100. Diga-se também, de passagem, que esses QIs diferem acentuadamente em termos de variabilidade em torno da média.

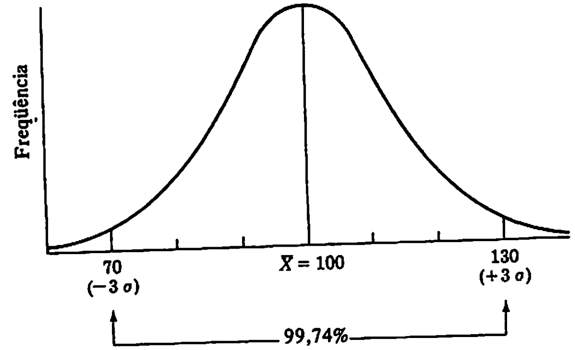


FIGURA 6.8 **Distribuição de QIs Masculinos**

Neste caso específico, vamos admitir que os QIs masculinos apresentem maior *heterogeneidade* do que os femininos; em outras palavras, a distribuição dos QIs masculinos contém uma porcentagem maior de escores extremos—representativos de sujeitos brilhantes e de sujeitos medíocres—enquanto que a distribuição de QIs femininos contém uma porcentagem maior de escores localizados próximos à média, isto é, ao redor do ponto de freqüência máxima, no centro.

Em virtude de o desvio padrão ser uma medida de variabilidade, essas diferenças ligadas a sexo deveriam refletir-se no valor do DP de cada distribuição de QIs. Poderíamos, assim, verificar, por exemplo, que DP = 10 para sujeitos do sexo masculino e que DP = 5 para os do sexo feminino.

Se conhecêssemos o desvio padrão de cada conjunto de escores QI e admitíssemos que cada conjunto tivesse distribuição normal, poderíamos estimar e, em seguida, comparar as porcentagens de machos e fêmeas localizadas numa dada amplitude de QIs.

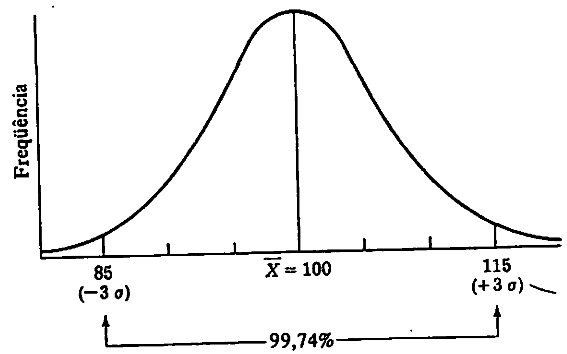


FIGURA 6.9 **Distribuição de QIs Femininos**

Por exemplo, se medirmos a linha base da distribuição de QIs masculinos em unidades DP (desvio padrão), ficaremos sabendo que 68,26% dos escores caem entre -1σ e $+1\sigma$, a partir da média. Como o desvio padrão é sempre expresso em unidades brutas e $\sigma = 10$, automaticamente descobrimos onde se localizam, na distribuição, os escores 90 e 100 (ou seja: $\bar{X} \pm \sigma = X$; portanto $100 - 10 = 90$ e $100 + 10 = 110$). Desse modo, 68,25% dos representantes do sexo masculino terão, nas condições propostas, QIs situados entre 90 e 110.

Distanciando-nos progressivamente de $\bar{X}$ no sentido das caudas, verificamos, tal como ilustra a Figura 6.8, que 99,74%, isto é, a quase totalidade dos sujeitos (machos), tem QIs entre 70 e 130 (entre -3σ e $+3\sigma$).

Da mesma forma, se examinarmos, a seguir, a distribuição dos QIs femininos—tal como vem ilustrado na Figura 6.9— verificaremos que 99,74% dos dados cairão entre os escores (QIs) 85 e 115 (também entre -3σ e $+3\sigma$). Assim, se compararmos as duas distribuições, veremos que a dos QIs femininos apresenta-se relativamente *homogênea*, uma vez que possui menor proporção de escores extremos em ambas as caudas. Tal diferença reflete-se no "tamanho relativo" de cada DP, e nos escores QIs que caem entre -3σ e $+3\sigma$, a contar da média.

USO DA TABELA B

Na discussão da distribuição normal lidamos, até este ponto, apenas com aquelas distâncias, contadas a partir da média, que são múltiplos exatos do desvio padrão. Em outras palavras, até aqui só trabalhamos com DPs iguais a 1, 2 ou 3, acima ou abaixo de $\bar{X}$. O problema que se coloca agora é o seguinte: que devemos fazer para determinar a porcentagem de sujeitos que se situam entre *duas quaisquer ordenadas* (não expressas por *números inteiros*)? Por exemplo, suponha o leitor que desejemos estabelecer a porcentagem da freqüência total que cai entre a média e, digamos, um escore bruto localizado a 1,40 DPs acima da média. Como bem demonstra a Figura 6.10, um escore bruto situado a 1,40 DPs acima da média é obviamente maior que 1 DP, porém menor que 2 DPs—contados, ambos a partir de $\bar{X}$. Sabemos, assim, que essa distância, calculada

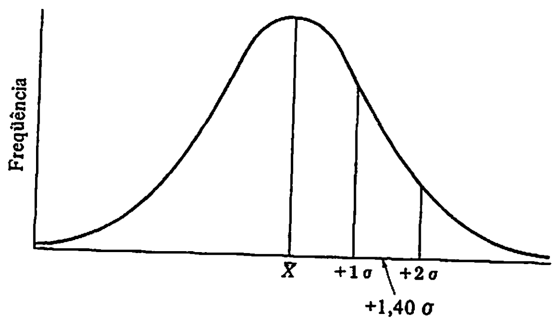


FIGURA 6.10 Localização de um Escore (Bruto) Localizado a 1,40 DPs Acima de $\bar{X}$

a partir da média, deveria incluir *mais* de 34,13%, porém *menos* de 47,72% do total da área sob a curva normal.

Para calcular a porcentagem *exata* (dos sujeitos que caem dentro) desse intervalo, devemos empregar a Tabela B do final do livro, tabela essa que fornece a porcentagem sob a curva normal entre a média e diversas distâncias sigmas (medidas a partir de $\bar{X}$). Tais distâncias sigmas (de 0,0 a 5,0) aparecem na coluna do lado esquerdo da Tabela B e foram calculadas até a primeira casa decimal. A segunda casa decimal aparece na primeira linha, no topo da tabela.

Observe-se que a simetria da curva normal torna possível calcular as porcentagens para *apenas um dos lados* da média*, ou seja, relativamente à metade da área da curva (50%). Os valores da Tabela B correspondem a qualquer dos lados.** Abaixo figura um extrato dessa tabela.

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	00,00	00,40	00,80	01,20	01,60	01,99	02,39	02,79	03,19	03,59
0,1	03,98	04,38	04,78	05,17	05,57	05,96	06,36	06,75	07,14	07,53
0,2	07,93	08,32	08,71	09,10	09,48	09,87	10,26	10,64	11,03	11,41
0,3	11,79	12,17	12,55	12,93	13,31	13,68	14,06	14,43	14,80	15,17
0,4	15,54	15,91	16,28	16,64	17,00	17,36	17,72	18,08	18,44	18,79

Para aprender a usar e entender a Tabela B, poderíamos, primeiro, tentar localizar a porcentagem de dados situados entre uma distância sigma igual a 1,0 e a média. (A razão disso. é simples: já sabemos que 34,13% da área total caem entre esses dois pontos na linha base.) Ao observar a Tabela B, verificamos que, de fato, ela indica que 34,13% da freqüência total caem entre a média e uma distância sigma igual a 1,00. Da mesma forma, vemos que a distância sigma igual a 2,00 abarca exatamente 47,72% da área total sob a curva, enquanto que a distância sigma igual a 2,01 corresponde a 47,78% dessa área total.

ESCORES PADRONIZADOS E A CURVA NORMAL

Estamos agora preparados para calcular a porcentagem da área total, sob a curva normal, associada a qualquer distância sigma que seja tomada a partir da média. Entretanto, ainda resta uma questão importante por responder: como determinamos a distância sigma de qualquer escore dado? Em outras palavras, como fazemos para traduzir nosso escore bruto—aquele que foi originalmente colhido no contato com os respondentes, por exemplo—em unidades de desvio padrão? Se desejássemos traduzir pés em jardas, simplesmente dividiríamos o número de pés por 3, uma vez que há 3 pés numa jarda.*** Da mesma forma, se estivéssemos traduzindo minutos em horas, dividiríamos o número de minutos por 60, já que há 60 minutos em cada hora. De maneira idêntica, podemos

* N.T.: Por questão de comodidade, o lado escolhido é o da *direita*, já que, assim, nenhum valor precisa vir antecedido de sinal.

** N.T.: Observe-se o que ficou dito na N. do T. anterior.

*** N.T.: 1 pé = 30,48 cm. 1 jarda = 3 pés e, portanto, igual a 3(30,48 cm) = 91,44 cm.

traduzir qualquer escore bruto em unidades de desvio padrão, dividindo a distância* que separa esse escore da média ($\bar{X}$) pelo desvio padrão da distribuição. Para ilustrar: vamos imaginar um escore bruto igual a 6 numa distribuição cuja média = 3 e cujo desvio padrão = 2. Calculando-se a diferença entre esse escore bruto e a média, obtém-se um escore-diferença representado por (6 − 3); esse cálculo simples mostra que um escore bruto igual a 6 está a 3 unidades acima da média–sendo que essa "medida" também é expressa em escores brutos. Se dividirmos esse escore-diferença (discrepância) pelo desvio padrão, que é igual a 2, veremos que o escore bruto "6" está a 1,5 (um e meio) desvios padrões acima da média. Em outras palavras, a distância sigma de um escore igual a 6, *nesta distribuição particular*, é 1,5. Note-se que, independentemente da "situação de mensuração", sempre há 3 pés numa jarda e 60 minutos numa hora. A constância que caracteriza todas essas medidas padronizadas não se aplica ao desvio padrão, uma vez que ele varia de distribuição para distribuição. Por essa razão, precisamos conhecer o desvio padrão de uma distribuição ou por *cálculo* ou por *estimativa* ou por *informação* de terceiros sempre que tenhamos em mente traduzir um determinado escore bruto em unidades de desvio padrão.

O processo que acabamos de ilustrar–ou seja, o cálculo da distância sigma a partir de $\bar{X}$–produz um valor chamado *escore z* ou *escore padronizado*, que indica, *em unidades de desvio padrão, o sentido e o grau com que um dado escore bruto se afasta da média da distribuição à qual ele pertence.* (Observe-se que a coluna do lado esquerdo da Tabela B no fim do livro vem encimada com *z*.) Assim, um escore *z* de +1,4 indica que o escore bruto fica a 1,4 DPs (ou quase 1½ DPs) *à direita* (acima) da média, enquanto que um escore *z* de −2,1 significa que o escore bruto correspondente cai *à esquerda* (abaixo) da média, num ponto ligeiramente superior a 2 DPs (ver Figura 6.11).

Obtemos um escore *z* através do cálculo do escore-diferença ($x = X - \bar{X}$)–que dá a distância de um X qualquer até a média–e, então, pela divisão dessa diferença por σ.

Em símbolos:

$$z = \frac{X - \bar{X}}{\sigma} \quad \text{ou} \quad \frac{x}{\sigma}, \text{ onde}$$

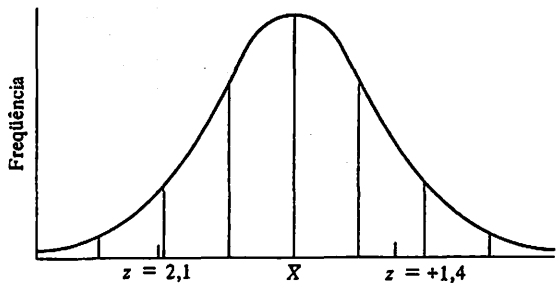


FIGURA 6.11 Posição de $z = -2,1$ e de $z = +1,4$ Numa Distribuição Normal

* N.T.: Essa distância entre o escore bruto e a média é a própria *discrepância.*

x = escore-diferença ou escore-discrepância
σ = desvio padrão da distribuição
z = escore padronizado

Exemplo 1
Vamos estudar a distribuição da renda anual de uma cidade norte-americana. A renda anual média é de US$5.000 e o desvio padrão, US$1.500. Admitindo-se que a distribuição da renda anual tenha distribuição normal, podemos traduzir um escore bruto de, por exemplo, US$7.000 em escore padrão do seguinte modo:

$$z = \frac{7.000 - 5.000}{1.500} = +1,33$$

Vemos, assim, que uma renda anual de US$7.000 corresponde a um 1,33 desvios padrões acima da média anual de US$5.000 (ver Figura 6.12).

Exemplo 2
Ao trabalhar com uma distribuição normal de escores representativos do grau de contentamento manifestado por um grupo de inquilinos beneficiários do serviço público de habitação, vamos supor que $\bar{X} = 10$ e DP = 2. (Por convenção, quanto maior o escore, maior a satisfação advinda do serviço público.) Para determinarmos quantos desvios padrões um escore igual a 3 dista da média ($\bar{X} = 10$), a primeira coisa que devemos fazer é obter a diferença entre esse escore e $\bar{X}$. Assim:

$$z = X - \bar{X} = 3 - 10 = -7$$

Dividindo, então, o resultado pelo desvio padrão, vem:

$$z = \frac{x}{\sigma} = -\frac{7}{2} = -3,5$$

Portanto, como bem demonstra a Figura 6.13, um escore bruto de 3, nessa distribuição, cai a 3,5 desvios padrões *abaixo* (à esquerda) da média.

Nota: conhecendo-se um escore *z* (padronizado), é sempre possível obter-se o escore bruto equivalente mediante o emprego da seguinte fórmula:

$$X = z\sigma + \bar{X}$$

Entrando com os dados do presente exemplo nessa fórmula, resulta:

$$X = (-3,5)(2) + 10 = -7 + 10 = 3$$

PROBABILIDADE E CURVA NORMAL

Como veremos a seguir, a curva normal pode ser usada em conjunção com os escores *z*

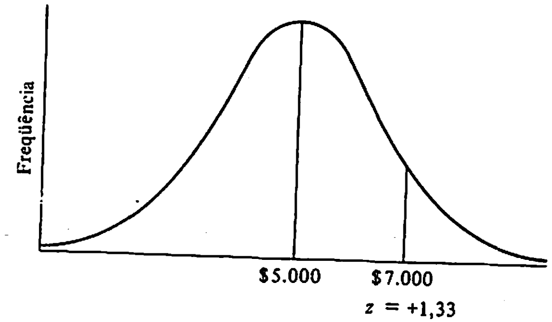


FIGURA 6.12 Posição de $z = 1,33$ (na Curva Normal), Relativa ao Escore Bruto US$7.000

(escores padronizados) e a Tabela B no cálculo da probabilidade associada a qualquer dado de uma distribuição. No contexto presente, o termo *probabilidade* refere-se à freqüência relativa de ocorrência de um dado ou evento qualquer; em outras palavras, *a probabilidade associada a um evento qualquer é o número de vezes que tal evento pode ocorrer em relação ao número total de eventos.* Redispondo os termos dessa definição:

$$\text{Probabilidade de um dado ou de um evento qualquer} = \frac{\text{Número de vezes que o dado ou evento pode ocorrer}}{\text{Número total de dados ou eventos}}$$

Assim, a probabilidade de extrair uma única carta (digamos, um "ás de espadas"), de um maço embaralhado de 52 cartas, é 1 em 52, uma vez que o evento "ás de espadas" só pode ocorrer uma vez no total (pois entre as 52 cartas só existe *um* "ás de espadas"). A probabilidade de obter-se "cara" com uma moeda honesta* (ou perfeitamente balanceada), numa só jogada é 1 em 2, já que "cara" só ocorre uma vez no conjunto total de possibilidades (2). Por igual raciocínio, se nos mandassem abrir um livro de 100 páginas numa página qualquer (por exemplo, pág. 23),** a probabilidade de conseguir, numa só tentativa e aleatoriamente, abrir o livro *naquela* página seria de 1 em 100.

A curva normal é uma distribuição na qual é possível determinar probabilidades associadas a *todos* os pontos da linha de base. Como já foi mencionado anteriormente, a curva normal é uma *distribuição de freqüências*; a freqüência total sob a curva é igual a 100%; essa curva apresenta uma área central que circun-

* N.T.: Além do termo "honesto"(a), muito comum em Estatística, ocorrem, ainda: *eqüiprovável, balanceado(a), não viesado(a).*

**N.T.: Esse raciocínio é falho! Se a idéia for abrir o livro na *página 23*, o mais "ingênuo" dos pesquisadores procurará abri-lo na "primeira metade" e, aí, o 100 fica sem sentido. Melhor seria suprimir a sugestão quanto à página, deixando apenas *qualquer.*

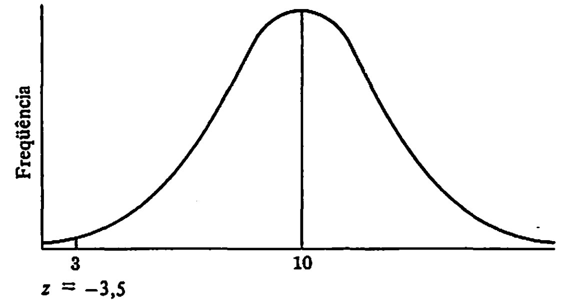


FIGURA 6.13 Posição de $z = -3,5$ Relativa ao Escore Bruto = 3

da a média, onde se localizam os escores mais freqüentes, e há, ainda, áreas menores progressivamente mais próximas de ambas as extremidades (caudas), onde encontramos, em pequenas proporções, escores muito altos ou muito baixos. Então, em termos probabilísticos, podemos dizer que a probabilidade decresce à medida que, na linha de base, nos afastamos da média em ambos os sentidos. Desse modo, dizer que 68,26% da freqüência total sob a curva normal caem entre -1σ e $+1\sigma$, a partir da média, é o mesmo que dizer que a probabilidade é de cerca de 68 em 100 de que um escore bruto qualquer caia dentro desse intervalo ($-1\sigma \longmapsto +1\sigma$). De forma análoga, dizer que 95,44% da freqüência total sob a curva normal caem entre -2σ e $+2\sigma$, a contar da média, é o mesmo que dizer que a probabilidade é de aproximadamente 95 em 100 de que um escore bruto qualquer venha a situar-se dentro desse intervalo ($-2\sigma \longmapsto +2\sigma$) e assim por diante.

O que acabamos de fazer foi repetir precisamente o mesmo conceito de probabilidade ou *freqüência relativa* que já estudamos quando extraímos uma carta de um baralho completo, quando jogamos uma moeda ou quando abrimos de forma aleatória um livro numa página qualquer. Note-se, contudo, que as probabilidades associadas às áreas sob a curva normal são sempre expressas com relação a 100%—que correspondem à área *total* sob a curva (por exemplo, 68 em 100; 95 em 100; 99 em 100 e assim por diante). Por essa razão e para garantir que em todo o resto deste livro o termo seja entendido de modo padronizado, vamos sempre falar em *probabilidade como sendo o número de vezes que um dado evento pode ocorrer num conjunto de 100 repetições.* Assim, a probabilidade de sortear um ás de espadas de um baralho bem misturado é de 1,92 em 100 ($\frac{1}{52}$); a probabilidade de tirar "cara" no lance de uma moeda é de 50 em 100 ($\frac{1}{2}$). Note-se, além disso, que é usual expressar probabilidades sob a forma decimal (P). Por exemplo, podemos dizer que $P = 0,50$—isto é, ($\frac{50}{100}$)—de tirar "cara" num só lance; analogamente, podemos dizer que $P = 0,68$—ou seja, ($\frac{68}{100}$)—de que um escore bruto caia entre -1σ e $+1\sigma$ sob a curva normal.

Expressa sob a forma de razão (quociente), *a probabilidade será sempre um número que oscila entre 0 e 1.* A probabilidade de ocorrência de um evento é 0 quando estamos absolutamente seguros de que ele *não ocorrerá*; é 1 quando estamos convencidos de que sem dúvida nenhuma ele *ocorrerá.* O problema é que os pesquisadores nunca estão totalmente seguros a respeito de coisa alguma! Em conseqüência, podemos, via de regra, esperar encontrar probabilidades iguais a 0,60, 0,25 ou 0,05; mas raras vezes é possível esperar reduzir a probabilidade a 0 ou, por outro lado, elevá-la a 1.

Uma característica importante das probabilidades reside na *regra da soma*, que estabelece o seguinte: *a probabilidade de ocorrência de um evento qualquer, num conjunto de alternativas, é igual à soma das probabilidades separadas.* Suponham, por exemplo, que desejemos encontrar a probabilidade de extrair, dum maço bem embaralhado de 52 cartas, *ou* o "ás de espadas", *ou* a "rainha de ouros" *ou* o "rei de copas" num único lance. Pela soma das probabilidades isoladas–($\frac{1}{52}+\frac{1}{52}+\frac{1}{52}$)–verificamos que a probabilidade de obtenção de qualquer dessas cartas, numa única tentativa é igual a $\frac{3}{52}$, ou seja, $P = 0{,}06$. Em outras palavras, temos 6 chances em 100 de obter, numa única tentativa, ou um "ás de espadas", *ou* uma "dama de ouros" *ou* um "rei de copas" (ver Figura 6.14).

A regra da soma sempre pressupõe que os eventos sejam *mutuamente excludentes*, em outras palavras, que dois eventos não podem jamais ocorrer em simultaneidade. Por exemplo, num baralho de 52 cartas, não é possível que, ao mesmo tempo, uma carta seja de "espadas", de "ouros" *e* de "copas". Analogamente, uma moeda, jogada uma só vez, não pode produzir em conjunto "cara" *e* "coroa".

Admitida a mútua exclusão de eventos, podemos dizer que as probabilidades associadas a cada uma das possíveis ocorrências de um evento sempre têm 1 por soma. Essa afirmação estabelece que *algo tem que ocorrer.* Se não sair "cara", sai "coroa"; se não for um "ás", será um "rei", "rainha", "valete", "dez" e por aí afora. No lançamento de uma moeda, a probabilidade de sair "cara" é de $\frac{1}{2}$ (isto é, $P = 0{,}50$). Naturalmente, a probabilidade de sair "coroa" é também de $\frac{1}{2}$ (ou seja, $P = 0{,}50$). Então, se somarmos as probabilidades relativas a todas as ocorrências, verificaremos que a probabilidade de sair "cara" *ou* "coroa" é 1 (isto é, $\frac{1}{2}+\frac{1}{2} = 1$).

Outra propriedade importante das probabilidades é a *regra da multiplicação*: aí, o problema consiste em determinar a probabilidade de realização de dois ou mais eventos *em sucessão*, um após o outro. A regra da multiplicação estabelece o seguinte:

Probabilidade de extrair um "ás de espadas" = $\frac{1}{52}$

Probabilidade de extrair uma "dama de ouros" = $\frac{1}{52}$ +

Probabilidade de extrair um "rei de copas" = $\frac{1}{52}$

Probabilidade de extrair *ou* um "ás de espadas", *ou* uma "dama de ouros" *ou* um rei de copas" — $\frac{3}{52}$; isto é, $P = 0{,}06$

FIGURA 6.14 Exemplo Ilustrativo da Regra da Soma: Trata-se, Aqui, de Extrair num Só Lance, *ou* um "Ás de Espadas", *ou* uma "Dama de Ouros" *ou* um "Rei de Copas"

Probabilidade de sair "cara" no primeiro lance → $\frac{1}{2}$

Probabilidade de sair "cara" no segundo lance → $\frac{1}{2}$ X

Probabilidade de saírem "caras" nas duas jogadas sucessivas → $\frac{1}{4}$ (isto é, $P = 0{,}25$)

FIGURA 6.15 Ilustração da Regra da Multiplicação: Cálculo da Probabilidade de Saírem "Caras" em 2 Jogadas Sucessivas da Mesma Moeda

a probabilidade de obtenção de eventos mutuamente excludentes em série é igual ao produto das probabilidades separadas. Em lugar de "ou ... ou", a regra da multiplicação sugere "primeiro, segundo, terceiro, ..."

Por exemplo, qual a probabilidade de obter "cara" em duas jogadas sucessivas de uma moeda (isto é, "cara" tanto no primeiro lance quanto no segundo)? Visto que esses eventos são independentes, o evento relacionado com o primeiro lance não influencia o segundo. Na primeira jogada da moeda, a probabilidade de sair "cara" é igual a $\frac{1}{2}$ (ou seja, $P = 0{,}50$); na segunda jogada, a probabilidade desse mesmo evento ainda é igual a $\frac{1}{2}$ (isto é, $P = 0{,}50$). Portanto, a probabilidade de conseguirmos "cara" em dois lances sucessivos de uma moeda é igual a $(\frac{1}{2})(\frac{1}{2}) = \frac{1}{4}$ (ou $P = 0{,}25$) (ver figura 6.15).

Para aplicarmos o conceito precedente de probabilidade com referência à curva normal, vamos voltar a um exemplo anterior. Recorde-se o leitor de que naquele problema relacionado com a distribuição de renda anual de uma cidade (norte-americana), um particular escore bruto teve de ser transformado no seu eqüivalente z. Tal distribuição de rendas tinha média de US$5.000 e desvio padrão de US$1.500.

Ao aplicarmos a fórmula da "escala z", verificamos, anteriormente, que uma renda anual de US$7.000 situava-se a 1,33 DPs acima da média (US$5.000); ou seja,

$$z = \frac{7.000 - 5.000}{1.500} = +1{,}33$$

Vamos agora determinar a probabilidade associada a um escore situado entre US$5.000–a própria média–e US$7.000. Em outras palavras, qual a probabilidade de escolher-se de forma aleatória, numa só tentativa, uma pessoa dessa cidade cuja renda anual caia entre US$5.000 e US$7.000? O problema está ilustrado graficamente na Figura 6.16 (observe-se a área sombreada sob a curva) e pode ser resolvido (em matemática) pela aplicação da fórmula dos escores z, combinada com a Tabela B (fim do livro).

PASSO 1: Transformar o escore bruto (US$7.000) em escore z*

* N.T.: Além de "escore z", é comum falar-se em "escore reduzido", "(variável) normal reduzida" ou, simplesmente, valor da "escala z".

em torno do qual se equilibram as discrepâncias positivas e negativas. A fim de que seja possível perceber essa característica da média, devemos primeiro entender o conceito de *discrepância*, que nada mais é senão a *distância da média a qualquer escore bruto*. Para calcular uma determinada discrepância, basta subtrair a média aritmética de qualquer escore bruto dado. Em símbolos:

$$x = X - \bar{X}\text{, onde}$$

x = escore-discrepância ou simplesmente discrepância (sempre simbolizada por x, isto é, "xis minúsculo")
X = qualquer escore bruto da distribuição
$\bar{X}$ = média aritmética.

TABELA 4.4 **Discrepâncias de um Conjunto de Escores Brutos com Relação a $\bar{X}$**

X	x		
9	+3	+5	
8	+2		
6	0		$\bar{X} = 6$
4	−2	−5	
3	−3		

Uma vez que $\bar{X} = 6$ para o conjunto de escores brutos 9, 8, 6, 4 e 3, o valor $X = 9$ fica exatamente três unidades *acima* da média 6 (ou $X - \bar{X} = 9 - 6 = +3$). De modo análogo, o escore bruto 4 fica duas unidades *abaixo* da média (ou $X - \bar{X} = 4 - 6 = -2$). Conclusão: quanto maior a discrepância x, tanto maior a distância entre um determinado escore bruto e a média da distribuição.

Considerando-se a média como o ponto de equilíbrio da distribuição, pode-se dizer, agora, que a soma das discrepâncias situadas acima da média é igual, em valor absoluto (descartados os sinais "menos"), à soma das discrepâncias que ficam abaixo dela. Vamos voltar a um exemplo anterior–o conjunto de escores 9, 8, 6, 4, 3, onde $\bar{X} = 6$. Se a média dessa distribuição for seu "centro de gravidade", então, desprezando os sinais "menos" e somando as discrepâncias positivas (discrepâncias relativas aos escores 8 e 9), o resultado deve ser igual ao obtido mediante a soma das discrepâncias negativas (isto é, relativas aos escores 4 e 3). Como evidencia a Tabela 4.4, é precisamente esse caso–já que a soma das discrepâncias abaixo de $\bar{X}$ (−5) é igual à soma das discrepâncias acima de $\bar{X}$ (+5).

Vamos a mais um exemplo: 4 é a média de 1,2, 3, 4, 5, 6 e 7. Observe-se que a soma das discrepâncias abaixo desse escore é −6, enquanto que a soma das discrepâncias acima dele é +6. Voltaremos a examinar o conceito de discrepância nos Capítulos 5 e 6.

Cálculo da Média de uma Distribuição de Freqüências (com Dados Isolados)

Para obter-se a média de um pequeno número de escores, a fórmula básica $\bar{X} = \Sigma X/N$ serve perfeitamente. Entretanto, quando estamos diante de um número maior de dados, pode ser mais prático e menos demorado calcular a média de uma distribuição de freqüências (dados isolados) pela fórmula

$$\bar{X} = \frac{\Sigma fX}{N}\text{, onde}$$

$\bar{X}$ = média
X = um escore bruto qualquer da distribuição (isto é, a própria variável)
fX = um escore qualquer multiplicado pela respectiva freqüência de ocorrência
ΣfX = soma dos fX's
N = total de escores.

A Tabela 4.5 ilustra o cálculo da média de um conjunto de dados isolados (porém afetados de freqüências).

TABELA 4.5 **Cálculo de $\bar{X}$ de uma Distribuição de Freqüências Simples (Dados Isolados)**

X	f	fX	
8	2	16	
7	3	21	
6	5	30	
5	6	30	$\bar{X} = \frac{\Sigma fX}{N} = \frac{132}{28} = 4{,}71$
4	4	16	
3	4	12	
2	3	6	
1	1	1	
	$N = 28$	$\Sigma fX = 132$	

COMPARAÇÃO ENTRE MODA, MEDIANA E MÉDIA

Há um momento em que o pesquisador procura uma medida de tendência central para a sua situação particular de pesquisa. Empregará ele a moda, a mediana ou a média? Sua decisão envolve vários fatores tais como:

1. o nível de mensuração;
2. o aspecto ou forma de sua distribuição de dados e
3. o objetivo da pesquisa.

Nível de Mensuração

Uma vez que a moda requer apenas o conhecimento das freqüências, ela pode ser calculada para qualquer conjunto de dados, sejam eles do nível nominal, ordinal ou intervalar. Por exemplo, poderíamos estabelecer que a categoria modal do nível nominal da variável "filiação religiosa" (protestante, católico ou judeu) é protestante, se a maioria dos nossos respondentes se identificassem como tais. Similarmente, poderíamos dizer que a maioria dos estudantes freqüentadores de uma dada universidade apresenta uma nota média 2,5 (Mo = 2,5).

A mediana exige uma ordenação de categorias, da mais alta à mais baixa. Por essa razão, ela só pode ser obtida a partir de dados ordinais ou intervalares, mas *nunca* de dados nominais. Ilustrando: poderíamos descobrir que a renda anual mediana de dentistas de uma pequena cidade é da ordem de Cr$ 289.000,00. Esse resultado fornece-nos um critério sólido para o exame da tendência central de nossos dados. Contrariamente, faria pouco sentido calcular a mediana de escalas de filiação religiosa (protestante, católico ou judeu), sexo (masculino ou feminino), ou país de origem (Inglaterra, Polônia, França ou Alemanha), onde não foi feita uma atribuição de postos (escalonamento).

O uso da média está restrito exclusivamente a dados intervalares. Calcular a média de dados ordinais ou nominais dá origem a um resultado sem sentido, que em geral não indica a tendência central. Que sentido faria o cálculo da média de uma distribuição onde a variável fosse filiação religiosa ou sexo? Embora seja menos óbvio, é igualmente inapropriado calcular a média de dados que podem ser *graduados* porém não *medidos*.

Forma da Distribuição

O aspecto ou forma de uma distribuição é outro fator que pode influenciar o pesquisador na escolha de uma medida de tendência central. Numa distribuição unimodal e perfeitamente simétrica, a moda, a mediana e a média serão idênticas, uma vez que o ponto de freqüência máxima (Mo) é também o valor "mais" central (Md) e o "centro de gravidade" dos dados ($\bar{X}$). Como demonstra a Figura 4.2, as medidas de tendência central coincidirão no ponto central, exatamente no "pico" da distribuição simétrica.

Quando o pesquisador trabalha com uma distribuição simétrica, a escolha de uma medida de tendência central baseia-se principalmente nos objetivos particulares de sua investigação e no nível em que seus dados foram colhidos. Entretanto, quando ele trabalha com uma distribuição assimétrica, sua decisão é muito influenciada pelo formato dos dados de que dispõe.

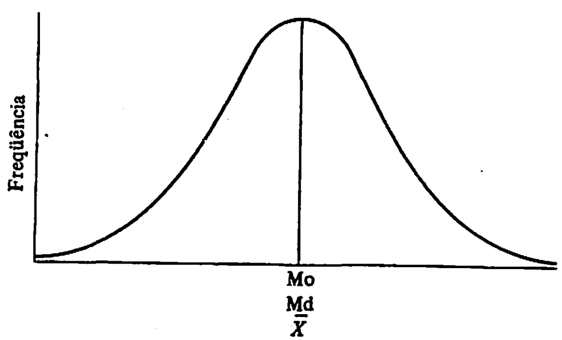


FIGURA 4.2 **Distribuição Unimodal, Simétrica, com a Localização da Moda, da Mediana e da Média (onde Mo = Md = $\bar{X}$)**

A Figura 4.3 mostra que a moda, a mediana e a média não coincidem em distribuições assimétricas, *muito embora suas posições relativas permaneçam constantes;* dito de outra forma, no sentido "pico"/"cauda", a ordem é sempre da moda para a média, com a mediana entre esses dois pontos. A moda é a que cai mais perto do "pico" da curva, já que ela corresponde ao ponto da escala cujo valor é mais freqüente. A média, ao contrário, é a que está mais próxima da "cauda", onde se situam os escores extremos (e poucos). Por essa razão, o escore médio numa distribuição positivamente assimétrica (Figura 4.3(a)) localiza-se entre os escores mais altos; a média numa distribuição negativamente assimétrica (Figura 4.3(b)) cai perto dos escores mais baixos.

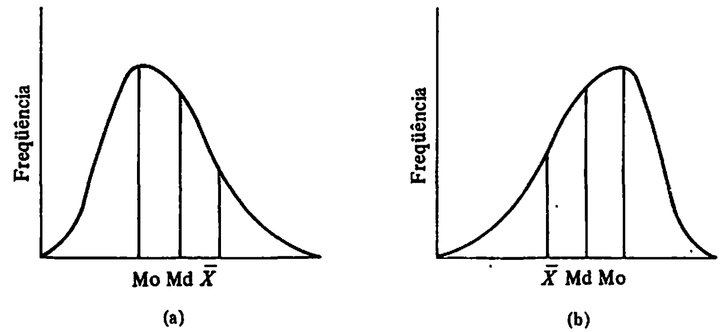


FIGURA 4.3 **Posições Relativas de Medidas de Tendência Central: (a)–Distribuição Positivamente Assimétrica e (b)–Distribuição Negativamente Assimétrica**

Enquanto que a média é muitíssimo influenciada pelos escores extremos (em ambos os sentidos), a mediana sofre pouca ou nenhuma influência de alterações nos valores extremos. Isto porque a média considera todos os escores de qualquer distribuição, enquanto que a mediana (por definição) só se preocupa com o valor numérico do escore que ocupa a posição "mais" central. Como ilustrado abaixo, a mudança do valor de um escore extremo de 10 (distribuição a) para 95 (distribuição b), não modifica de maneira nenhuma o valor da mediana (Md = 7,5), enquanto que a média pula de 7,63 para 18,25.

distribuição A:	5 6 6 7 8 9 10 10	Md = 7,5	$\bar{X}$ = 7,63
distribuição B:	5 6 6 7 8 9 10 95	Md = 7,5	$\bar{X}$ = 18,25

Numa distribuição assimétrica, a mediana sempre cai em algum lugar entre a média e a moda. É essa característica que a torna a medida de tendência central mais desejável para descrever uma distribuição assimétrica. Para ilustrar essa vantagem da mediana, voltemo-nos para a Tabela 4.6 e examinemos o salário anual "médio" dos empregados que trabalham numa pequena corporação. Se trabalhássemos no departamento de relações públicas dessa corporação e desejássemos angariar-lhe uma imagem pública favorável, calcularíamos, talvez, a média, a fim de demonstrar que o empregado "típico" ganha Cr$ 306.000,00 (por ano), e é, portanto, relativamente bem pago. Por outro lado, se nós fôssemos representantes sindicais e estivéssemos procurando melhorar os níveis salariais, iríamos provavelmente empregar a moda para demonstrar que o salário anual "médio" é de apenas Cr$ 17.000,00, o que representa uma quantia bastante reduzida. Para finalizar, se fôssemos pesquisadores sociais desejosos de dar uma informação acurada do salário "médio" dos empregados dessa corporação, empregaríamos, inteligentemente, a mediana (Cr$ 51.000,00), uma vez que esse valor cai entre as outras medidas de tendência central e, portanto, oferece um quadro mais equilibrado da estrutura salarial. O método mais aceitável seria indicar as três medidas de tendência central e permitir que os interessados interpretassem os resultados. Infelizmente é verdade que poucos pesquisadores sociais—não contando os relações públicas e os líderes sindicais—indicam mais do que uma única medida de tendência central.

TABELA 4.6 Medidas de Tendência Central numa Distribuição Assimétrica de Salários Anuais

Salário Medido em Cruzeiros (Cr$)	
1.700.000,00	
425.000,00	
170.000,00	$\bar{X}$ = Cr$ 306.000,00
85.000,00	
17.000,00	Md = Cr$ 51.000,00
17.000,00	
17.000,00	Mo = Cr$ 17.000,00

Ainda mais grave é o fato de que alguns relatórios de pesquisa não especificam de forma exata a medida de tendência central—moda, mediana ou média—que foi usada para calcular a quantia ou posição "média" de um grupo de escores. Como demonstrou a ilustração anterior, uma interpretação razoável de certos dados pode tornar-se literalmente impossível sem tal informação.

Salientou-se páginas atrás que algumas distribuições de freqüências podem ser caracterizadas como bimodais, uma vez que elas contêm dois pontos de freqüência máxima. Para descrever adequadamente distribuições bimodais, é, em geral, útil identificar *ambas* as modas; aspectos importantes de tais distribuições podem ser encobertos se a mediana ou a média forem usadas.

Considerem a situação em que um pesquisador realizou entrevistas pessoais com 26 respondentes de renda baixa, a fim de estabelecer seus conceitos ideais de tamanho de família. A cada um deles foi dito o seguinte: "Suponha que você pudesse decidir exatamente o tamanho de sua família. Incluindo todas as crianças e os adultos, quantas pessoas você gostaria que ela tivesse?" Como demonstra a Tabela 4.7, os resultados desse estudo indicaram amplas preferências por tamanho de família, preferências que variaram desde *viver sozinho* (1) até *viver com muitas pessoas* (10). Usando ou a média ou a mediana, poderíamos concluir que a família ideal de um respondente "médio" consistiria de seis pessoas ($\bar{X}$ = 5,58; Md = 6). Entretanto, sabendo que a distribuição é bimodal, percebemos que, de fato, houve *duas* concepções ideais de tamanho de família

nesse grupo de respondentes: uma com um número bastante grande de pessoas (Mo = 8) e outra com apenas algumas pessoas (Mo = 3).

TABELA 4.7 Concepções Ideais de Tamanho de Família numa Amostra de 26 Respondentes de Renda Baixa. (Distribuição Bimodal)

Tamanho Ideal de Família	*f*
10	1
9	2
8	6
7	3
6	2
5	1
4	2
3	6
2	2
1	1
	N = 26

Objetivo da Pesquisa

Até aqui, discutimos a escolha de uma medida de tendência central em termos do nível de mensuração e da forma de uma distribuição de escores. Perguntamos agora: que o pesquisador espera fazer com a medida de tendência central escolhida? Se ele procura uma medida descritiva que seja rápida, simples, ainda que grosseira, ou se ele estiver trabalhando com uma distribuição bimodal, ele geralmente empregará a moda. Na maioria das situações com que um pesquisador se defronta, entretanto, a moda só é útil como um indicador preliminar de tendência central, que pode ser obtida com rapidez a partir de um exame cuidadoso dos dados. Se ele procura uma medida de tendência central que seja exata,* a decisão fica geralmente entre a mediana e a média.

* N.T.: Muito embora o uso popular tenha tornado sinônimas as palavras *precisão* e *exatidão*, do ponto de vista formal estamos diante de coisas diferentes. Um instrumento é *preciso* quando, em sucessivas mensurações da mesma magnitude, ele apresenta um *erro constante*. Um instrumento preciso caracteriza-se pela *precisão*.

Um instrumento é *exato* quando, em sucessivas mensurações da mesma magnitude, o *erro* é *variável* (para mais ou para menos), porém a média aritmética de todas as mensurações representa o valor mais provável da dita magnitude. Por isso, a média aritmética representa o valor *exato* buscado. O instrumento que produz mensurações exatas caracteriza-se pela *exatidão*.

Exemplificando: Se um objeto pesa *realmente* 2 kg, pesado com uma balança exata repetidas vezes, as mensurações resultantes poderão ser: 2,1 kg; 2,2 kg; 1,9 kg; 1,8 kg; etc., onde a média aritmética produz o valor real buscado. Assim:

$$\frac{2{,}1\text{ kg} + 2{,}2\text{ kg} + 1{,}9\text{ g} + 1{,}8\text{ kg}}{4} = 2\text{ kg}$$

Seja agora outra balança com a seguinte característica: em cada quilo ela erra, para mais, 10 gramas. Então, esse objeto, cujo peso real é 2 kg, pesará *sempre* 2 kg + 20 g. Se a pesagem for repetida *n* vezes ($n > 1$), o erro será consistente e constante: 20 g.

Na descrição de uma distribuição assimétrica, o pesquisador normalmente escolhe a mediana (como, aliás, já foi mencionado em outro local), uma vez que ela tende a dar um quadro balanceado* dos escores extremos. Além disso, a mediana é algumas vezes usada como um ponto da distribuição que divide os escores em duas categorias–*ambas* com o *mesmo número* de sujeitos. Por exemplo, poderíamos dividir os respondentes (Tabela 4.7) em dois grupos: aqueles que preferem uma família grande (valores superiores à mediana) e os que preferem família pequena (valores menores que a mediana).

Sempre que se deseja uma medida exata de distribuições simétricas, a média tem preferência sobre a mediana pelo fato de poder ser facilmente usada em análises estatísticas mais avançadas, tais como as que aparecem em capítulos subseqüentes deste livro. Acrescente-se a isso o fato de ser a média mais estável que a mediana, ou seja, se de uma dada população forem extraídas diferentes amostras, haverá menor variação nas médias do que nas medianas. Essa vantagem da média–muito embora talvez ainda não compreendida ou avaliada pelo leitor–tornar-se-á mais evidente em estudos posteriores do tópico que trata do problema das *decisões em Estatística* (ver Capítulo 7).

CÁLCULO DA MODA, DA MEDIANA E DA MÉDIA NUMA DISTRIBUIÇÃO DE FREQÜÊNCIAS COM DADOS AGRUPADOS

Numa distribuição de freqüências com dados agrupados, a moda é o ponto médio do intervalo de classe que apresenta a freqüência mais alta. De acordo com essa definição, a moda da distribuição constante da Tabela 4.8 é 72, uma vez que esse valor corresponde ao ponto médio do intervalo que aparece mais vezes (ele ocorre 17 vezes)**

TABELA 4.8 Localização da Moda numa Distribuição de Freqüências com Dados Agrupados

Intervalo de Classe	*Ponto Médio*	*f*
95-99	97	3
90-94	92	2
85-89	87	4
80-84	82	7
75-79	77	12
70-74	72	17
65-69	67	12
60-64	62	5
55-59	57	5
50-54	52	4
		$N = 71$

*N.T.: Balanceado significa aqui "que não sofre a influência de valores extremos".

**N.T.: Rigorosamente, o ponto médio da 6ª classe (de cima para baixo) é 71,5, se a variável for contínua. Para que esse ponto médio seja 72 é preciso que, nas condições da tabela, sejam desprezados os limites do intervalo, ou seja, 70 e 74. Isso porque a classe foi indicada pela notação 70–74, que é a maneira tradicional de indicar um intervalo aberto (dos dois lados).

Para localizar a mediana de dados agrupados numa distribuição de freqüências, devemos (1) encontrar o intervalo de classe que contém a mediana e (2) interpolar.

Passo 1–Para localizar o intervalo mediano, construímos primeiro uma distribuição de freqüências acumuladas, tal como figura na terceira coluna da Tabela 4.9. A partir do intervalo que contém os valores mais baixos (as idades menores–20/29), vamos somando as freqüências até chegar àquele intervalo que contém o ponto (o valor) que divide a distribuição em duas partes iguais, isto é, o ponto central. No presente exemplo, $N = 100$. Portanto, devemos procurar o qüinquagésimo valor, que é: $N/_2 = 100/2 = 50$. Consultando a coluna das freqüências acumuladas (de baixo para cima), observamos que 26 sujeitos têm idades de 39 anos ou menos. Vemos também que o qüinquagésimo sujeito situa-se no intervalo 40/49, já que esse é o intervalo de classe cujas freqüências acumuladas contém 53–portanto, mais do que a metade dos sujeitos. Em outras palavras: com relação às freqüências acumuladas, do vigésimo sétimo ao qüinquagésimo terceiro sujeitos encontramo-nos no intervalo 40/49. Esse é o intervalo mediano.*

TABELA 4.9 Distribuição de Freqüências de Dados Agrupados (idades)

Intervalo	*f*	*fa*
60-69	15	100
50-59	32	85
40-49	27	53
30-39	16	26
20-29	10	10
	$N = 100$	

Passo 2–Para encontrar o valor exato da mediana, aplicamos a fórmula

$$\text{Mediana} = \begin{pmatrix}\text{Limite inferior real}\\ \text{do intervalo}\\ \text{mediano}\end{pmatrix} + \left(\frac{\dfrac{N}{2} - \begin{matrix}\text{Freqüência acumulada}\\ \text{corrrespondente ao}\\ \text{intervalo anterior ao}\\ \text{que contém a mediana}\end{matrix}}{\begin{matrix}\text{Freqüência absoluta do}\\ \text{intervalo mediano}\end{matrix}} \right) \times$$

$$\times \begin{pmatrix}\text{Amplitude}\\ \text{do intervalo}\\ \text{mediano}\end{pmatrix}$$

* N.T.: Melhor seria dizer apenas o seguinte: o intervalo mediano é o intervalo que corresponde, na coluna de freqüências acumuladas, ao valor que contém (pela primeira vez) metade do total de sujeitos da distribuição. Sendo fa_i uma freqüência acumulada de ordem i, o intervalo mediano será obtido por correspondência com a relação $fa_i \geqslant N/2$. O intervalo mediano é também chamado de *lugar mediano*.

Para os dados da Tabela 4.9*, a mediana é calculada como segue:

$$\text{Mediana} = 39{,}5 + \left(\frac{50-26}{27}\right)10 = 39{,}5 + 8{,}89 = 48{,}39 \text{ anos.}$$

Para calcular a média de uma distribuição de freqüências com dados agrupados, podemos usar uma versão modificada da fórmula vista anteriormente (ver Tabela 4.5). Conforme exemplificado abaixo, o símbolo X não será mais usado para designar um escore, mas, sim, o *ponto médio de um intervalo de classe*. Portanto,

$$\bar{X} = \frac{\Sigma fX}{N}, \text{ onde}$$

$\bar{X}$ = média
X = ponto médio de um intervalo de classe
fX = ponto médio de um intervalo de classe multiplicado pela respectiva freqüência absoluta
N = número total de escores.

Podemos ilustrar o cálculo da média de dados agrupados usando como referência a seguinte distribuição.**

Intervalo	*f*
17 ⊢20	1
14 ⊢17	2
11 ⊢14	3
8 ⊢11	5
5 ⊢8	4
2 ⊢5	2
	$N = 17$

PASSO 1: Calcular o ponto médio de cada intervalo de classe.

Intervalo	*X = Ponto Médio*
17 ⊢20	18
14 ⊢17	15
11 ⊢14	12
8 ⊢11	9
5 ⊢8	6
2 ⊢5	3

*N.T.: A variável "idade" é uma variável *contínua*. Por isso, qualquer *limite inferior real* é obtido a partir do *limite inferior aparente menos 0,5*. Para ajudar a "enxergar" este fato, o ideal seria adotar a seguinte notação nos intervalos de classe: 40 ⊢50, 50 ⊢60, etc. Assim fica claro que os limites inferiores reais são, respectivamente, 39,5 e 49,5, que os superiores são, na mesma ordem, 49,4 e 59,4.

**N.T.: Veja-se a propósito, o que ficou dito na nota do tradutor, página 24, Cap. 2, sobre a notação de intervalo.

PASSO 2: Multiplicar cada ponto médio pela respectiva freqüência absoluta, a fim de obter ΣfX.

Intervalo	*X = Ponto Médio*	*f*	*fX*
17 –20	18	1	18
14 –17	15	2	30
11 –14	12	3	36
8 –11	9	5	45
5 –8	6	4	24
2 –5	3	2	6
		$N = 17$	$\Sigma fX = 159$

PASSO 3: "Entrar" com o resultado do Passo 2 na fórmula de $\bar{X}$.

$$\bar{X} = \frac{\Sigma fX}{N} = \frac{159}{17} = 9{,}35$$

RESUMO

Este capítulo introduziu as três medidas de tendência central mais conhecidas, vale dizer, medidas do que é "médio" ou "típico" num conjunto de dados. A moda foi definida como sendo a categoria ou escore que ocorre com maior freqüência; a mediana, como o ponto que divide a distribuição ao meio; a média, como a soma de todos os valores dividida pelo número de parcelas (isto é, de valores do conjunto). Essas medidas de tendência central foram comparadas relativamente a: nível de mensuração (da variável), forma da distribuição (dos dados) e objetivo da pesquisa. Podemos resumir as condições que regem a escolha de uma dessas medidas de tendência central da seguinte maneira:

Moda:

1. nível de mensuração: nominal, ordinal ou intervalar;
2. forma da distribuição: mais apropriada para distribuições bimodais;*
3. objetivo: permite obter uma medida de tendência central rápida, simples, embora grosseira.

Mediana:

1. nível de mensuração: ordinal ou intervalar;
2. forma da distribuição: mais adequada para distribuições muito assimétricas;

* N.T.: Melhor seria dizer *multimodais*, já que a "bimodalidade" está implícita na "multimodalidade". Observe-se, ainda, que nada impede que a *moda* seja a medida "preferível" mesmo numa distribuição *unimodal*. O que rege isso é o *objetivo* da pesquisa.

3. objetivo: é uma medida de tendência central "confiável"; pode, às vezes, ser usada em operações estatísticas mais avançadas ou para "quebrar" uma distribuição em duas categorias distintas (por exemplo: alto X baixo).

Média:

1. nível de mensuração: intervalar (no mínimo);
2. forma da distribuição: mais apropriada para distribuições unimodais simétricas;
3. objetivo: medida de tendência central exata; pode freqüentemente ser usada em operações estatísticas mais avançadas, tais como os testes para tomada de decisões que aparecerão mais adiante neste livro.

PROBLEMAS

1. Sete empregados horistas numa companhia de porte médio ganham Cr$ 153,00; Cr$ 136,00; Cr$ 153,00; Cr$ 68,00; Cr$ 17,00; Cr$ 102,00 e Cr$ 51,00. Calcule (a) o salário-hora modal; (b) o salário-hora mediano e (c) o salário-hora médio.
2. Suponha que a companhia do problema 1 tenha recrutado mais um empregado pelo salário-hora de Cr$ 17,00, resultando, assim, a seguinte distribuição: Cr$ 153,00; Cr$ 136,00; Cr$ 153,00; Cr$ 68,00; Cr$ 17,00; Cr$ 102,00; Cr$ 51,00; e Cr$ 17,00. Calcule (a) o salário-hora modal; (b) o salário-hora mediano; (c) o salário-hora médio.
3. Para os escores 205, 6, 5, 5, 5, 2, e 1 calcule (a) a moda; (b) a mediana e (c) a média. Que medida de tendência central *não* deveria ser usada para descrever (isto é, para analisar) esse conjunto de escores? Por quê?
4. Seis estudantes num seminário de sociologia foram testados por meio de um instrumento que produz mensurações de nível intervalar. O objetivo da pesquisa era mensurar suas atitudes com relação a um grupo minoritário. Suas respostas, numa escala de 1 a 10 (quanto mais alto o escore, mais favorável a atitude), foram: 5, 2, 6, 3, 1 e 1.
   Relativamente aos escores acima, calcule (a) a moda; (b) a mediana e (c) a média. No geral, até que ponto esses estudantes são favoráveis a esse grupo minoritário?
5. Dados os escores 10, 12, 14, 8, 6, 7, 10, 10, encontre (a) a moda; (b) a mediana e (c) a média.
6. Dados os escores 3, 3, 4, 3, 1, 6, 5, 6, 6, 4, encontre (a) a moda; (b) a mediana e (c) a média.
7. Dados os escores 8, 8, 7, 9, 10, 5, 6, 8, 8, encontre (a) a moda, (b) a mediana e (c) a média.
8. Dados os escores 5, 4, 6, 6, 1, 3, encontre (a) a moda, (b) a mediana e (c) a média.
9. Dados os escores 8, 6, 10, 12, 1, 3, 4, 4, encontre (a) a moda, (b) a mediana e (c) a média.
10. Dados os escores 12, 12, 1, 12, 5, 6, 7, encontre (a) a moda, (b) a mediana e (c) a média.
11. Qual é a discrepância de cada um dos seguintes escores com relação a $\bar{X} = 20{,}5$? (a) $X = 20{,}5$; (b) $X = 33{,}0$; (c) $X = 15{,}0$; (d) $X = 21{,}0$.
12. Qual é a discrepância de cada um dos seguintes escores com relação a $\bar{X} = 3{,}0$? (a) $X = 4{,}0$; (b) $X = 2{,}5$; (c) $X = 6{,}3$; (d) $X = 3{,}0$.

13. Qual é a discrepância de cada um dos seguintes escores com relação a $\bar{X} = 15$? (a) $X = 22{,}5$; (b) $X = 3$; (c) $X = 15$; (d) $X = 10{,}5$.
14. Os escores relativos a atitudes de 31 estudantes frente a um grupo minoritário, foram dispostos na seguinte distribuição de freqüências (quanto maior o escore mais favorável a atitude:

Escores Relativos e Atitudes	*f*
7	3
6	4
5	6
4	7
3	5
2	4
1	2
	$N = 31$

Calcule (a) a moda; (b) a mediana; (c) a média.

15. Trinta e uma crianças matriculadas numa terceira série do primeiro grau de uma zona urbana indicaram, numa pesquisa, o número de irmãos e/ou irmãs que viviam com cada um deles em casa. Os dados resultantes foram dispostos na seguinte tabela de distribuição de freqüências:

Número de Irmãos/irmãs	*f*
5	6
4	7
3	9
2	5
1	4
	$N = 31$

Para essa classe de 31 crianças, calcule (a) o número modal de irmãos/irmãs; (b) o número mediano de irmãos/irmãs; (c) a média de irmãos/irmãs.

16. Dada a seguinte distribuição de freqüências, calcule (a) a moda; (b) a mediana e (c) a média:

Escores	*f*
10	3
9	4
8	6
7	8
6	9
5	7
4	5
3	2
2	1
1	1
	$N = 46$

17. Dada a seguinte distribuição de freqüências, calcule (a) a moda; (b) a mediana e (c) a média:

Intervalo de Classe	*f*
20-24	2
15-19	4
10-14	8
5-9	5
	$N = 19$

18. Dada a seguinte distribuição de freqüências, calcule (a) a moda; (b) a mediana e (c) a média:

Intervalo de Classe	*f*
90-99	16
80-89	17
70-79	15
60-69	3
50-59	2
40-49	3
	$N = 56$

19. Dada a seguinte distribuição de freqüências, calcule (a) a moda; (b) a mediana; (c) a média:

Intervalo de Classe	*f*
17-19	2
14-16	3
11-13	6
8-10	5
5-7	1
	$N = 17$

5

Medidas de Variabilidade

No Capítulo 4, vimos que a moda, a mediana e a média podiam ser usadas para resumir, num único número, aquilo que é "médio" ou "típico" numa distribuição. Quando empregada sozinha, entretanto, qualquer medida de tendência central fornece apenas uma visão incompleta de um conjunto de dados e, portanto, pode confundir ou distorcer tanto quanto esclarecer.

Com vistas a ilustrar essa situação, admitam que Honolulu (Havaí) e Houston (Texas) tenham quase a mesma temperatura média diária de 75°.* Será que, por isso, podemos admitir que a temperatura é basicamente a mesma em ambas as localidades? Ou não será possível que enquanto uma cidade é melhor para natação a outra o seja para atividades externas?

Como bem ilustra a Figura 5.1, a temperatura em Honolulu varia muito pouco ao longo do ano, oscilando, em geral, entre 70°F e 80°F. Por outro lado, a temperatura em Houston pode diferir estacionalmente, isto é, apresentar-se baixa em janeiro—cerca de 40°F—e alta em julho e agosto—bem perto de 100°F. Desnecessário dizer, pois, que as praias de Houston não estão apinhadas de gente durante todo o ano!

Vamos a um outro exemplo. Suponham que, numa particular cidade, tanto ladrões quanto professores secundários tenham uma renda média anual de Cr$ 136.000,00. Será que essa informação indica que as duas distribuições de renda são *necessariamente* semelhantes? Muito ao contrário, poder-se-ia descobrir que elas diferem—e *muito*—num outro aspecto importante, qual seja, o fato de as rendas dos professores *concentrarem-se* ao redor de Cr$ 136.000,00, enquanto que as dos ladrões espalham-se mais, o que reflete, portanto, maiores oportunidades para prisões, desemprego, pobreza e, em *alguns casos*, fortunas excepcionais.

* N.T.: A falta de uma unidade de medida pode, às vezes, confundir o leitor. Este livro, de autor norte-americano, traduz, em muitos pontos, aspectos da cultura vigente em países anglo-saxões; entre esses aspectos, destaca-se a *escala de temperatura Fahrenheit*, cuja característica é a seguinte: 32°F correspondem ao gelo fundente e 212°F, à agua em ebulição. Em graus *Celsius* (centígrados), 75°F correspondem a 23,9°C (aprox.). A fórmula de conversão usada foi a seguinte:

$$°C = \frac{5}{9}\,(°F - 32°).$$

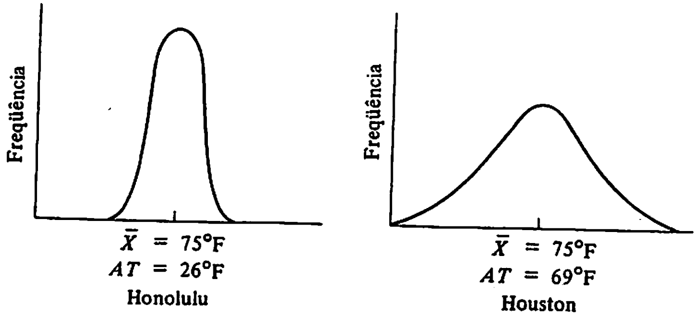


FIGURA 5.1 **Diferenças de Variabilidade: Distribuições de Temperaturas em Locais Distintos: Honolulu e Houston (Valores Aproximados)**

Tal fato demonstra que necessitamos, além de uma medida de tendência central, de um índice que indique o grau de dispersão dos escores em torno do centro da distribuição (isto é, em torno da média). Numa palavra, precisamos de uma medida indicativa do que costumeiramente se chama *variabilidade* (também designada como *variação* ou *dispersão*). Voltando a um exemplo anterior, poderíamos dizer que a distribuição de temperaturas em Houston (Texas) tem *maior variabilidade* do que a distribuição de temperaturas em Honolulu (Havaí). Da mesma forma, podemos dizer que a distribuição de rendas entre professores apresenta *menos variabilidade* do que a distribuição de rendas entre ladrões. O presente capítulo discute apenas as medidas de variabilidade mais conhecidas: a *amplitude total*, o *desvio médio* e o *desvio padrão.*

AMPLITUDE TOTAL

Pode-se obter uma medida de variabilidade rápida (cômoda), embora não muito exata, pelo cálculo da chamada *amplitude total* (*AT*), que nada mais é senão a diferença entre o maior e o menor escore da distribuição. Por exemplo, se a temperatura anual mais alta em Honolulu foi de 88°F e a mais baixa, 62°F, a amplitude total da temperatura anual em Honolulu foi de 26°F (isto é, 88°F – 62°F = 26°F). Se o dia mais quente em Houston apresentou 102°F e o mais frio, 33°F, a amplitude total da temperatura anual em Houston foi de 69°F (isto é, 102°F – 33°F = 69°F).

A vantagem da amplitude total–cálculo rápido e fácil–constitui também sua mais importante desvantagem. Por outras palavras, a amplitude total é inteiramente dependente de *apenas* dois escores: o maior e o menor num dado conjunto de valores. Como resultado, a amplitude total fornece, via de regra, um mero índice grosseiro da variabilidade de uma distribuição. Por exemplo, $AT = 98$ no seguinte conjunto de dados: 2, 6, 7, 7, 10, 12, 13, 100 ($AT = 100 - 2 = 98$); entretanto, $AT = 12$ neste outro conjunto: 2, 6, 7, 7, 10, 12, 13, 14 ($AT = 14 - 2 = 12$). Portanto, pela simples troca de um *único* valor (14 em lugar de 100), fizemos com que a amplitude total flutuasse bruscamente de 98 para 12. Qualquer medida que seja tão afetada pelo escore de um único respondente (ou por um único valor da variável) não pode, por certo, fornecer uma idéia precisa quanto à variabilidade; quando muito, é possível considerá-la como um índice preliminar ou, até, pouco exato.

DESVIO MÉDIO

No capítulo anterior, *desvio** foi definido como sendo a distância entre qualquer escore bruto e a média da distribuição. Aprendemos, assim, que, para calcular discrepâncias (desvios), bastava subtrair a média (aritmética) de qualquer escore bruto: $x = (X - \bar{X})$. Se desejarmos agora obter uma medida de variabilidade que leve em conta *todos* os escores da distribuição (e não *apenas dois*), poderemos tomar o *valor absoluto* de cada discrepância (isto é, das distâncias com relação à média), somar esses valores e, então, dividir o total pelo número de escores. O resultado será o *desvio médio.* Em símbolos:

$$\text{DM} = \frac{\Sigma|x|}{N}\text{, onde}$$

DM = desvio médio
$\Sigma|x|$ = soma dos valores absolutos das discrepâncias (ou seja, soma das discrepâncias *sem* considerar-se a presença dos sinais + e –)
N = total de escores (ou de sujeitos, respondentes ou dados).

Nota importante–No cálculo de $\Sigma|x|$ *devemos* ignorar os sinais "mais" (+) e "menos" (–) e somar apenas os valores absolutos. Este procedimento é *correto,* porque a soma das discrepâncias reais (Σx)–onde os sinais indicam o sentido do desvio com relação à média**–é sempre igual a zero.*** As discrepâncias positivas e as negativas tendem a cancelar-se (compensar-se) e não podem, por isso, ser usadas para descrever ou comparar a variabilidade de distribuições. Por outro lado, a soma dos valores absolutos das discrepâncias tende a tornar-se maior à medida que a variabilidade da distribuição aumenta.

Podemos agora ilustrar, passo a passo, o procedimento para o cálculo do desvio médio. Para tanto, consideremos o seguinte conjunto de escores: 9, 8, 6, 4, 2 e 1.

PASSO 1: Calcular a média (aritmética) da distribuição.

X	
9	$\bar{X} = \frac{\Sigma X}{N}$
8	
6	$= \frac{30}{6}$
4	
2	$= 5$
1	
$\Sigma X = 30$	

* N.T.: Insista-se no seguinte: *desvio, afastamento* e *discrepância* são expressões sinônimas. Entretanto, é prática comum usar o termo *discrepância* para designar qualquer x_i que resulte de $(X_i - \bar{X})$.

** N.T.: Os valores afetados de + (mais) estão situados à *direita* do zero; os afetados de – (menos), à *esquerda.* Nos termos do texto, a *média* (aritmética), à semelhança do zero, funciona como uma *origem.* Então: à direita da média, discrepâncias positivas; à esquerda, discrepâncias negativas.

*** N.T.: Dizer que a soma das discrepâncias é sempre igual a zero é o mesmo que dizer, em símbolos, o seguinte: $\Sigma x = 0$. Ora, $\Sigma x = \Sigma(X - \bar{X})$. Então: $\Sigma(X - \bar{X}) = 0$.

PASSO 2: Subtrair de cada escore bruto a média da distribuição e somar as discrepâncias sem levar em conta os sinais (isto é, "fazer de conta" que *todas* as discrepâncias têm o sinal +).

X	x
9	+4
8	+3
6	+1
4	−1
2	−3
1	−4
$\Sigma X = 30$	$\Sigma\|x\| = 16$

PASSO 3: Dividir $\Sigma|x|$ por N, a fim de garantir uma eqüidistribuição que leve em conta o número de dados envolvido.

$$DM = \frac{\Sigma|x|}{N} = \frac{16}{6} = 2{,}67$$

De acordo com o procedimento precedente, vemos que o desvio médio para o conjunto 9, 8, 6, 4, 2 e 1 é 2,67. Isso indica que, em média, os escores dessa distribuição oscilam 2,67 unidades em torno da média.

Com vistas a entender melhor a utilidade do desvio médio, vamos voltar ao problema das distribuições (a), (b) e (c) de rendas diárias, tal como figuram na Tabela 5.1.

TABELA 5.1 Variabilidade em Distribuições de Rendas Diárias, onde a Média é Fixa (Cr\$340,00)

Distribuição (a)		*Distribuição* (b)		*Distribuição* (c)	
X Cr\$	$\|x\|$	X Cr\$	$\|x\|$	X Cr\$	$\|x\|$
340	0	391	+51	595	+255
340	0	374	+34	510	+170
340	0	357	+17	425	+ 85
340	0	340	0	340	0
340	0	323	−17	255	− 85
340	0	306	−34	170	−170
340	0	289	−51	85	−255
	$\Sigma\|x\| = 0$		$\Sigma\|x\| = 204$		$\Sigma\|x\| = 1.020$
$\bar{X}$ = Cr\$340		$\bar{X}$ = Cr\$340		$\bar{X}$ = Cr\$340	
AT = Cr\$0		AT = Cr\$102		AT = Cr\$510	
DM = Cr\$0		DM = Cr\$29,14		DM = Cr\$145,71	
Nenhuma Variabilidade		Pouca Variabilidade		Muita Variabilidade	

Observem, antes de mais nada, que a média de cada distribuição é Cr\$340,00. Notem também que parece haver diferenças marcantes em termos de variabilidade entre essas distribuições–diferenças que podem ser detectadas com o auxílio da amplitude total e do desvio médio.

Vamos examinar, primeiro, a distribuição de rendas (a), onde todos os valores são idênticos. Uma vez que todos os escores dessa distribuição são expressos pelo mesmo valor numérico (Cr\$ 340,00), podemos dizer que ela não apresenta variabilidade alguma. Todas as pessoas (isto é, as sete da amostra (a)) ganharam a mesma importância naquele determinado dia. Como resultado, a amplitude total é 0 (zero) e não há absolutamente nenhum desvio com relação à média (DM = 0). As distribuições (b) e (c) apresentam variabilidade. De forma específica, a distribuição (b) tem uma amplitude total de Cr\$ 102,00 e um desvio médio de Cr\$ 29,14; a distribuição (c) apresenta uma amplitude total de Cr\$ 510,00 e um desvio médio de Cr\$ 145,71. Podemos, pois, dizer que a distribuição (b) contém menos variabilidade que a distribuição (c), ou seja, as rendas que fazem parte da distribuição (b) são, entre si, mais semelhantes que as rendas da distribuição (c).

DESVIO PADRÃO

Por motivos que logo se tornarão aparentes, o desvio médio já não é muito usado pelos pesquisadores; na maioria dos casos, ele foi abandonado como "medida de variabilidade" em favor de um "parente próximo" mais eficiente: o *desvio padrão*. Como teremos a oportunidade de verificar, entretanto, o desvio médio não pode ser considerado uma perda de tempo, pois, na pior das hipóteses, ele oferece-nos uma boa base para a compreensão da natureza do desvio padrão.

Em discussões anteriores, vimos que o desvio médio evita o problema de números negativos que cancelam os positivos; isso foi conseguido graças à convenção de *ignorarem-se* os sinais "mais" (+) e "menos" (−), somando, em seguida, os valores absolutos dos desvios com relação à média. Tal procedimento para a criação de uma medida de variabilidade tem a nítida desvantagem de impedir que esses valores absolutos sejam sempre úteis em análises estatísticas mais avançadas (pois eles não comportam facilmente manipulações algébricas).

Para superar esse problema e obter uma medida de variabilidade que seja mais conveniente (isto é, ajustada, "usável") em procedimentos estatísticos avançados, podemos *elevar ao quadrado as discrepâncias reais* (com os respectivos sinais) *e somá-las a seguir* (Σx^2). Como ilustrado na Tabela 5.2, esse procedimento permite "fugir" à influência dos sinais, já que o quadrado de qualquer número é sempre *positivo*.

Após termos somado os quadrados das discrepâncias, podemos *dividir o total por N, a fim de garantir uma eqüidistribuição desse resultado (total) relativamente a todos os escores envolvidos.* O valor que se obtém dessa operação é conhecido pelo nome de *média quadrática** (ou média das discrepâncias ao quadrado). (*Nota*–Recorde-se que procedimento semelhante foi adotado no cálculo do desvio médio, quando $\Sigma|x|$ foi dividido por N.) Continuando com a ilustração da Tabela 5.2, vemos que

$$\frac{\Sigma x^2}{N} = \frac{52}{6} = 8{,}67$$

TABELA 5.2 Elevação de Discrepâncias ao Quadrado para a Eliminação de Números Negativos. (Dados da Tabela 5.1)

X	x	x^2
9	+4	16
8	+3	9
6	+1	1
4	−1	1
2	−3	9
1	−4	16
	$\Sigma x = 0$	$\Sigma x^2 = 52$

Surge, agora, um problema adicional. Como conseqüência direta do fato de termos elevado as discrepâncias ao quadrado, a unidade de medida sofreu uma alteração, o que dificulta a interpretação do resultado 8,67. De fato, temos 8,67, mas 8,67 unidades de quê? Com vistas, então, a voltar à unidade de medida original, *extraímos a raiz quadrada da média quadrática*, donde resulta:

$$\sqrt{\frac{\Sigma x^2}{N}} = \sqrt{8{,}67} = 2{,}94$$

Podemos agora definir o desvio padrão–resultado das operações atrás realizadas–como sendo a *raiz quadrada da média das discrepâncias ao quadrado.* Indicando o desvio padrão por DP ou pela letra grega minúscula σ** (sigma), vem:

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}} \quad \text{onde}$$

* N.T.: Na verdade, a *média quadrática* é muito mais conhecida pelo nome de *variância*. Em símbolos, $S^2(X) = \frac{\Sigma x^2}{N}$, onde $S^2(X)$ representa a variância. Às vezes, em lugar de $S^2(X)$ aparece $\sigma^2(X)$; quando tal ocorrer, interprete-se assim: $S^2(X)$ é a variância da *amostra* e $\sigma^2(X)$, da *população*.

** N.T.: Releia-se a nota de rodapé anterior. O que ficou dito relativamente à variância também vale para o desvio padrão. Observe-se, ainda, que o Autor não faz, a esta altura, nenhuma distinção na maneira de notar o desvio padrão, quer seja ele calculado para uma população, quer para uma amostra. Essa distinção será feita, entretanto, no Capítulo 7. Mais uma observação: o *desvio padrão* pode ser definido como sendo a *raiz quadrada positiva da variância.* Assim:

$$+\sqrt{\text{Variância}} = \text{Desvio padrão.}$$

σ = desvio padrão
Σx^2 = soma das discrepâncias ao quadrado
N = total de escores da distribuição.

Para resumir: o procedimento para o cálculo do desvio padrão não difere muito do método que vimos anteriormente para o cálculo do desvio médio. Com referência ao exemplo presente, os seguintes passos devem ser seguidos:

PASSO 1: Calcular a média da distribuição.

X
9
8
6
4
2
1
$\Sigma X = 30$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{30}{6} = 5$$

PASSO 2: Subtrair de cada escore bruto o valor da média da distribuição para, dessa forma, obterem-se as discrepâncias.

X	x
9	+4
8	+3
6	+1
4	−1
2	−3
1	−4

PASSO 3: Elevar ao quadrado cada discrepância e somá-las.

X	x	x^2
9	+4	16
8	+3	9
6	+1	1
4	−1	1
2	−3	9
1	−4	16
		$\Sigma x^2 = 52$

PASSO 4: Dividir por N e extrair a raiz quadrada do resultado.

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{\frac{52}{6}} = \sqrt{8{,}67} = 2{,}94$$

Podemos agora dizer que o desvio padrão do conjunto de escores 9, 8, 6, 4, 2 e 1 é 2,94.

Fórmula para o Cálculo do Desvio Padrão a Partir de Dados Brutos

Até este ponto, a única fórmula usada para o cálculo do desvio padrão tem sido $\sqrt{\Sigma x^2/N}$. No caso particular de estar disponível uma máquina de calcular, há um método mais fácil de obter-se o DP–um método que prescinde das discrepâncias e que trabalha diretamente com os dados brutos.

A fórmula desse novo método é

$$\sigma = \sqrt{\frac{\Sigma X^2}{N} - \overline{X}^2}$$

onde

σ = desvio padrão
ΣX^2 = soma dos quadrados dos escores brutos (*Importante: primeiro*, elevar cada escore bruto ao quadrado; *depois*, efetuar a soma.)
N = número total de escores
$\overline{X}^2$ = média aritmética ao quadrado.

O procedimento passo a passo para o cálculo do DP mediante a aplicação da fórmula anterior pode ser ilustrado a partir dos dados da Tabela 5.2.

PASSO 1: Quadrar (isto é, elevar ao quadrado) cada escore bruto separadamente e somar.

X	X^2
9	81
8	64
6	36
4	16
2	4
1	1
	$\Sigma X^2 = 202$

PASSO 2: Calcular a média aritmética e elevá-la ao quadrado.

X	
9	
8	$\overline{X} = \frac{\Sigma X}{N} = \frac{30}{6} = 5$
6	
4	
2	$\overline{X}^2 = 25$
1	
$\Sigma X = 30$	

PASSO 3: "Entrar" com os resultados dos Passos 1 e 2 na fórmula.

$$\sigma = \sqrt{\frac{\Sigma X^2}{N} - \overline{X}^2} = \sqrt{\frac{202}{6} - 25} = \sqrt{33{,}67 - 25{,}00} = \sqrt{8{,}67} =$$

$$= 2{,}94$$

Como podemos observar acima, a aplicação da nova fórmula aos dados da Tabela 5.2 produz uma resposta idêntica à obtida com o método original.

Cálculo do Desvio Padrão (DP) de uma Distribuição de Freqüências com Dados Isolados

Para obter-se o desvio padrão de uma distribuição de freqüências com dados isolados, aplique-se a fórmula

$$\sigma = \sqrt{\frac{\Sigma f X^2}{N} - \overline{X}^2}$$

Vamos, a título de ilustração, calcular o desvio padrão da seguinte distribuição:

Escores	*f*
7	1
6	2
5	3
4	5
3	2
2	2
1	1
	$N = 16$

PASSO 1: Multiplicar cada escore (X) pela respectiva freqüência (f), a fim de obter fX.

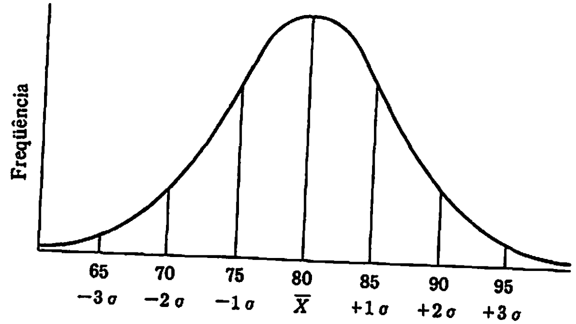


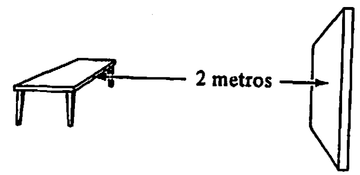
FIGURA 5.2 Mensuração da Linha de Base em Unidades de Desvio Padrão. (Neste exemplo, o desvio padrão (σ) é 5 e a média ($\bar{X}$) é 80)

unidades "sigma"). Por exemplo, poderíamos somar um desvio padrão ao valor da média, a fim de determinar qual escore bruto localiza-se a exatamente um desvio padrão dela (isto é, um escore bruto distante da média *um sigma*). Como ficou demonstrado na Figura 5.2, se $\bar{X} = 80$ e DP = 5, então o escore bruto 85 fica a exatamente um desvio padrão *acima* da média (80 + 5 = 85); isso equivale a dizer que o escore 85 dista dela $+1\sigma$. Essa direção é "mais" (+) porque todas as discrepâncias *acima* da média são positivas; todas as discrepâncias *abaixo* da média são "menos" (–) ou negativas.

Podemos continuar "graduando" a linha de base mediante a adição de mais um desvio padrão ao escore bruto 85. Tal procedimento dá-nos o escore bruto 90, que se situa a exatamente dois desvios padrões acima da média (85 + 5 = 90). Por igual procedimento, adiciona-se um desvio padrão ao escore bruto 90 e obtemos 95—que representa o escore bruto localizado a exatamente três desvios padrões acima da média. Para continuar esse processo do "lado negativo" (abaixo da média), vamos subtraindo, a partir de $\bar{X}$, um desvio padrão, dois desvios padrões etc., donde resulta: 75 (localizado a -1σ), 70 (localizado a -2σ), e assim por diante.

Como bem ilustra a Figura 5.3, o processo de medir a linha de base em unidades de desvio padrão é, em muitos aspectos, semelhante à mensuração da distância em metros entre uma mesa e uma dada parede. No entanto, a analogia falha em pelo menos um

(a)



(b) $\bar{X} = 90$ ← $+2\sigma$ → $\bar{X} = 80$ $\sigma = 5$ unidades de escore bruto

FIGURA 5.3 Mensuração de Distâncias: (a) entre uma mesa e uma parede, onde a unidade de medida usada é o metro; (b) entre um escore bruto e a média da distribuição, onde a unidade de medida vem expressa em unidades de desvio padrão

ponto importante: enquanto que metros, centímetros sempre têm *tamanho constante*, o valor do desvio padrão varia de distribuição para distribuição. Aliás, se não fosse assim, não poderíamos usar o desvio padrão—como ilustramos páginas atrás—na comparação de distribuições quanto à sua variabilidade (por exemplo, DP = Cr$ 85.000,00 para a distribuição da renda de professores secundários; DP = Cr$ 255.000,00 para a distribuição da renda de larápios). Por essa razão, precisamos calcular o "tamanho" do desvio padrão de *todas* as distribuições com que estejamos trabalhando. Também em conseqüência deste fato, é geralmente mais difícil compreender o significado do desvio padrão, em contraste com o significado de metros ou centímetros como unidades de mensuração. No capítulo seguinte voltaremos a examinar o conceito de desvio padrão.

COMPARAÇÃO ENTRE AMPLITUDE TOTAL, DESVIO MÉDIO E DESVIO PADRÃO

A amplitude total pode ser considerada um mero índice preliminar ou grosseiro da variabilidade de uma distribuição. Sua obtenção é rápida e fácil; todavia, ela não é muito "confiável". Pode ser aplicada a dados intervalares ou ordinais.

A amplitude total tem utilidade no que diz respeito ao cálculo do desvio padrão. A Figura 5.2 ilustra que seis desvios padrões cobrem quase que inteiramente a distância que vai desde o menor escore até o maior numa distribuição (de -3σ a $+3\sigma$). Esse fato isolado já nos dá um método conveniente para estimar (mas *não* calcular) o desvio padrão. De modo geral, o tamanho do desvio padrão é cerca de um sexto ($^1/_6$) do tamanho da amplitude total. Por exemplo, se a amplitude total for 36, espera-se que o DP caia bem perto de 6; se a amplitude total for 6, o DP deve de estar próximo de 1.

Essa regra pode assumir considerável importância aos olhos do aluno que deseja "testar" se seu resultado tem grande probabilidade de estar correto. Tomemos um caso extremo: se $AT = 10$ e o DP (calculado) = 12, devemos ter cometido algum erro, pois o DP não pode ser maior que a amplitude total. Uma nota de alerta: a regra do "um sexto" só se aplica a distribuições compostas de grande número de escores. Quando o número de escores for pequeno, via de regra, haverá necessidade de menos desvios padrões para cobrir toda a amplitude da distribuição.

Enquanto que a amplitude total é calculada a partir de *apenas dois escores*, tanto o desvio padrão quanto o desvio médio levam em conta *todos os escores* da distribuição. Apesar da sua relativa estabilidade, o desvio médio não é muito usado em pesquisa, uma vez que ele não se presta a muitas análises estatísticas avançadas. Por outro lado, o cálculo do desvio padrão implica a utilização de um procedimento aceitável do ponto de vista matemático, com vistas a contornar o problema dos sinais (+; –): quadrar as discrepâncias em vez de simplesmente ignorar os sinais. Como resultado, o desvio padrão acabou por tornar-se o passo inicial na obtenção de certas medidas estatísticas, em particular no terreno das inferências. Essa característica do desvio padrão será explorada em detalhes em capítulos subseqüentes, especialmente nos de números 6 e 7.

Apesar da sua utilidade como medida de dispersão "confiável", o desvio padrão também apresenta algumas desvantagens. Comparado com outras medidas de variabilidade, seu cálculo é não só *difícil* como *demorado*. Todavia, essas desvantagens estão dia a dia sendo superadas pelo crescente uso ou de máquinas de calcular de alta velocidade ou de computadores programados para a execução de análises estatísticas. O desvio padrão (da mesma forma que o desvio médio) também tem a característica de ser uma medida intervalar, donde decorre que sua utilização torna-se *impossível* com dados *nominais* ou *intervalares*—via de regra utilizados no trabalho de muitos pesquisadores.

CÁLCULO DA AMPLITUDE TOTAL, DO DESVIO MÉDIO E DO DESVIO PADRÃO DE DADOS AGRUPADOS

Estejam os dados agrupados ou não, a amplitude total é sempre a diferença entre o maior e o menor escore. Seu cálculo não requer o uso de nenhum método especial ou fórmula.

Com o propósito de ilustrar passo a passo o procedimento para a obtenção do desvio médio numa distribuição de freqüências com dados agrupados, considere-se a seguinte tabela*:

Intervalo de Classe	f
17–19	1
14–16	2
11–13	3
8–10	5
5–7	4
2–4	2
	$N = 17$

PASSO 1: Calcular o ponto médio de cada intervalo de classe.

Intervalo	*X = Ponto Médio*
17–19	18
14–16	15
11–13	12
8–10	9
5–7	6
2–4	3

PASSO 2: Determinar a média da distribuição.

X = ponto médio	f	fX	
18	1	18	
15	2	30	$\bar{X} = \frac{\Sigma fX}{N}$
12	3	36	
9	5	45	$= \frac{159}{17}$
6	4	24	
3	2	6	$= 9{,}35$
		$\Sigma fX = 159$	

* N.T.: É sempre interessante especificar, na coluna onde figuram os intervalos de classe, a *unidade* em que a variável foi mensurada.

PASSO 3: Calcular o desvio (= a discrepância) de cada ponto médio com relação à média.

X = Ponto Médio	$X - \bar{X} = \lvert x \rvert$
18	8,65
15	5,65
12	2,65
9	0,35
6	3,35
3	6,35

PASSO 4: Multiplicar cada discrepância pela freqüência correspondente ao intervalo de classe e, em seguida, somar esses produtos.

Intervalo	f	x	fx
17–19	1	8,65	8,65
14–16	2	5,65	11,30
11–13	3	2,65	7,95
8–10	5	0,35	1,75
5–7	4	3,35	13,40
2–4	2	6,35	12,70
	$N = 17$		$\Sigma fx = 55{,}75$

PASSO 5: Dividir por N.

$$\text{DM} = \frac{\Sigma f\lvert x\rvert}{N} = \frac{55{,}75}{17} = 3{,}28$$

Chegamos, assim, a um desvio médio de 3,28.

Para o cálculo do desvio padrão de dados agrupados, podemos usar uma fórmula que se utiliza apenas dos dados brutos da tabela inicial. Em símbolos, temos:

$$\sigma = \sqrt{\frac{\Sigma fX^2}{N} - \bar{X}^2}$$

onde

σ = desvio padrão
f = freqüências dos intervalos de classe
X = ponto médio de cada intervalo de classe
N = número total de escores
$\bar{X}^2$ = quadrado da média (aritmética).

O procedimento passo a passo para o cálculo do desvio padrão pode ser ilustrado, por exemplo, com relação à seguinte tabela de dados agrupados:

Intervalo de Classe	*f*
17–19	1
14–16	2
11–13	3
8–10	5
5–7	4
2–4	2

PASSO 1: Multiplicar cada ponto médio pela freqüência do intervalo de classe correspondente e somar esses produtos.

Intervalo de Classe	*f*	*Ponto Médio (X)*	*fX*
17–19	1	18	18
14–16	2	15	30
11–13	3	12	36
8–10	5	9	45
5–7	4	6	24
2–4	2	3	6
	$N = 17$		$\Sigma fX = 159$

PASSO 2: Calcular a média aritmética e elevá-la ao quadrado.

$$\bar{X} = \frac{\Sigma fX}{N} = \frac{159}{17} = 9{,}35$$

$$\bar{X}^2 = 87{,}42$$

PASSO 3: Multiplicar cada ponto médio por fX e somar esses produtos.

Intervalo de Classe	*f*	*Ponto médio* (X)	*fX*	*fX*^{2}
17–19	1	18	18	324
14–16	2	15	30	450
11–13	3	12	36	432
8–10	5	9	45	405
5–7	4	6	24	144
2–4	2	3	6	18
				$\Sigma fX^2 = 1773$

PASSO 4: "Entrar" com os resultados dos Passos 2 e 3 na fórmula.

$$\sigma = \sqrt{\frac{\Sigma fX^2}{N} - \bar{X}^2} = \sqrt{\frac{1773}{17} - 87{,}42} = \sqrt{104{,}29 - 87{,}42} =$$

$$= \sqrt{16{,}87} = 4{,}11$$

O desvio padrão, conforme sugere a própria fórmula, é 4,11.

RESUMO

Introduzimos, neste capítulo, o conceito de *amplitude total*, de *desvio médio* e de *desvio padrão*—que são três medidas de variabilidade ou, para dizer de outra forma, medidas que "mostram" a maneira como os escores agrupam-se em torno do centro da distribuição. A amplitude total foi definida como um indicador de variabilidade rápido e grosseiro, facilmente calculada a partir da diferença entre o maior e o menor escore da distribuição. O desvio médio—ou seja, a soma dos valores absolutos das discrepâncias (desvios), dividida por N—foi, desde o início, tratado como uma medida inadequada de variação, do ponto de vista matemático, apesar de constituir sólida base para a compreensão do desvio padrão; este, por sua vez, foi definido como sendo a raiz quadrada da variância da distribuição. O desvio padrão é uma medida de variabilidade *confiável*, de nível intervalar, que pode ser utilizada em operações estatísticas avançadas, descritivas ou inferenciais. O significado amplo do desvio padrão será objeto de estudos posteriores: nas discussões relacionadas com a curva normal e nas generalizações de amostras para populações.

PROBLEMAS

1. Um grupo de 5 estudantes obteve as seguintes notas de exame, numa escala de 10 pontos: 7, 5, 3, 2 e 1. Relativamente a esse conjunto de escores, calcule (a) a amplitude total; (b) o desvio médio; (c) o desvio padrão.
2. Numa escala destinada a mensurar atitudes para com o problema da segregação racial, duas classes de universitários obtiveram os seguintes pontos:

Classe A	*Classe B*
4	3
6	3
2	2
1	1
1	4
1	2

Compare a variabilidade de atitudes relativamente à segregação racial entre os membros das classes A e B mediante a determinação do seguinte: (a) amplitude total para cada classe; (b) o desvio médio dos escores de cada classe; (c) o desvio padrão de cada classe. Qual das classes apresenta maior variabilidade de atitudes?

3. Para o seguinte grupo de escores: 3, 5, 5, 4, 1, calcule (a) a amplitude total; (b) o desvio médio; (c) o desvio padrão.

4. Dado o conjunto de escores: 1, 6, 6, 3, 7, 4, 10, calcule o desvio padrão.

5. Dado o conjunto de escores: 12, 12, 10, 9, 8, calcule o desvio padrão.

6. Calcule o desvio padrão relativo à seguinte distribuição freqüencial de escores:

X	f
5	3
4	5
3	6
2	2
1	2
	$N = 18$

7. Calcule o desvio padrão relativo à seguinte distribuição freqüencial de escores:

X	f
7	2
6	3
5	5
4	7
3	4
2	3
1	1
	$N = 25$

8. Calcule o desvio padrão relativo à seguinte distribuição freqüencial de escores:

X	f
10	2
9	5
8	8
7	7
6	4
5	3
	$N = 29$

9. Calcule, relativamente à seguinte tabela de dados agrupados por freqüências, (a) a amplitude total; (b) o desvio médio; (c) o desvio padrão.

Intervalo de Classe	f
90–99	6
80–89	8
70–79	4
60–69	3
50–59	2
	$N = 23$

10. Relativamente à seguinte tabela de dados agrupados por freqüências, calcule: (a) a amplitude total; (b) o desvio médio; (c) o desvio padrão.

Intervalo de Classe	f
17–19	2
14–16	3
11–13	6
8–10	5
5–7	1

11. Relativamente à seguinte tabela de dados agrupados por freqüências, calcule: (a) a amplitude total; (b) o desvio médio; (c) o desvio padrão.

Intervalo de Classe	f
20–24	2
15–19	4
10–14	8
5–9	5
	$N = 19$

PARTE II
Da Descrição à Tomada de Decisões

6

A Curva Normal

Nos capítulos anteriores, vimos que as distribuições de freqüências podem assumir uma grande variedade de formas. Algumas são *simétricas* ou livres de alongamento; outras são negativa ou positivamente alongadas; outras, ainda, apresentam-se com mais de uma "corcova", e assim por diante. Apesar dessa grande diversidade, há uma distribuição de freqüências com a qual muitos de nós já nos familiarizamos, se não por outra razão, pelo fato de termos sido classificados de acordo com ela por nossos professores. Essa distribuição, comumente chamada *curva normal*, é um modelo teórico ou ideal que resulta muito mais de uma equação matemática do que de um real delineamento de pesquisa com posterior coleta de dados.[1] Entretanto, a utilidade da curva normal para o pesquisador pode ser evidenciada através de suas aplicações a efetivas situações de pesquisa.

Como veremos neste capítulo, por exemplo, a curva normal pode ser usada na descrição de distribuições de escores, na interpretação do desvio padrão e em afirmações relacionadas com a noção de probabilidade. Em capítulos subseqüentes, veremos que a curva normal constitui um ingrediente essencial para a tomada de decisões estatísticas, a partir da qual o pesquisador pode generalizar para populações as conclusões a que tenha chegado ao lidar com amostras. Antes de prosseguirmos com a discussão de técnicas para tomada de decisões, faz-se necessário que conheçamos alguns dados relacionados com as propriedades da curva normal.

[1] A curva normal pode ser construída a partir da seguinte fórmula:

$$Y = \frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\bar{X}}{\sigma}\right)^2} \text{, onde}$$

Y = ordenada correspondente a um dado valor de X (isto é, a ordenada representa a freqüência com que X ocorre na distribuição)
π = 3,1416
e = 2,7183

CARACTERÍSTICAS DA CURVA NORMAL

Como pode ser caracterizada a curva normal? Quais as propriedades que a distinguem de outras distribuições? Tal como indica a Figura 6.1, a curva normal é um tipo de curva simétrica, suave, cuja forma lembra um sino e, por isso, é amplamente conhecida por "curva em sino". É possível que o aspecto mais marcante dessa curva seja a sua *simetria*; se "dobrássemos" a curva em seu ponto central (que corresponde à freqüência máxima), daríamos origem a duas metades, sendo que cada uma delas seria a imagem espelhada da outra.*

Além disso, a curva normal é *unimodal*, isto é, possui um só (pico ou) ponto de freqüência máxima; esse ponto, por sua vez, é aquele situado no meio da distribuição (curva), em que a média, a mediana e a moda coincidem. (Recorde-se que, conforme o Capítulo 3, a média, a mediana e a moda ocorrem em pontos diferentes (isto é, não coincidentes) numa distribuição alongada (assimétrica).) A partir do topo (central, arredondado), a curva normal "cai" gradualmente até formar as caudas (duas, uma de cada lado), que se estendem de forma indefinida, aproximando-se cada vez mais da linha de base (eixo das abscissas) sem, entretanto, jamais tocá-la.**

CURVAS NORMAIS: O MODELO E O MUNDO REAL

Poderíamos perguntar: até que ponto as distribuições de dados reais (ou seja, de dados coletados por pesquisadores ao longo de suas pesquisas) ajustam-se ou aproximam-se à figura (ao formato) da curva normal? A título de ilustração, vamos imaginar que todos os fenômenos sociais, psicológicos e físicos sejam normalmente distribuídos. Como seria esse mundo hipotético?

Se atentássemos para as características físicas dos seres humanos, estatura, por

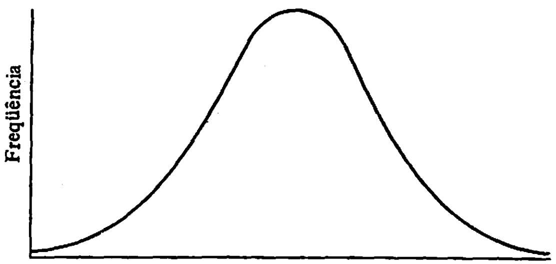


FIGURA 6.1 Formato da Curva Normal

* N.T.: Essa "dobra" representa, de maneira mais formal, um *eixo de simetria*. Tomando-se esse eixo como referência, observa-se que a curva normal é *bicaudal*. Então, para retomar a idéia do autor, dizer que uma metade é a imagem espelhada da outra é o mesmo que dizer o seguinte: "dobrando-se" a curva em seu eixo de simetria, a cauda da *esquerda* ajusta-se (sobrepõe-se) perfeitamente à cauda da *direita*.

** N.T.: Diz-se, por essa razão, que o *campo de variação* de uma distribuição normal estende-se de $-\infty$ a $+\infty$

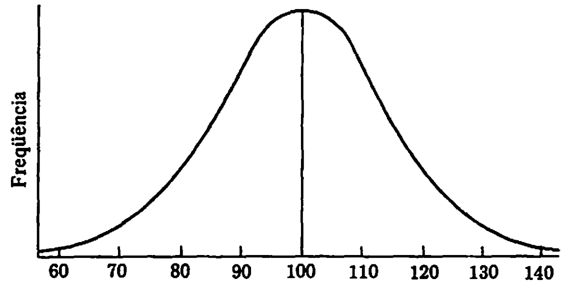


FIGURA 6.2 Distribuição Hipotética de Quocientes de Inteligência (QIs)

exemplo, veríamos que a maioria dos adultos estaria na faixa que vai de 152 cm (aprox.) até 183 cm (aprox.), com muito pouca gente menor que 152 cm ou maior que 183 cm. A Figura 6.2 demonstra que o QI também seria previsível–a maioria dos QIs situando-se entre 90 e 110; como podemos observar, há uma "descida" gradual dos escores para ambas as caudas, com pouquíssimos "gênios" que têm QI superior a 140 e, da mesma forma, pouquíssimas criaturas menos privilegiadas, cujos QIs estão abaixo de 60. Por igual raciocínio, relativamente poucos sujeitos poderiam ser considerados políticos extremistas–de direita ou de esquerda–enquanto que a tendência política da maioria seria considerada moderada.

Finalmente, mesmo o desgaste dos pisos, resultante do fluxo de transeuntes, lembra a distribuição normal: a maior parte do desgaste ocorre no centro dos pisos (degraus etc.), enquanto que nos lados, à medida que nos afastamos do centro, o desgaste vai-se tornando cada vez menor.

Os leitores terão, a esta altura, observado que o mundo hipotético da curva normal não difere de forma radical do mundo "real" em que vivemos no momento. Fenômenos tais como estatura, QI, orientação política, desgaste dos pisos etc. aproximam-se, na prática, até que muito bem da distribuição normal teórica. Pelo fato de tantos fenômenos terem essa característica–isto é, pelo fato de ela ocorrer tão freqüentemente na natureza (e por outras razões que logo se tornarão aparentes)–pesquisadores de diferentes campos têm feito uso extensivo da curva normal, aplicando-a aos dados que eles coletam e analisam.

Observe-se, porém, que alguns fenômenos no campo social–como em qualquer outro–simplesmente não se ajustam à noção teórica da distribuição normal. Muitas distribuições são assimétricas; outras têm mais de uma moda; outras são simétricas, mas não têm a forma de "sino". Como exemplo concreto, consideremos a distribuição de riqueza no mundo. É fato bem conhecido que "os que têm" superam de longe "os que não têm". Assim, como ilustra a Figura 6.3, a distribuição de riquezas (indicada pela renda *per capita*) é de extrema assimetria (pelo menos na aparência), de sorte que apenas uma pequena proporção da população mundial recebe porção significativa da renda total. De forma análoga, especialistas em demografia dizem-nos que os Estados Unidos da América do Norte tornaram-se, nos últimos tempos, uma terra de jovens e velhos. Do ponto de vista econômico, essa distribuição de idades representa um fardo pesado

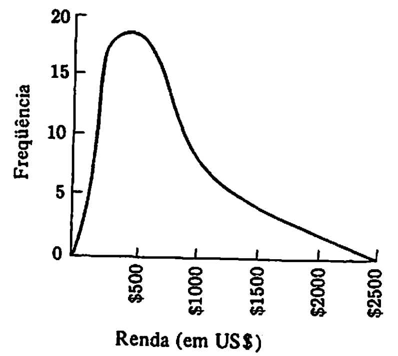


FIGURA 6.3 Distribuição da Renda "Per Capita" (Nações do Mundo)

para um grupo relativamente pequeno de trabalhadores, isto porque, sendo todos cidadãos de meia-idade, têm a seu encargo um número assustador tanto de velhos (aposentados) quanto de jovens (ainda em período escolar).

Naquelas circunstâncias em que temos boas razões para esperar grandes divergências da normalidade–como, por exemplo, no caso da idade e da renda–a curva normal não pode ser usada como "modelo" para os dados coletados. Vemos, assim, que não é possível aplicá-la com liberdade a todas as distribuições que o pesquisador obtém, e deve, ao contrário, ser usada com uma boa dose de bom senso. Felizmente os estatísticos sabem que grande quantidade de fenômenos de interesse segue o modelo normal.

A ÁREA SOB A CURVA NORMAL

A fim de podermos empregar a curva normal na solução de problemas, precisamos, antes, aprender o significado da expressão "área sob a curva normal": *é aquela porção do plano, compreendida entre a curva e a linha de base, que corresponde, em qualquer distribuição normal, a 100% dos dados considerados.* A Figura 6.4 ilustra essa característica.

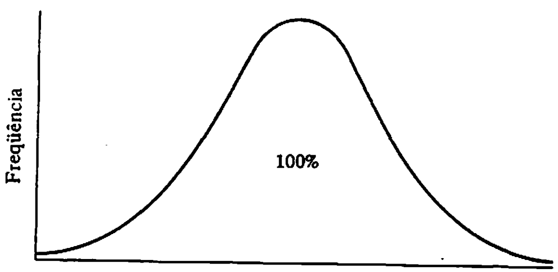


FIGURA 6.4 A Área sob a Curva Normal

Poderíamos limitar uma porção dessa área total se traçássemos, a partir de dois pontos quaisquer tomados na linha de base, segmentos perpendiculares até a própria curva. Por exemplo, usando a média como ponto de partida, poderíamos traçar um segmento em $\bar{X}$ e um outro no ponto que coincide com 1 DP (1 distância sigma) acima (isto é, à direita) de $\bar{X}$. Esta porção sombreada da curva normal ilustrada na Figura 6.5 abrange 34,13% da freqüência total.

Da mesma forma, podemos dizer que 47,72% dos sujeitos sob a curva normal caem entre $\bar{X}$ e 2 DP acima (à direita) da média; igualmente, que 49,87% caem entre $\bar{X}$ e 3 DP acima da média (ver Figura 6.6).

Como veremos mais adiante, *uma proporção constante da área total sob a curva normal cairá entre a média e qualquer distância dada a contar de $\bar{X}$–desde que a mensuração seja feita em unidades de desvio padrão.* Esta afirmação é verdadeira, independentemente da média e do desvio padrão da distribuição particular que estejamos estudando e aplica-se a todos e quaisquer dados que tenham distribuição normal. Assim, a área sob a curva normal compreendida entre $\bar{X}$ e 1 DP acima da média vai *sempre* incluir 34,13% da totalidade dos dados, quer estejamos discutindo a distribuição de estaturas, de inteligência, de filiação partidária ou de desgaste nos degraus das portas à entrada de nossas casas. O requisito básico, em cada situação, é que estejamos trabalhando com escores que tenham *realmente* distribuição *normal.*

A natureza simétrica da curva normal leva-nos a tirar outra conclusão importante: *qualquer distância medida em "sigmas", acima ou abaixo da média, contém a mesma porção da área sob a curva.* Então, se 34,13% da área total situam-se entre a média e 1 DP *acima* de $\bar{X}$, também 34,13% da área total situam-se entre a média e 1 DP *abaixo* de $\bar{X}$; se 47,72% situam-se entre a média e 2 DP *acima* de $\bar{X}$, também 47,72% situam-se entre a média e 2 DP *abaixo* de $\bar{X}$; finalmente, se 49,87% situam-se entre a média e 3 DP *acima* de $\bar{X}$, também 49,87% situam-se entre a média e 3 DP *abaixo* de $\bar{X}$. Em outras palavras, como ilustra a Figura 6.7, 68,26% da área total sob a curva normal (34,13% + 34,13% = 68,26%) caem entre -1σ e $+1\sigma$, sendo a média (aritmética), $\bar{X}$, o ponto de referência; 95,44% da área total (47,72% + 47,72%) caem entre -2σ e $+2\sigma$ a partir de $\bar{X}$; 99,74% da área total–que, aliás, é praticamente toda a área sob a curva–caem entre -3σ e $+3\sigma$ (sempre $\bar{X}$ como ponto de partida). Podemos

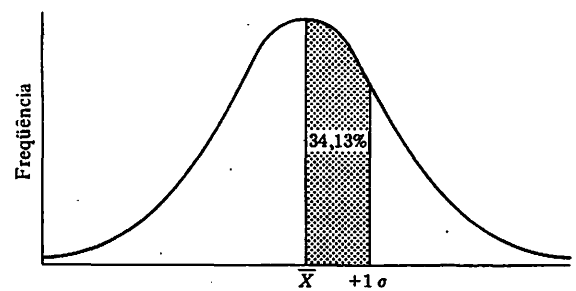


FIGURA 6.5 Porcentagem da Área Total sob a Curva Normal Limitada por $\bar{X}$ e por um Desvio Padrão Acima de $\bar{X}$

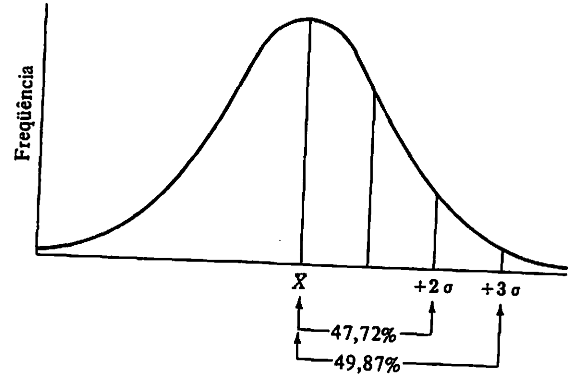


FIGURA 6.6 Porcentagem da Área Total sob a Curva Normal Limitada por $\bar{X}$ e por Dois Desvios Padrões Acima de $\bar{X}$.

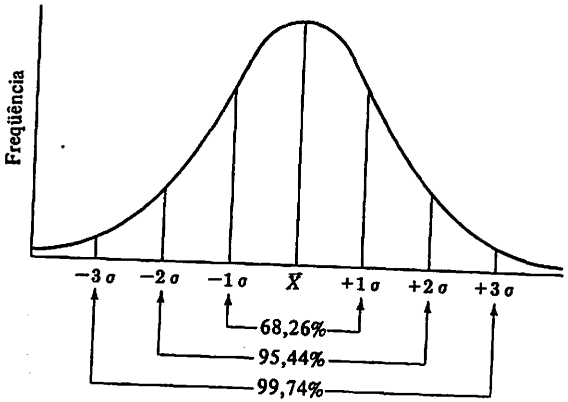


FIGURA 6.7 Porcentagens da Área Total sob a Curva Normal Compreendidas Entre $\pm 1\sigma$, $\pm 2\sigma$ e $\pm 3\sigma$

dizer, então, que 6 DPs abarcam, para efeitos práticos, a totalidade dos dados sob qualquer curva normal (mais de 99%).

TORNANDO A NOÇÃO DE DESVIO PADRÃO MAIS CLARA: UMA ILUSTRAÇÃO

Uma importante característica da curva normal é auxiliar na interpretação e compreensão do desvio padrão. Para entendermos melhor essa característica, vamos examinar o que os antropólogos nos dizem a respeito de diferenças de QIs ligadas a sexo. Muito embora isso contrarie a opinião dos chauvinistas, há provas de que homens e mulheres têm um QI médio aproximadamente igual a 100. Diga-se também, de passagem, que esses QIs diferem acentuadamente em termos de variabilidade em torno da média.

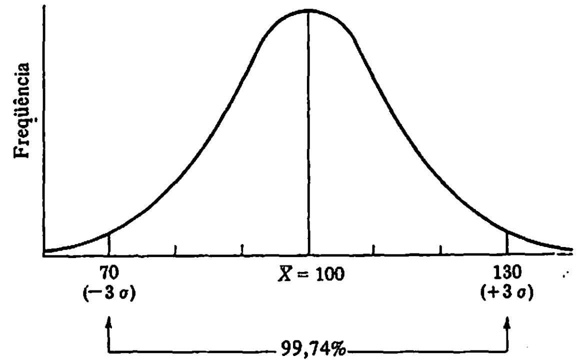


FIGURA 6.8 Distribuição de QIs Masculinos

Neste caso específico, vamos admitir que os QIs masculinos apresentem maior *heterogeneidade* do que os femininos; em outras palavras, a distribuição dos QIs masculinos contém uma porcentagem maior de escores extremos–representativos de sujeitos brilhantes e de sujeitos medíocres–enquanto que a distribuição de QIs femininos contém uma porcentagem maior de escores localizados próximos à média, isto é, ao redor do ponto de freqüência máxima, no centro.

Em virtude de o desvio padrão ser uma medida de variabilidade, essas diferenças ligadas a sexo deveriam refletir-se no valor do DP de cada distribuição de QIs. Poderíamos, assim, verificar, por exemplo, que DP = 10 para sujeitos do sexo masculino e que DP = 5 para os do sexo feminino.

Se conhecêssemos o desvio padrão de cada conjunto de escores QI e admitíssemos que cada conjunto tivesse distribuição normal, poderíamos estimar e, em seguida, comparar as porcentagens de machos e fêmeas localizadas numa dada amplitude de QIs.

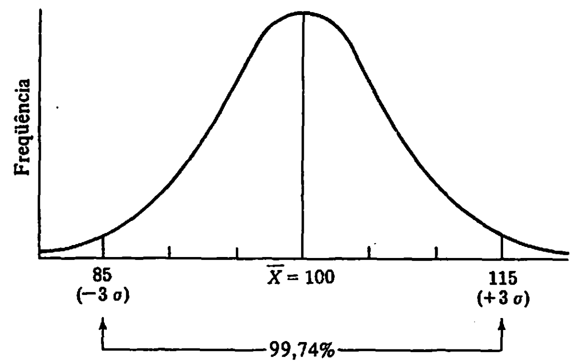


FIGURA 6.9 Distribuição de QIs Femininos

Por exemplo, se medirmos a linha base da distribuição de QIs masculinos em unidades DP (desvio padrão), ficaremos sabendo que 68,26% dos escores caem entre -1σ e $+1\sigma$, a partir da média. Como o desvio padrão é sempre expresso em unidades brutas e $\sigma = 10$, automaticamente descobrimos onde se localizam, na distribuição, os escores 90 e 100 (ou seja: $\bar{X} \pm \sigma = X$; portanto $100 - 10 = 90$ e $100 + 10 = 110$). Desse modo, 68,25% dos representantes do sexo masculino terão, nas condições propostas, QIs situados entre 90 e 110.

Distanciando-nos progressivamente de $\bar{X}$ no sentido das caudas, verificamos, tal como ilustra a Figura 6.8, que 99,74%, isto é, a quase totalidade dos sujeitos (machos), tem QIs entre 70 e 130 (entre -3σ e $+3\sigma$).

Da mesma forma, se examinarmos, a seguir, a distribuição dos QIs femininos–tal como vem ilustrado na Figura 6.9– verificaremos que 99,74% dos dados cairão entre os escores (QIs) 85 e 115 (também entre -3σ e $+3\sigma$). Assim, se compararmos as duas distribuições, veremos que a dos QIs femininos apresenta-se relativamente *homogênea*, uma vez que possui menor proporção de escores extremos em ambas as caudas. Tal diferença reflete-se no "tamanho relativo" de cada DP, e nos escores QIs que caem entre -3σ e $+3\sigma$, a contar da média.

USO DA TABELA B

Na discussão da distribuição normal lidamos, até este ponto, apenas com aquelas distâncias, contadas a partir da média, que são múltiplos exatos do desvio padrão. Em outras palavras, até aqui só trabalhamos com DPs iguais a 1, 2 ou 3, acima ou abaixo de $\bar{X}$. O problema que se coloca agora é o seguinte: que devemos fazer para determinar a porcentagem de sujeitos que se situam entre *duas quaisquer ordenadas* (não expressas por *números inteiros*)? Por exemplo, suponha o leitor que desejemos estabelecer a porcentagem da freqüência total que cai entre a média e, digamos, um escore bruto localizado a 1,40 DPs acima da média. Como bem demonstra a Figura 6.10, um escore bruto situado a 1,40 DPs acima da média é obviamente maior que 1 DP, porém menor que 2 DPs–contados, ambos a partir de $\bar{X}$. Sabemos, assim, que essa distância, calculada

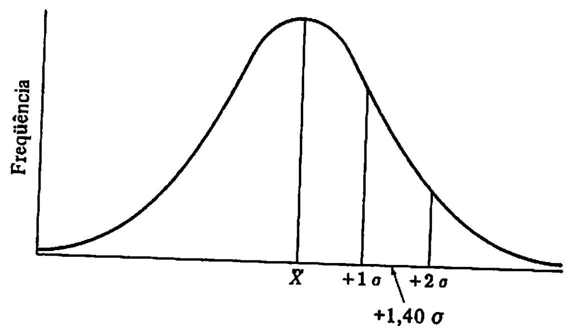


FIGURA 6.10 Localização de um Escore (Bruto) Localizado a 1,40 DPs Acima de $\bar{X}$

a partir da média, deveria incluir *mais* de 34,13%, porém *menos* de 47,72% do total da área sob a curva normal.

Para calcular a porcentagem *exata* (dos sujeitos que caem dentro) desse intervalo, devemos empregar a Tabela B do final do livro, tabela essa que fornece a porcentagem sob a curva normal entre a média e diversas distâncias sigmas (medidas a partir de $\bar{X}$). Tais distâncias sigmas (de 0,0 a 5,0) aparecem na coluna do lado esquerdo da Tabela B e foram calculadas até a primeira casa decimal. A segunda casa decimal aparece na primeira linha, no topo da tabela.

Observe-se que a simetria da curva normal torna possível calcular as porcentagens para *apenas um dos lados* da média*, ou seja, relativamente à metade da área da curva (50%). Os valores da Tabela B correspondem a qualquer dos lados.** Abaixo figura um extrato dessa tabela.

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	00,00	00,40	00,80	01,20	01,60	01,99	02,39	02,79	03,19	03,59
0,1	03,98	04,38	04,78	05,17	05,57	05,96	06,36	06,75	07,14	07,53
0,2	07,93	08,32	08,71	09,10	09,48	09,87	10,26	10,64	11,03	11,41
0,3	11,79	12,17	12,55	12,93	13,31	13,68	14,06	14,43	14,80	15,17
0,4	15,54	15,91	16,28	16,64	17,00	17,36	17,72	18,08	18,44	18,79

Para aprender a usar e entender a Tabela B, poderíamos, primeiro, tentar localizar a porcentagem de dados situados entre uma distância sigma igual a 1,0 e a média. (A razão disso é simples: já sabemos que 34,13% da área total caem entre esses dois pontos na linha base.) Ao observar a Tabela B, verificamos que, de fato, ela indica que 34,13% da freqüência total caem entre a média e uma distância sigma igual a 1,00. Da mesma forma, vemos que a distância sigma igual a 2,00 abarca exatamente 47,72% da área total sob a curva, enquanto que a distância sigma igual a 2,01 corresponde a 47,78% dessa área total.

ESCORES PADRONIZADOS E A CURVA NORMAL

Estamos agora preparados para calcular a porcentagem da área total, sob a curva normal, associada a qualquer distância sigma que seja tomada a partir da média. Entretanto, ainda resta uma questão importante por responder: como determinamos a distância sigma de qualquer escore dado? Em outras palavras, como fazemos para traduzir nosso escore bruto–aquele que foi originalmente colhido no contato com os respondentes, por exemplo–em unidades de desvio padrão? Se desejássemos traduzir pés em jardas, simplesmente dividiríamos o número de pés por 3, uma vez que há 3 pés numa jarda.*** Da mesma forma, se estivéssemos traduzindo minutos em horas, dividiríamos o número de minutos por 60, já que há 60 minutos em cada hora. De maneira idêntica, podemos

* N.T.: Por questão de comodidade, o lado escolhido é o da *direita*, já que, assim, nenhum valor precisa vir antecedido de sinal.

** N.T.: Observe-se o que ficou dito na N. do T. anterior.

*** N.T.: 1 pé = 30,48 cm. 1 jarda = 3 pés e, portanto, igual a 3(30,48 cm) = 91,44 cm.

traduzir qualquer escore bruto em unidades de desvio padrão, dividindo a distância* que separa esse escore da média ($\bar{X}$) pelo desvio padrão da distribuição. Para ilustrar: vamos imaginar um escore bruto igual a 6 numa distribuição cuja média = 3 e cujo desvio padrão = 2. Calculando-se a diferença entre esse escore bruto e a média, obtém-se um escore-diferença representado por (6 – 3); esse cálculo simples mostra que um escore bruto igual a 6 está a 3 unidades acima da média–sendo que essa "medida" também é expressa em escores brutos. Se dividirmos esse escore-diferença (discrepância) pelo desvio padrão, que é igual a 2, veremos que o escore bruto "6" está a 1,5 (um e meio) desvios padrões acima da média. Em outras palavras, a distância sigma de um escore igual a 6, *nesta distribuição particular*, é 1,5. Note-se que, independentemente da "situação de mensuração", sempre há 3 pés numa jarda e 60 minutos numa hora. A constância que caracteriza todas essas medidas padronizadas não se aplica ao desvio padrão, uma vez que ele varia de distribuição para distribuição. Por essa razão, precisamos conhecer o desvio padrão de uma distribuição ou por *cálculo* ou por *estimativa* ou por *informação* de terceiros sempre que tenhamos em mente traduzir um determinado escore bruto em unidades de desvio padrão.

O processo que acabamos de ilustrar–ou seja, o cálculo da distância sigma a partir de $\bar{X}$–produz um valor chamado *escore z* ou *escore padronizado*, que indica, *em unidades de desvio padrão, o sentido e o grau com que um dado escore bruto se afasta da média da distribuição à qual ele pertence.* (Observe-se que a coluna do lado esquerdo da Tabela B no fim do livro vem encimada com z.) Assim, um escore z de +1,4 indica que o escore bruto fica a 1,4 DPs (ou quase 1½ DPs) *à direita* (acima) da média, enquanto que um escore z de –2,1 significa que o escore bruto correspondente cai *à esquerda* (abaixo) da média, num ponto ligeiramente superior a 2 DPs (ver Figura 6.11).

Obtemos um escore z através do cálculo do escore-diferença ($x = X - \bar{X}$)–que dá a distância de um X qualquer até a média–e, então, pela divisão dessa diferença por σ.

Em símbolos:

$$z = \frac{X - \bar{X}}{\sigma} \quad \text{ou} \quad \frac{x}{\sigma}\text{, onde}$$

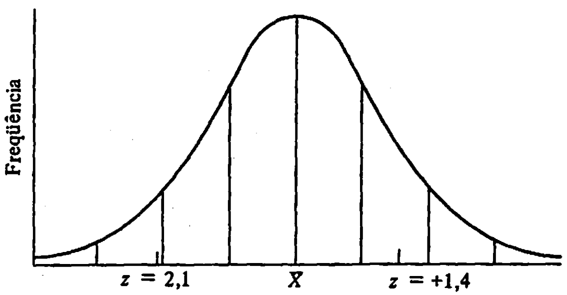


FIGURA 6.11 Posição de $z = -2,1$ e de $z = +1,4$ Numa Distribuição Normal

* N.T.: Essa distância entre o escore bruto e a média é a própria *discrepância*.

x = escore-diferença ou escore-discrepância
σ = desvio padrão da distribuição
z = escore padronizado

Exemplo 1

Vamos estudar a distribuição da renda anual de uma cidade norte-americana. A renda anual média é de US\$5.000 e o desvio padrão, US\$1.500. Admitindo-se que a distribuição da renda anual tenha distribuição normal, podemos traduzir um escore bruto de, por exemplo, US\$7.000 em escore padrão do seguinte modo:

$$z = \frac{7.000 - 5.000}{1.500} = +1,33$$

Vemos, assim, que uma renda anual de US\$7.000 corresponde a um 1,33 desvios padrões acima da média anual de US\$5.000 (ver Figura 6.12).

Exemplo 2

Ao trabalhar com uma distribuição normal de escores representativos do grau de contentamento manifestado por um grupo de inquilinos beneficiários do serviço público de habitação, vamos supor que $\bar{X} = 10$ e DP = 2. (Por convenção, quanto maior o escore, maior a satisfação advinda do serviço público.) Para determinarmos quantos desvios padrões um escore igual a 3 dista da média ($\bar{X} = 10$), a primeira coisa que devemos fazer é obter a diferença entre esse escore e $\bar{X}$. Assim:

$$z = X - \bar{X} = 3 - 10 = -7$$

Dividindo, então, o resultado pelo desvio padrão, vem:

$$z = \frac{x}{\sigma} = -\frac{7}{2} = -3,5$$

Portanto, como bem demonstra a Figura 6.13, um escore bruto de 3, nessa distribuição, cai a 3,5 desvios padrões *abaixo* (à esquerda) da média.

Nota: conhecendo-se um escore z (padronizado), é sempre possível obter-se o escore bruto equivalente mediante o emprego da seguinte fórmula:

$$X = z\sigma + \bar{X}$$

Entrando com os dados do presente exemplo nessa fórmula, resulta:

$$X = (-3,5)(2) + 10 = -7 + 10 = 3$$

PROBABILIDADE E CURVA NORMAL

Como veremos a seguir, a curva normal pode ser usada em conjunção com os escores z

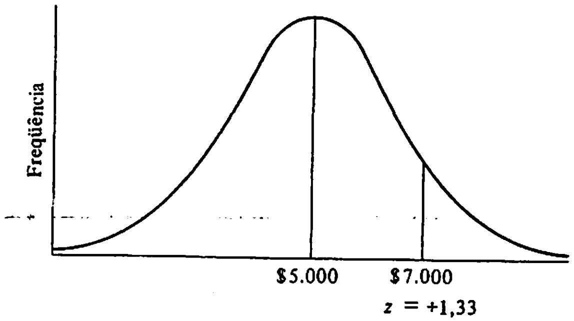


FIGURA 6.12 Posição de $z = 1{,}33$ (na Curva Normal), Relativa ao Escore Bruto US$7.000

(escores padronizados) e a Tabela B no cálculo da probabilidade associada a qualquer dado de uma distribuição. No contexto presente, o termo *probabilidade* refere-se à freqüência relativa de ocorrência de um dado ou evento qualquer; em outras palavras, *a probabilidade associada a um evento qualquer é o número de vezes que tal evento pode ocorrer em relação ao número total de eventos.* Redispondo os termos dessa definição:

$$\text{Probabilidade de um dado ou de um evento qualquer} = \frac{\text{Número de vezes que o dado ou evento pode ocorrer}}{\text{Número total de dados ou eventos}}$$

Assim, a probabilidade de extrair uma única carta (digamos, um "ás de espadas"), de um maço embaralhado de 52 cartas, é 1 em 52, uma vez que o evento "ás de espadas" só pode ocorrer uma vez no total (pois entre as 52 cartas só existe *um* "ás de espadas"). A probabilidade de obter-se "cara" com uma moeda honesta* (ou perfeitamente balanceada), numa só jogada é 1 em 2, já que "cara" só ocorre uma vez no conjunto total de possibilidades (2). Por igual raciocínio, se nos mandassem abrir um livro de 100 páginas numa página qualquer (por exemplo, pág. 23),** a probabilidade de conseguir, numa só tentativa e aleatoriamente, abrir o livro *naquela* página seria de 1 em 100.

A curva normal é uma distribuição na qual é possível determinar probabilidades associadas a *todos* os pontos da linha de base. Como já foi mencionado anteriormente, a curva normal é uma *distribuição de freqüências*; a freqüência total sob a curva é igual a 100%; essa curva apresenta uma área central que circun-

* N.T.: Além do termo "honesto"(a), muito comum em Estatística, ocorrem, ainda: *eqüiprovável, balanceado(a), não viesado(a).*

**N.T.: Esse raciocínio é falho! Se a idéia for abrir o livro na *página 23*, o mais "ingênuo" dos pesquisadores procurará abri-lo na "primeira metade" e, aí, o 100 fica sem sentido. Melhor seria suprimir a sugestão quanto à página, deixando apenas *qualquer.*

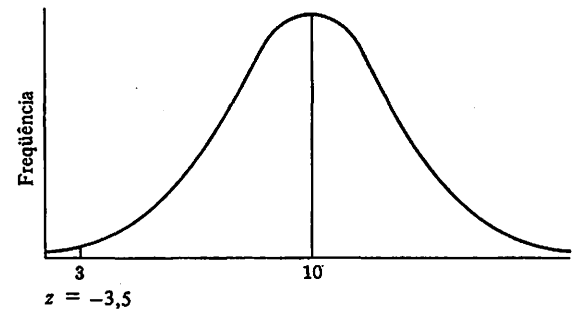


FIGURA 6.13 Posição de $z = -3{,}5$ Relativa ao Escore Bruto = 3

da a média, onde se localizam os escores mais freqüentes, e há, ainda, áreas menores progressivamente mais próximas de ambas as extremidades (caudas), onde encontramos, em pequenas proporções, escores muito altos ou muito baixos. Então, em termos probabilísticos, podemos dizer que a probabilidade decresce à medida que, na linha de base, nos afastamos da média em ambos os sentidos. Desse modo, dizer que 68,26% da freqüência total sob a curva normal caem entre -1σ e $+1\sigma$, a partir da média, é o mesmo que dizer que a probabilidade é de cerca de 68 em 100 de que um escore bruto qualquer caia dentro desse intervalo ($-1\sigma \longmapsto +1\sigma$). De forma análoga, dizer que 95,44% da freqüência total sob a curva normal caem entre -2σ e $+2\sigma$, a contar da média, é o mesmo que dizer que a probabilidade é de aproximadamente 95 em 100 de que um escore bruto qualquer venha a situar-se dentro desse intervalo ($-2\sigma \longmapsto +2\sigma$) e assim por diante.

O que acabamos de fazer foi repetir precisamente o mesmo conceito de probabilidade ou *freqüência relativa* que já estudamos quando extraímos uma carta de um baralho completo, quando jogamos uma moeda ou quando abrimos de forma aleatória um livro numa página qualquer. Note-se, contudo, que as probabilidades associadas às áreas sob a curva normal são sempre expressas com relação a 100%—que correspondem à área *total* sob a curva (por exemplo, 68 em 100; 95 em 100; 99 em 100 e assim por diante). Por essa razão e para garantir que em todo o resto deste livro o termo seja entendido de modo padronizado, vamos sempre falar em *probabilidade como sendo o número de vezes que um dado evento pode ocorrer num conjunto de 100 repetições.* Assim, a probabilidade de sortear um ás de espadas de um baralho bem misturado é de 1,92 em 100 ($\frac{1}{52}$); a probabilidade de tirar "cara" no lance de uma moeda é de 50 em 100 ($\frac{1}{2}$). Note-se, além disso, que é usual expressar probabilidades sob a forma decimal (P). Por exemplo, podemos dizer que $P = 0{,}50$—isto é, ($\frac{50}{100}$)—de tirar "cara" num só lance; analogamente, podemos dizer que $P = 0{,}68$—ou seja, ($\frac{68}{100}$)—de que um escore bruto caia entre -1σ e $+1\sigma$ sob a curva normal.

Expressa sob a forma de razão (quociente), *a probabilidade será sempre um número que oscila entre 0 e 1.* A probabilidade de ocorrência de um evento é 0 quando estamos absolutamente seguros de que ele *não ocorrerá*; é 1 quando estamos convencidos de que sem dúvida nenhuma ele *ocorrerá.* O problema é que os pesquisadores nunca estão totalmente seguros a respeito de coisa alguma! Em conseqüência, podemos, via de regra, esperar encontrar probabilidades iguais a 0,60, 0,25 ou 0,05; mas raras vezes é possível esperar reduzir a probabilidade a 0 ou, por outro lado, elevá-la a 1.

Uma característica importante das probabilidades reside na *regra da soma*, que estabelece o seguinte: *a probabilidade de ocorrência de um evento qualquer, num conjunto de alternativas, é igual à soma das probabilidades separadas.* Suponham, por exemplo, que desejemos encontrar a probabilidade de extrair, dum maço bem embaralhado de 52 cartas, *ou* o "ás de espadas", *ou* a "rainha de ouros" *ou* o "rei de copas" num único lance. Pela soma das probabilidades isoladas–$(\frac{1}{52}+\frac{1}{52}+\frac{1}{52})$–verificamos que a probabilidade de obtenção de qualquer dessas cartas, numa única tentativa é igual a $\frac{3}{52}$, ou seja, $P = 0{,}06$. Em outras palavras, temos 6 chances em 100 de obter, numa única tentativa, ou um "ás de espadas", *ou* uma "dama de ouros" *ou* um "rei de copas" (ver Figura 6.14).

A regra da soma sempre pressupõe que os eventos sejam *mutuamente excludentes*, em outras palavras, que dois eventos não podem jamais ocorrer em simultaneidade. Por exemplo, num baralho de 52 cartas, não é possível que, ao mesmo tempo, uma carta seja de "espadas", de "ouros" *e* de "copas". Analogamente, uma moeda, jogada uma só vez, não pode produzir em conjunto "cara" *e* "coroa".

Admitida a mútua exclusão de eventos, podemos dizer que as probabilidades associadas a cada uma das possíveis ocorrências de um evento sempre têm 1 por soma. Essa afirmação estabelece que *algo tem que ocorrer.* Se não sair "cara", sai "coroa"; se não for um "ás", será um "rei", "rainha", "valete", "dez" e por aí afora. No lançamento de uma moeda, a probabilidade de sair "cara" é de $\frac{1}{2}$ (isto é, $P = 0{,}50$). Naturalmente, a probabilidade de sair "coroa" é também de $\frac{1}{2}$ (ou seja, $P = 0{,}50$). Então, se somarmos as probabilidades relativas a todas as ocorrências, verificaremos que a probabilidade de sair "cara" *ou* "coroa" é 1 (isto é, $\frac{1}{2}+\frac{1}{2} = 1$).

Outra propriedade importante das probabilidades é a *regra da multiplicação*: aí, o problema consiste em determinar a probabilidade de realização de dois ou mais eventos *em sucessão*, um após o outro. A regra da multiplicação estabelece o seguinte:

Probabilidade de extrair um "ás de espadas" $= \frac{1}{52}$

Probabilidade de extrair uma "dama de ouros" $= \frac{1}{52} +$

Probabilidade de extrair um "rei de copas" $= \frac{1}{52}$

Probabilidade de extrair *ou* um "ás de espadas", *ou* uma "dama de ouros" *ou* um rei de copas" $\frac{3}{52}$; isto é, $P = 0{,}06$

FIGURA 6.14 Exemplo Ilustrativo da Regra da Soma: Trata-se, Aqui, de Extrair num Só Lance, *ou* um "Ás de Espadas", *ou* uma "Dama de Ouros" *ou* um "Rei de Copas"

Probabilidade de sair "cara" no primeiro lance → $\frac{1}{2}$

Probabilidade de sair "cara" no segundo lance → $\frac{1}{2}$ ×

Probabilidade de saírem "caras" nas duas jogadas sucessivas → $\frac{1}{4}$ (isto é, $P = 0{,}25$)

FIGURA 6.15 Ilustração da Regra da Multiplicação: Cálculo da Probabilidade de Saírem "Caras" em 2 Jogadas Sucessivas da Mesma Moeda

a probabilidade de obtenção de eventos mutuamente excludentes em série é igual ao produto das probabilidades separadas. Em lugar de "ou ... ou", a regra da multiplicação sugere "primeiro, segundo, terceiro, ..."

Por exemplo, qual a probabilidade de obter "cara" em duas jogadas sucessivas de uma moeda (isto é, "cara" tanto no primeiro lance quanto no segundo)? Visto que esses eventos são independentes, o evento relacionado com o primeiro lance não influencia o segundo. Na primeira jogada da moeda, a probabilidade de sair "cara" é igual a $\frac{1}{2}$ (ou seja, $P = 0{,}50$); na segunda jogada, a probabilidade desse mesmo evento ainda é igual a $\frac{1}{2}$ (isto é, $P = 0{,}50$). Portanto, a probabilidade de conseguirmos "cara" em dois lances sucessivos de uma moeda é igual a $(\frac{1}{2})(\frac{1}{2}) = \frac{1}{4}$ (ou $P = 0{,}25$) (ver figura 6.15).

Para aplicarmos o conceito precedente de probabilidade com referência à curva normal, vamos voltar a um exemplo anterior. Recorde-se o leitor de que naquele problema relacionado com a distribuição de renda anual de uma cidade (norte-americana), um particular escore bruto teve de ser transformado no seu eqüivalente z. Tal distribuição de rendas tinha média de US\$ 5.000 e desvio padrão de US\$ 1.500.

Ao aplicarmos a fórmula da "escala z", verificamos, anteriormente, que uma renda anual de US\$ 7.000 situava-se a 1,33 DPs acima da média (US\$ 5.000); ou seja,

$$z = \frac{7.000 - 5.000}{1.500} = +1{,}33$$

Vamos agora determinar a probabilidade associada a um escore situado entre US\$ 5.000–a própria média–e US\$ 7.000. Em outras palavras, qual a probabilidade de escolher-se de forma aleatória, numa só tentativa, uma pessoa dessa cidade cuja renda anual caia entre US\$ 5.000 e US\$ 7.000? O problema está ilustrado graficamente na Figura 6.16 (observe-se a área sombreada sob a curva) e pode ser resolvido (em matemática) pela aplicação da fórmula dos escores z, combinada com a Tabela B (fim do livro).

PASSO 1: Transformar o escore bruto (US\$ 7.000) em escore z*

* N.T.: Além de "escore z", é comum falar-se em "escore reduzido", "(variável) normal reduzida" ou, simplesmente, valor da "escala z".

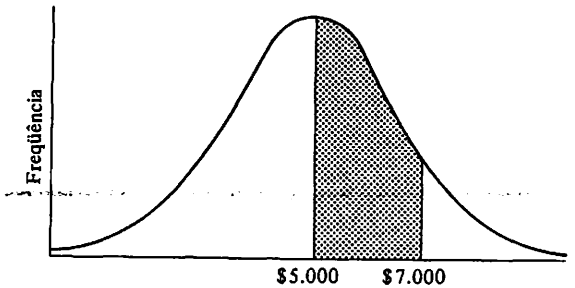


FIGURA 6.16 Proporção da Área Total (sob a Curva Normal) à qual Desejamos Associar uma Probabilidade de Ocorrência

$$z = \frac{X - \bar{X}}{\sigma} = \frac{7.000 - 5.000}{1.500} = +1{,}33$$

Vemos, assim, que o escore bruto US$ 7.000 está a 1,33 DPs acima da média.

PASSO 2: Usando a Tabela B, procurar a porcentagem da freqüência total (sob a curva) que cai entre o escore z ($z = +1{,}33$) e a média.

Na Tabela B, verificamos que 40,82% (quase 41%) da população total dessa cidade ganha entre US$ 5.000 e US$ 7.000 (ver Figura 6.17). Assim, se caminharmos com a vírgula duas casas para a esquerda e arredondarmos, veremos que a probabilidade é de 41 em 100; $P = 0{,}41$ de que, ao extrair aleatoriamente um sujeito dessa população, sua renda anual esteja compreendida nesses limites.

No exemplo anterior, o problema era determinar a probabilidade associada à distância entre a média e um particular valor sigma. Muitas vezes, entretanto, é possível que queiramos achar a porcentagem da área associada a um escore bruto particular;

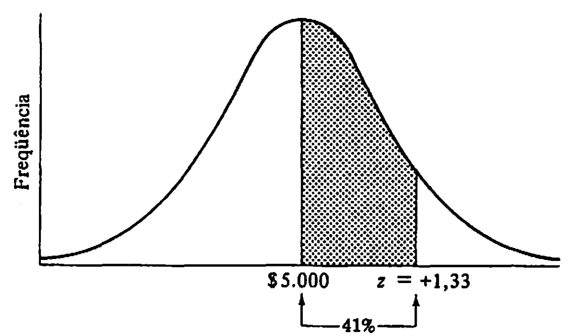


FIGURA 6.17 Porcentagem da Área Total sob a Curva Entre $\bar{X}$ = US$ 5.000 e $z = +1{,}33$

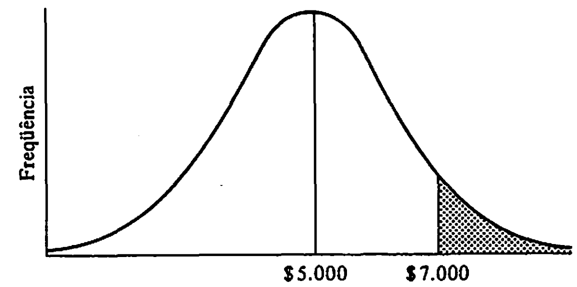


FIGURA 6.18 Porção da Área Total (sob a Curva Normal) Cuja Probabilidade de Ocorrência Desejamos Estabelecer

outras vezes, a porcentagem da área *até* esse escore; finalmente, a porcentagem da área *além* dele, caso em que a idéia de "caudalidade" deve estar presente.* Por exemplo, no caso em pauta, talvez desejássemos saber qual a probabilidade associada ao evento "renda anual = US$ 7.000 ou *mais que* US$ 7.000".

Este problema pode ser ilustrado graficamente, conforme a Figura 6.18 (ver área sombreada). Neste caso, o procedimento é seguir os passos 1 e 2 acima, a fim de obter o escore z e encontrar a porcentagem sob a curva normal entre $\bar{X}$ = US$ 5.000 e $z = 1{,}33$ (ver Tabela B). No presente caso, entretanto, precisamos ir um passo além e *subtrair* de 50% a porcentagem obtida na Tabela B—ou seja, aquela porcentagem da área que se localiza à direita (ou à esquerda) de $\bar{X}$. Esta afirmação é verdadeira porque a *Tabela B sempre registra a porcentagem da área que vai de $\bar{X}$ até um dado z* (e nunca a porcentagem associada a um z particular ou à área situada além dele).

Portanto, subtraindo 40,82% de 50%, verificamos que pouco mais de 9% (9,18%) caem *em* ou *além de* US$ 7.000. Em termos probabilísticos podemos dizer (caminhando com a vírgula duas casas para a esquerda) que só há pouco mais de 9 chances em 100 (isto é, $P = 0{,}09$) de encontrarmos casualmente, nessa cidade, um sujeito cuja renda anual seja *igual a* ou *maior que* US$ 7.000.

Já mencionamos anteriormente que qualquer distância sigma *acima* da média contém uma proporção de sujeitos (dados) idêntica à encontrada na mesma distância sigma *abaixo* da média. Por essa razão, o procedimento para calcular probabilidades associadas a pontos (valores, dados, escores) abaixo de $\bar{X}$ é idêntico ao adotado nos exemplos acima.

Por exemplo, a porcentagem da freqüência total entre $z = -1{,}33$ (US$ 3.000) e a média é idêntica à porcentagem entre $z = 1{,}33$ (US$ 7.000) e a média. Portanto, sabemos que a probabilidade de sortear um sujeito dessa população (cidade) cuja renda anual caia entre US$ 3.000 e US$ 5.000 é de $P = 0{,}41$. Por igual raciocínio, a porcen-

* N.T.: "Caudalidade" é um neologismo necessário. Ter presente a "caudalidade" significa observar se o escore em foco está *à direita* ou *à esquerda* na curva. Fala-se, aí, em *unicaudalidade*. Também é possível, porém, em casos especiais, pensar-se em *bicaudalidade*, circunstâncias em que dois escores são considerados simultaneamente.

tagem da freqüência total *em* ou *além de* $z = -1{,}33$ (ou seja, US$ 3.000 ou *menos*) é igual à porcentagem *em* ou *além de* $z = +1{,}33$ (US$ 7.000 ou *mais*). Desse modo, sabemos que $P = 0{,}09$ de obtermos aleatoriamente alguém dessa cidade com uma renda anual de US$ 3.000 ou menos.

Podemos usar a regra da adição para encontrar a probabilidade de obtenção de mais de uma porção da área sob a curva normal. Por exemplo, já determinamos que $P = 0{,}09$ tanto para rendas de US$ 3.000 ou menos quanto para rendas de US$ 7.000 ou mais. Para encontrarmos a probabilidade de obter "US$ 3.000 ou menos" *ou* "US$ 7.000 ou mais", simplesmente somamos as probabilidades individuais (isto é, separadas):

$$P = 0{,}09 + 0{,}09 = 0{,}18$$

De modo similar, podemos calcular a probabilidade de sortear alguém cuja renda caia entre US$ 3.000 e US$ 7.000; para tanto, basta somar as probabilidades associadas ao escore $z = 1{,}33$ de cada lado da média.* Portanto,

$$P = 0{,}41 + 0{,}41 = 0{,}82$$

Observe-se que $0{,}82 + 0{,}18 = 1$, resultado que corresponde à soma de todas as ocorrências possíveis relacionadas com um mesmo fenômeno.

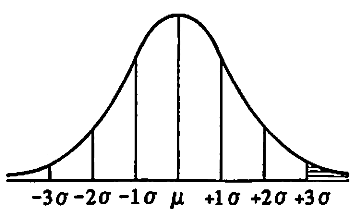
A aplicação da regra da multiplicação à curva normal pode ser ilustrada com o seguinte problema: qual a probabilidade de sortearmos de forma aleatória, nessa população, quatro sujeitos cujas rendas anuais sejam de US$ 7.000 ou mais? Já sabemos que a probabilidade de extrair casualmente *um* sujeito (dessa população) cuja renda anual seja *no mínimo* igual a US$ 7.000 é $P = 0{,}09$. Portanto,

$$P = (0{,}09)(0{,}09)(0{,}09)(0{,}09) = (0{,}09)^4 = 0{,}00007$$

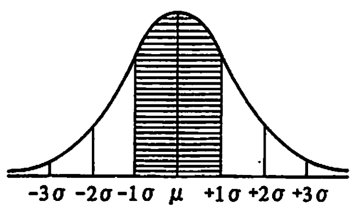
Pela aplicação da regra da multiplicação, vemos que a probabilidade de extração aleatória de quatro sujeitos de renda igual ou superior a US$ 7.000 corresponde a 7 (chances) em 100.000.

† EXERCÍCIOS RESOLVIDOS

1. Qual a probabilidade de z pertencer à região hachurada?

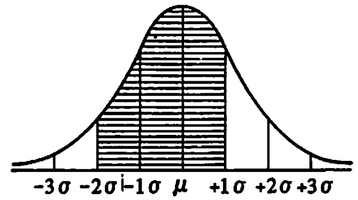


(a)

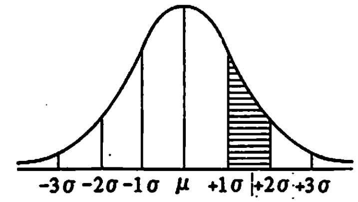


(b)

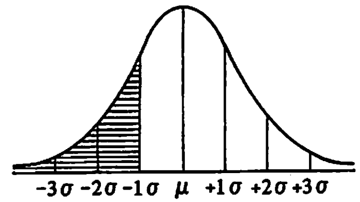
* N.T.: Aqui, por exemplo, o problema exige que pensemos em "bicaudalidade".



(c)



(d)



(e)

Solução

a) Neste exercício, como nos demais, o eixo das abscissas já vem graduado em desvios padrões (σ). Isto quer dizer que a solução é quase imediata, bastando consultar a Tabela B — e respeitar a nota de rodapé.
A região hachurada começa no 3º desvio padrão. Logo, o que o problema pede é

$P(3 \leqslant z)$

Ora,

$P(0 \leqslant z \leqslant 3) = 0{,}4987$
$P(0 \leqslant z < +\infty) = 0{,}5000$ } Então, subtraindo a menor probabilidade da maior, vem: 0,0013

b) Para que z pertença à região hachurada, basta que, em cada metade da curva, ocorra o seguinte:

$P(0 \leqslant z \leqslant 1) = 0{,}3413$ ← Metade à direita.
$P(-1 \leqslant z \leqslant 0) = 0{,}3413$ ← Metade à esquerda.

Logo, somando:

$P(-1 \leqslant z \leqslant 1) = 0{,}6826$

c) $P(-2 \leqslant z \leqslant 0) = 0{,}4772$
$P(0 \leqslant z \leqslant 1) = 0{,}3413$

$P(-2 \leqslant z \leqslant 1) = 0{,}4772 + 0{,}3413 = 0{,}8185$
↑
Isto eqüivale a $P[(-2 \leqslant z \leqslant 0)$ ou $(0 \leqslant z \leqslant 1)]$.

d) O problema pede $P(1 \leqslant z \leqslant 2)$. E essa probabilidade pode-se obter a partir de

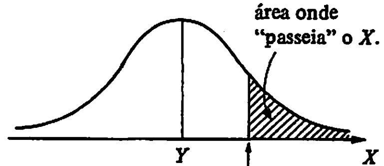
$$P(0 \leqslant z \leqslant 2) - P(0 \leqslant z \leqslant 1) = 0{,}4772 - 0{,}3413 = 0{,}1359$$

e) O problema pede $P(-\infty < z \leqslant -1)$.
Ora, $P(-\infty < z \leqslant 0) = 0{,}5000$. Logo

$$P(-\infty < z \leqslant -1) = P(-\infty < z \leqslant 0) - P(-1 \leqslant z \leqslant 0) = \\ = 0{,}5000 - 0{,}3413 = 0{,}1587$$

2. Sabe-se que X tem distribuição normal com média Y e variância 9. Se $P(X \geqslant 28) = 0{,}1587$, qual o valor de Y?

Solução



$$X \rightarrow N(\underbrace{Y}_{\text{Média}};\ \underbrace{9}_{\text{Variância}})$$

$$z = \frac{X - \bar{X}}{\sigma} = \frac{28 - Y}{\sqrt{9}} = \frac{28 - Y}{3}$$

Ora, $P(Y \leqslant X < +\infty) = 0{,}5000$
$P(28 \leqslant X) = 0{,}1587$

Logo: $P(Y \leqslant X \leqslant 28) = 0{,}3413$

Mas, então: $P(Y \leqslant X \leqslant 28) = 0{,}3413 = P(0 \leqslant z \leqslant 1)$.
Conseqüentemente, $z = 1$.

Então: $1 = \dfrac{28 - Y}{3}$, donde

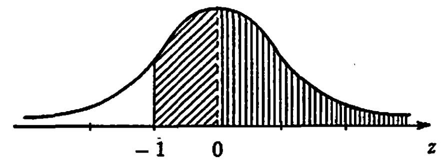
$$28 - Y = 3$$
$$-Y = -28 + 3$$
$$-Y = -25$$
$$Y = 25$$

Vem agora a pergunta: z não poderia ser -1?

Vejamos:

$$-1 = \frac{28 - Y}{3}$$
$$28 - Y = -3$$
$$-Y = -3 - 28$$
$$-Y = -31$$
$$Y = 31$$

Então: se $X \rightarrow N(31; 9)$, $P(X \geqslant 28) = P(z \geqslant -1)$.
Graficamente:



Ora, $P(z \geqslant -1)$ = área hachurada, isto é:

$$P(-1 \leqslant z \leqslant 0) + P(0 \leqslant z < +\infty) = 0{,}3413 + 0{,}5000 = 0{,}8413$$

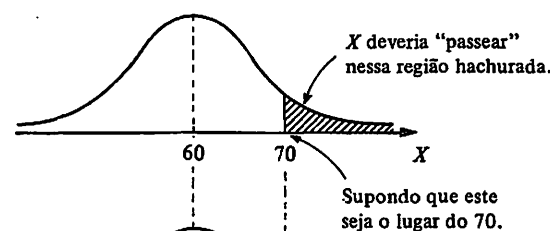
(Valor *rejeitado*, pois não coincide com 0,1587 – dado pelo problema.)

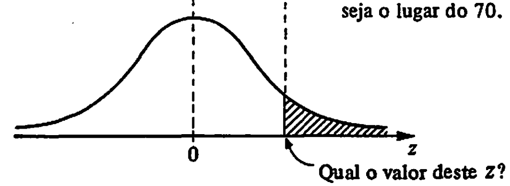
Conclusão: z não pode ser -1. A única resposta é, pois, $Y = 25$.

3. Sabe-se que X tem distribuição normal com média 60 e variância M. Sabe-se também que $P(X \geqslant 70) = 0{,}0475$. Qual o valor de M? Resposta arredondada para o inteiro mais próximo.

Solução

Um gráfico pode ajudar a pensar:





Ora, se a probabilidade correspondente à metade (direita) da curva é 0,5000, é evidente que a probabilidade associada à região *não hachurada* é

$$\begin{array}{r} 0{,}5000 \\ -0{,}0475 \\ \hline 0{,}4525 \end{array}$$

Localizando a probabilidade 0,4525 na Tabela B, observamos que o z correspondente é 1,67.

Então: $z = \dfrac{X - \overline{X}}{\sigma}$ (ver pág. 90)

$$1{,}67 = \frac{70 - 60}{\sigma}$$

$$70 - 60 = 1{,}67\,\sigma$$

$$\frac{10}{1{,}67} = \sigma \therefore \sigma \cong 5{,}988$$

Ora, se M é a variância, $M = \sigma^2 = (5{,}988)^2 = 35{,}8 \cong 36$

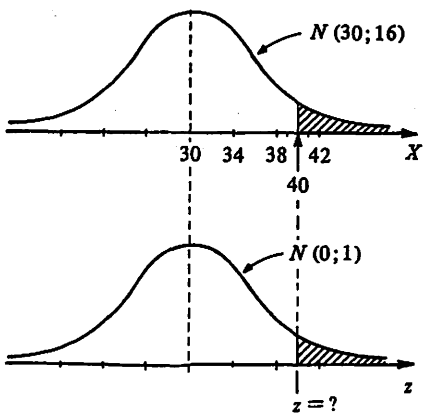
4. X tem distribuição normal com os seguintes parâmetros:

média aritmética = 30
variância = 16

Qual a probabilidade de $(X \geqslant 40)$?

Solução

Se $X \rightarrow N\,(30;\,16)$ e o problema pede $P(X \geqslant 40)$, a primeira coisa que devemos fazer é construir dois gráficos, uma para X e outro para z, com o objetivo de visualizar o que pede o problema.



Então, procurar $P\,(X \geqslant 40)$ é o mesmo que procurar $P\,(z \geqslant 2{,}5)$. Por que $z = 2{,}5$? Pelo seguinte motivo:

$z = \dfrac{X - \overline{X}}{\sigma}$ (ver pág. 90), embora, a rigor, a fórmula devesse ser $z = \dfrac{X_i - \mu}{\sigma}$.

Então, substituindo, vem:

$$z = \frac{40 - 30}{\sqrt{16}} = \frac{10}{4} = 2{,}5$$

Recorrendo à Tabela B, temos:

$$\begin{aligned} P(0 \leqslant z < +\infty) &= 0{,}5000 = p_1 \\ P(0 \leqslant z \leqslant 2{,}5) &= \underline{0{,}4938} = p_2 \\ p_1 - p_2 &= 0{,}0062 \end{aligned}$$

† DISTRIBUIÇÃO BINOMIAL

Vamos examinar o seguinte problema: quais os resultados possíveis quando se joga 1 moeda "honesta" (= eqüiprovável) 2 vezes? Ora, do ponto de vista estritamente lógico, jogar 1 moeda honesta 2 vezes é o mesmo que jogar 2 moedas honestas *1 vez*. Então:

Resultados da Moeda 1 (M_1) \ *Resultados da Moeda 2 (M_2)*	C	K
C	CC	CK
K	KC	KK

C = Cara
K = Coroa

Aqui estão os resultados procurados.

Chamando de X o número de vezes em que ocorre um particular evento – por exemplo, sair "cara" – os resultados acima podem ser escritos de maneira mais sintética. Assim:

Valores de X: (em "caras")	0	1	2
Probabilidades associadas a cada um desses X:	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{1}{4}$

↑ Zero "caras" é o mesmo que duas "coroas".

↑ Note-se que aqui a ordem dos elementos não tem a menor importância.

Esses resultados podem ser resumidos ainda mais:

X: 0 , 1 , 2

$$P(X=i): \frac{1}{4}, \quad \frac{2}{4}, \quad \frac{1}{4}, \quad \text{onde } i=0,\ 1,\ 2$$

Para melhor visualização, esses resultados podem ser dispostos em colunas, o que acaba lembrando um histograma*:

	CK		
CC	KC	KK	
↑	↑	↑	X (Número
(X = 2)	(X = 1)	(X = 0)	de "caras")

Vamos agora examinar os resultados possíveis quando se joga 1 moeda honesta 3 vezes. Procedendo de modo análogo ao adotado no problema anterior, resulta o seguinte:

colunas

Moeda 3 (M_3) \ *Moedas 1 e 2* (M_1 e M_2)	CC	CK	KC	KK
C	CCC	CCK	CKC	CKK
K	KCC	KCK	KKC	KKK

linhas

Observação: na composição do quadro, escrever *linha* em primeiro lugar e *coluna* em segundo.

* *Não* se vá pensar que o gráfico de uma distribuição binomial seja um *histograma*. Como a variável X na distribuição binomial é *discreta*, o gráfico constitui-se de *pontos* apenas.

Associando-se probabilidades aos resultados obtidos, vem que:

X:	0	1	2	3
$P(X = i)$:	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$
	↑	↑	↑	↑
	$P(X = 3)$	$P(X = 2)$	$P(X = 1)$	$P(X = 0)$

Graficamente:

	KCC	KKC		
	CKC	KCK		
CCC	CCK	CKK	KKK	
↑	↑	↑	↑	X (Número
(X = 3)	(X = 2)	(X = 1)	(X=0)	de "caras")

Examinemos agora os resultados possíveis produzidos pelo lançamento de 1 moeda honesta 4 vezes. Lembrar que 3 moedas (resultados anteriores) passarão juntas pela mesma entrada, ficando a segunda entrada reservada para a quarta moeda. Assim:

M_4 \ $M_1 M_2 M_3$	CCC	CCK	CKC	CKK	KCC	KCK	KKC	KKK
C	CCCC	CCCK	CCKC	CCKK	CKCC	CKCK	CKKC	CKKK
K	KCCC	KCCK	KCKC	KCKK	KKCC	KKCK	KKKC	KKKK

Associando-se probabilidades, vem que:

X:	0	1	2	3	4
$P(X = i)$:	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{1}{16}$

Graficamente:

		KKCC			
		KCKC			
	KCCC	KCCK	KKKC		
	CKCC	CKKC	KKCK		
	CCKC	CKCK	KCKK		
CCCC	CCCK	CCKK	CKKK	KKKK	
↑	↑	↑	↑	↑	X (Número
(X = 4)	(X = 3)	(X = 2)	(X = 1)	(X = 0)	de "caras")

Nota: Observar que à medida que o número de moedas cresce a disposição gráfica vai-se aproximando da curva normal. Quando $n \geqslant 30$, isto é, quando o número de repetições é igual ou superior a 30, praticamente se confunde a binomial com a normal.

Esses exercícios que fizemos até aqui pressupõem o seguinte:

1. os valores de X são sempre números inteiros;
2. os eventos são *independentes* e *mutuamente excludentes*;
3. em todas as fases do processo a idéia de *reposição* está presente;
4. não se leva em conta a *ordem* de ocorrência dos elementos.

Dois eventos dizem-se *independentes* quando a ocorrência de um não influi na ocorrência do outro. Por *mutuamente excludentes* entenda-se que se ocorrer um deles o outro *não* ocorrerá.

Reposição diz respeito a amostragem. Há reposição sempre que o elemento sorteado é *recolocado* na amostra *antes* do sorteio seguinte. Como é fácil perceber, esse procedimento visa a garantir que em todo o processo as probabilidades (associadas aos elementos que dele participam) mantenham-se *constantes*.

Observe-se que se a ordem dos elementos não influi no resultado final, CK = KC, CCK = KCC = CKC etc. Introduzindo-se a notação de potência, os resultados anteriores podem ser escritos assim:

$$n = 2 \longrightarrow C^2 + 2CK + K^2$$
$$n = 3 \longrightarrow C^3 + 3C^2K + 3CK^2 + K^3$$
$$n = 4 \longrightarrow C^4 + 4C^3K + 6C^2K^2 + 4CK^3 + K^4$$

Observando-se esses polinômios verifica-se que eles se originam do desenvolvimento de $(a + b)^n$, que é o Binômio de Newton, donde a distribuição chamar-se "binomial".

De fato,

$$(C + K)^2 = C^2 + 2CK + K^2$$
$$(C + K)^3 = C^3 + 3C^2K + 3CK^2 + K^3$$
$$(C + K)^4 = C^4 + 4C^3K + 6C^2K^2 + 4CK^3 + K^4$$

Num desenvolvimento binomial, sugere-se que fique em primeiro lugar o evento cuja ocorrência esteja associada à idéia de *fracasso* e, em segundo lugar, o evento cuja ocorrência esteja associada à idéia de *sucesso*. Nunca será demais recordar que, no campo das probabilidades, *sucesso* diz respeito à ocorrência do particular evento que esteja sendo estudado; por exclusão, *fracasso* designará a ocorrência de qualquer evento que não o estudado. Como exemplo, calculemos a probabilidade de saírem 2 "caras", em qualquer ordem, no lançamento de 1 moeda honesta 5 vezes. Então, de acordo com o enunciado do problema, a variável de interesse, X, à qual está associada a idéia de *sucesso*, é: *número de ocorrências de "cara"*. Montando o binômio, resulta que

$$[P(F) + P(S)]^n = (q + p)^n$$

onde

$P(F)$	abrevia a expressão "probabilidade do evento-fracasso";
$P(S)$	abrevia a expressão "probabilidade do evento-sucesso";
q	representa a probabilidade associada a $P(F)$;
p	representa a probabilidade associada a $P(S)$;
n	representa o número de repetições (ou número de moedas, ou tamanho da amostra).

Como no caso em pauta $n = 5$, vem que:

$$(q+p)^5 = \underset{\substack{\uparrow \\ P(X=0)}}{q^5} + \underset{\substack{\uparrow \\ P(X=1)}}{5q^4p} + \underset{\substack{\uparrow \\ P(X=2)}}{10q^3p^2} + \underset{\substack{\uparrow \\ P(X=3)}}{10q^2p^3} + \underset{\substack{\uparrow \\ P(X=4)}}{5qp^4} + \underset{\substack{\uparrow \\ P(X=5)}}{p^5}$$

Notar que os valores de X coincidem com os expoentes de p.

O problema pede $P(X = 2$ caras). Por isso, o termo do desenvolvimento binomial que leva à solução é $10q^3p^2$. Ora, se a moeda for honesta, $p = q = 0{,}5$. Daí decorre que:

$$10\,q^3p^2 = 10\,(0{,}5)^3\,(0{,}5)^2 = 10\,(0{,}5)^5 = 10\,(0{,}03125) = 0{,}3125.$$

Vamos aproveitar esse mesmo desenvolvimento binomial e responder às seguintes perguntas:

a) Qual a probabilidade de $X < 3$?
b) Qual a probabilidade de $X \geqslant 3$?
c) Qual a probabilidade de X ser, no máximo, igual a 1?
d) Qual a probabilidade de $X = 4$?

Soluções

a) $P(X < 3)$ é o mesmo que $P(X = 0)$ ou $P(X = 1)$ ou $P(X = 2)$. Como a expressão "ou" equivale a uma *soma*, vem que:

$$P(X < 3) = P(X=0) + P(X = 1) + P(X = 2) =$$
$$= q^5 + 5q^4p + 10q^3p^2 =$$
$$= (0{,}5)^5 + 5\,(0{,}5)^4(0{,}5) + 10\,(0{,}5)^3(0{,}5)^2 =$$
$$= 0{,}03125 + 0{,}15625 + 0{,}3125 = 0{,}5$$

Esse resultado poderia ter sido obtido por simples inspeção do desenvolvimento de $(q + p)^n$. Como não existe probabilidade maior que 1, $(q + p) = 1$. Como $q = p = 0{,}5$ no caso de uma moeda honesta, os três primeiros termos correspondem à *metade* do desenvolvimento binomial – e a metade de 1 é 0,5.

b) $P(X \geqslant 3) = P(X = 3) + P(X = 4) + P(X = 5)$
Por igual raciocínio, $P(X \geqslant 3) = 0{,}5$.

c) $P(X = 1, \text{no máximo}) = P(X=0) + P(X=1) = q^5 + 5q^4p =$

$= (0{,}5)^5 + 5\,(0{,}5)^4(0{,}5) = 0{,}03125 + 0{,}15625 =$

$= 0{,}1875$

d) $P(X=4) = 5qp^4 = 5\,(0{,}5)\,(0{,}5)^4 = 5\,(0{,}5)^5 =$

$= 5\,(0{,}03125) = 0{,}15625$

Note-se que nos casos em que $p = q$, o desenvolvimento binomial é *simétrico*. Portanto, no presente exercício, $P(X = 4) = P(X = 1)$. Conseqüentemente, o resultado acima poderia ter sido obtido a partir de $P(X = 1) = 0{,}15625$, valor esse encontrado na solução do item anterior (c).

Termo Geral

Terão notado os leitores que à medida que n cresce aumenta também a dificuldade de expansão de $(q + p)^n$. Imagine-se, a título de exemplo, a complicação que resultaria do desenvolvimento de $(q + p)^{10}$ para, a seguir, calcular-se $P(X = 8)$ com $p = 0{,}2$.

Essa dificuldade pode ser contornada com a aplicação da fórmula do termo geral do desenvolvimento binomial, que é:

$$T_{k+1} = C_n^k q^{n-k} p^k \text{*}$$

Para a aplicação dessa fórmula é importante lembrar do seguinte:

T_{k+1} é o termo de ordem $k + 1$, ou seja, existem k termos que o precedem.

* Neste ponto conviria que o leitor recapitulasse algumas noções básicas de análise combinatória. (Veja, por exemplo, Hariki, Seiji e Onaga, Dulce; *Curso de Matemática*; editora HARBRA ltda., SP, 1980, volume 2, capítulo 8.)

$$C_n^k = \binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

Então, para calcular $P(X = 8)$ basta fazer $k = 8$ e entrar na fórmula:

$$T_9 = C_{10}^8 q^{10-8} p^8$$

$$T_9 = \frac{10!}{8!\,(10-8)!}(0{,}8)^2(0{,}2)^8 = 45\,(0{,}64)\,(0{,}0000025) =$$

$= 0{,}000072$ ou 0^+, conforme a Tabela K (pág. 369).

Observação: Se $p = 0{,}2$ (de acordo com o problema), $q = 0{,}8$, pois $(q + p) = 1$.

Tabelas de $B(n\,;p)$*

Como já se mencionou, o crescimento de n acarreta um aumento progressivo nos trabalhos de cálculo. Por isso, existem tabelas que fornecem diretamente as probabilidades, bastando localizar n, X e p. Uma dessas tabelas, reproduzida no fim deste livro, cobre várias possibilidades para $n \leqslant 31$.

A título de exemplificar seu uso, reproduzimos na página seguinte um segmento dessa tabela para $n = 4$ e $p = 0{,}5$.

Probabilidades Binomiais Simples

$p \rightarrow$	0,05	0,10	0,20	0,25	0,30	0,40	0,50	$n = 4$
$X = 0$	815	656	410	316	240	130	062	4
1	171	292	410	422	412	346	250	3
2	014	049	154	211	265	346	375	2
3	0^+	004	026	047	076	154	250	1
4	0^+	0^+	002	004	008	026	062	$0 = X$
$n = 4$	0,95	0,90	0,80	0,75	0,70	0,60	0,50	$\leftarrow p$

Esta tabela oferece probabilidades *isoladas*. Por exemplo: no lançamento de 1 moeda honesta 4 vezes, qual a probabilidade de saírem 3 coroas?

$$P(X = 3) = 0{,}250$$

Notar que na tabela não figura a parte 0, ..., a fim de facilitar-se a leitura.

* A binomial tem dois *parâmetros fundamentais*, isto é, dois valores a partir dos quais é possível montar a distribuição: n e p. A notação $B(n\,;p)$ significa: binomial com n repetições (independentes) e probabilidade de sucesso igual a p.

EXERCÍCIOS RESOLVIDOS

1. No lançamento de 1 moeda honesta 18 vezes, qual a probabilidade de saírem 12 "caras"?

Solução

Aqui $X = 12$ "caras", $n = 18$ e $p = 0{,}5$. A Tabela K (pág. 371), reproduzida a seguir, apresenta os seguinte valores:

$p = 0{,}5$

$p \rightarrow$	0,05	0,10	0,20	0,25	0,30	0,40	0,50	$n = 18$
$X = 0$	397	150	018	006	002	0^+	0^+	18
1	376	300	081	034	013	001	0^+	17
2	168	284	172	096	046	007	001	16
3	047	168	230	170	105	025	003	15
4	009	070	215	213	168	061	012	14
5	001	022	151	199	202	115	033	13
6	0^+	005	082	144	187	166	071	12
7	0^+	001	035	082	138	189	121	11
8	0^+	0^+	012	038	081	173	167	10
9	0^+	0^+	003	014	039	128	185	9
10	0	0^+	001	004	015	077	167	8
11	0^+	0^+	0^+	001	005	037	121	7
12	0^+	0^+	0^+	0^+	001	015	071	6
13	0^+	0^+	0^+	0^+	0^+	004	033	5
14	0^+	0^+	0^+	0^+	0^+	001	012	4
15	0^+	0^+	0^+	0^+	0^+	0^+	003	3
16	0^+	0^+	0^+	0^+	0^+	0^+	001	2
17	0^+	0^+	0^+	0^+	0^+	0^+	0^+	1
18	0^+	0^+	0^+	0^+	0^+	0^+	0^+	$0 = X$
$n = 18$	0,95	0,90	0,80	0,75	0,70	0,60	0,50	$\leftarrow p$

$n = 18$

$X = 12$

$P(X = 12) = 0{,}071$

Como saber se esse resultado está correto? É só aplicar

$$T_{k+1} = C_n^k \cdot q^{n-k} \cdot p^k$$

Fazendo $k = 12$, vem que:

$$T_{12+1} = C_{18}^{12} \cdot (0{,}5)^{18-12} \cdot (0{,}5)^{12}$$

$$T_{13} = \frac{18!}{12!\,(18-12)!} \cdot (0{,}5)^6 \cdot (0{,}5)^{12} = 18.564\,(0{,}0000038) =$$

$$= 0{,}0705432 \cong 0{,}071$$

2. No lançamento de 1 moeda honesta 4 vezes, qual a probabilidade de $(X \leqslant 3)$, sendo X o número de coroas?

Solução

Neste problema,

$$(X \leqslant 3) = (X = 3) \text{ ou } (X = 2) \text{ ou } (X = 1) \text{ ou } (X = 0)$$

$$p = 0{,}5$$
$$n = 4$$

Então

$$P(X \leqslant 3) = \underset{0{,}250}{P(X = 3)} + \underset{0{,}375}{P(X = 2)} + \underset{0{,}250}{P(X = 1)} + \underset{0{,}062}{P(X = 0)}$$

Probabilidades tiradas diretamente da Tabela K (pág. 367)

$$\therefore P(X \leqslant 3) = 0{,}250 + 0{,}375 + 0{,}250 + 0{,}062 = 0{,}937$$

Média Aritmética

A média aritmética de uma distribuição binomial é dada por

$$\mu = np$$

onde

- μ representa a média procurada (populacional);
- n representa o número de repetições do experimento;
- p representa a probabilidade associada ao evento-sucesso.

Exemplo

Calcular a média aritmética de $B\,(6; 0{,}5)$.

Solução

Entrando na fórmula com os dados do problema resulta que $\mu = 6\,(0{,}5) = 3{,}0$.

Variância

A variância de uma distribuição binomial é dada por

$$\sigma^2 = npq$$

onde

- σ^2 representa a variância procurada (populacional);
- n representa o número de repetições do experimento;
- p representa a probabilidade associada ao evento-sucesso;
- q representa a probabilidade associada ao evento-fracasso.

Exemplo

Calcular a variância de $B\,(6; 0{,}5)$.

Solução

Entrando na fórmula com os dados do problema resulta que

$$\sigma^2 = 6\,(0{,}5)\,(0{,}5) = 6\,(0{,}25) = 1{,}50$$

Aproximação da Binomial pela Normal

Quando np ou nq — sempre o *menor* — for $\geqslant 5$, a *normal* constituirá uma boa aproximação para a *binomial.* Essa propriedade é particularmente interessante quando se leva em conta que a partir de $n > 31$ poucas são as tabelas que registram as probabilidades associadas aos valores de X.

A fórmula resolutiva da binomial pela normal é:

$$z_i = \frac{X_i \pm 0{,}5 - np}{\sqrt{npq}}$$

Exemplo

Uma moeda honesta é lançada 20 vezes. Sendo X o número de "caras", determinar $P\,(12 \leqslant X \leqslant 14)$.

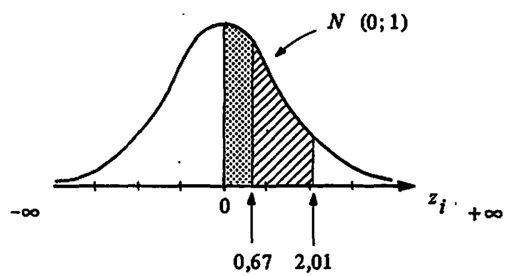
Solução

A variável X na distribuição binomial é *descontínua.* Para que seja possível trabalhar com a normal, é preciso transformar X (descontínua) em z (contínua). Assim:

$$z_1 = \frac{12 - 0{,}5 - 20\,(0{,}5)}{\sqrt{20\,(0{,}5)\,(0{,}5)}} = \frac{11{,}5 - 10}{\sqrt{5}} = \frac{1{,}5}{2{,}2361} = 0{,}67$$

$$z_2 = \frac{14 + 0{,}5 - 20\,(0{,}5)}{\sqrt{20\,(05)\,(0{,}5)}} = \frac{14{,}5 - 10}{\sqrt{5}} = \frac{4{,}5}{2{,}2361} = 2{,}01$$

Transpondo z_1 e z_2 para a curva normal reduzida (Tabela B), resulta o seguinte:



Então, de acordo com o problema, $P\,(12 \leqslant X \leqslant 14) = P\,(0{,}67 \leqslant z \leqslant 2{,}01)$. Portanto, a resposta procurada corresponde à *diferença* que resulta do seguinte:

a (z_2 = 2,01) corresponde a (área pontilhada) + (área hachurada) = 0,4778
a (z_1 = 0,67) corresponde a (área pontilhada) = 0,2486

$$\begin{pmatrix}\text{Área}\\\text{pontilhada}\end{pmatrix} + \begin{pmatrix}\text{Área}\\\text{hachurada}\end{pmatrix} - \begin{pmatrix}\text{Área}\\\text{pontilhada}\end{pmatrix} = \text{Diferença} = 0{,}2292$$

Resta agora verificar se a resposta encontrada é razoável, em termos de aproximação, quando comparada à que obteríamos se a resolução fosse pela binomial.

Consultando a Tabela K (pág. 370) — *Distribuição Binomial* — temos:

$$\begin{aligned} &P\,(X = 12) = 0{,}120\\ +\ &P\,(X = 13) = 0{,}074\\ &P\,(X = 14) = 0{,}037 \end{aligned}$$

$$P\,(X = 12) + P\,(X = 13) + P\,(X = 14) = 0{,}120 + 0{,}074 + 0{,}037 = 0{,}231$$

Comparando agora 0,231 (solução binomial) com 0,2292 (solução normal), verificamos a existência de um erro desprezível da ordem de 0,0018. Entenda-se: para efeitos *práticos*, essa aproximação é muito boa; entretanto, do ponto de vista *teórico*, a solução binomial é *preferível*, uma vez que os resultados binomiais são exatos.

EXERCÍCIOS RESOLVIDOS

1. Um casal deseja 8 crianças. Chamando de X o número de homens (filhos), qual a probabilidade de que

a) $X = 2$?
b) $2 < X < 4$?
c) $2 \leqslant X \leqslant 5$?
d) só nasçam mulheres?
e) nasçam, no mínimo, 4 homens?

*Solução**

Trata-se de uma binomial do tipo $B\,(8; \frac{1}{2})$, onde $p = 0{,}5$.

a) $P\,(X = 2) = 0{,}109$ (Basta consultar a Tabela K (pág. 369) com $(n = 8)$, $(p = 0{,}50)$ e $(X = 2)$.)

O exercício pode também ser feito de forma exaustiva, a partir do desenvolvimento de $(q + p)^8$. Assim:

$$(q+p)^8 = \underset{P(X=0)}{q^8} + \underset{P(X=1)}{8q^7p} + \underset{P(X=2)}{28q^6p^2} + \underset{P(X=3)}{56q^5p^3} + \underset{P(X=4)}{70q^4p^4} + \underset{P(X=5)}{56q^3p^5} +$$

$$+ \underset{P(X=6)}{28q^2p^6} + \underset{P(X=7)}{8qp^7} + \underset{P(X=8)}{p^8}$$

O termo que dá a $P\,(X = 2)$ é $28q^6p^2$. Substituindo então $p = q = 0{,}5$ e efetuando as operações, vem:

$$28\,(0{,}5)^6(0{,}5)^2 = 28\,(0{,}5)^{6+2} = 28\,(0{,}5)^8 = 28\,(0{,}0039062) =$$

$$= 0{,}1093736 \text{ ou } 0{,}109$$

b) $P\,(2 < X < 4) = P\,(X = 3) = 0{,}219$ (Veja a Tabela K (pág. 369)

c) $P\,(2 \leqslant X \leqslant 5) = P\,(X = 2) + P\,(X = 3) + P\,(X = 4) + P\,(X = 5)$

$$\underset{0{,}109}{\uparrow} + \underset{0{,}219}{\uparrow} + \underset{0{,}273}{\uparrow} + \underset{0{,}219}{\uparrow} = 0{,}820$$

d) P (só nasçam mulheres) $= P\,(X = 0) = q^8 = 0{,}004$

e) P (nasçam, no mínimo, 4 homens) =

* Admitir que $P(♂) = P(♀) = 0{,}5$. Admitir também que haja *independência* entre os nascimentos.

$$= P\,(X = 4) + P\,(X = 5) + P\,(X = 6) + P\,(X = 7) + P\,(X = 8)$$

$$\underset{0{,}273}{\uparrow} + \underset{0{,}219}{\uparrow} + \underset{0{,}109}{\uparrow} + \underset{0{,}031}{\uparrow} + \underset{0{,}004}{\uparrow} = 0{,}636$$

2. Seja a variável X distribuída binomialmente. Se $p = 0{,}25$ e $n = 30$,

a) quanto vale a média aritmética?
b) quanto vale a variância?
c) qual a probabilidade de $X = 10$?

Soluções

a) $\mu = np$
$\mu = 30\,(0{,}25) = 7{,}50$

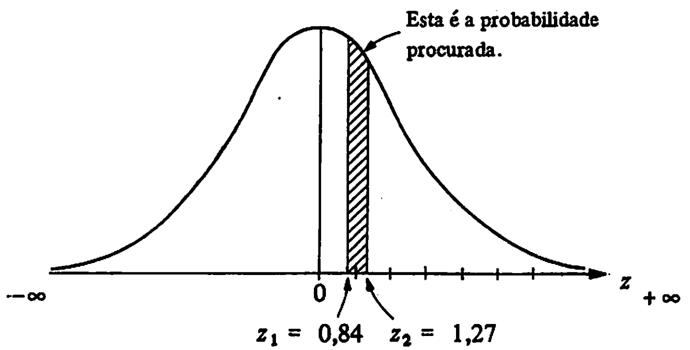
b) $\sigma^2 = npq$
$\sigma^2 = 30\,(0{,}25)\,(0{,}75) = 5{,}6250$

c) $P\,(X = 10) = P\,(z_1 \leqslant z_i \leqslant z_2)$

$$z_i = \frac{(X_i \pm 0{,}5) - np}{\sqrt{npq}}$$

$$z_1 = \frac{(10 - 0{,}5) - 7{,}5}{\sqrt{30\,(0{,}25)\,(0{,}75)}} = \frac{9{,}5 - 7{,}5}{2{,}37} = \frac{2}{2{,}37} = 0{,}84$$

$$z_2 = \frac{(10 + 0{,}5) - 7{,}5}{\sqrt{30\,(0{,}25)\,(0{,}75)}} = \frac{10{,}5 - 7{,}5}{2{,}37} = \frac{3}{2{,}37} = 1{,}27$$



A probabilidade associada a $z_2 = 1{,}27$ é $0{,}3980 = p_2$
A probabilidade associada a $z_1 = 0{,}84$ é $0{,}2995 = p_1$
A probabilidade da área assinalada $= p_2 - p_1 = 0{,}0985$

RESUMO

Procurou-se, neste capítulo, relacionar as propriedades da distribuição normal teórica com problemas do "mundo real", que apresentam interesse para a pesquisa. Demonstrou-se, assim, que a área sob a curva normal pode ser usada tanto na interpretação do desvio padrão quanto na elaboração de afirmações que envolvem a noção de probabilidade. A importância da distribuição normal tornar-se-á mais visível em capítulos subseqüentes deste livro.

PROBLEMAS

1. Em qualquer distribuição normal, qual a porcentagem da área total que cai (a) entre −1 DP e +1 DP; (b) entre −2 DPs e +2 DPs; (c) entre −3 DPs e +3 DPs?
2. Dado um conjunto de escores brutos com distribuição normal, onde $\bar{X}$ = 7,5 e DP = 1,3, exprimir cada um dos seguintes escores brutos em termos de escala z: (a) $X = 8{,}0$; (b) $X = 6{,}0$; (c) $X = 5{,}3$; (d) $X = 10{,}2$; (e) $X = 7{,}5$; (f) $X = 8{,}5$; (g) $X = 11$.
3. Seja uma distribuição de ganhos diários, onde $\bar{X}$ = Cr$ 178,50 e DP = Cr$ 47,60. Supondo-se distribuição normal, como podem ser expressas, em escala z, as seguintes diárias: (a) X = Cr$ 142,80; (b) X = Cr$ 187,00; (c) X = Cr$ 224,40; (d) X = Cr$ 85,00; (e) X = Cr$ 255,00; (f) X = Cr$ 195,50; (g) X = Cr$ 153,00.
4. Mesmos dados básicos do problema 3, acima. Determinar: (a) a porcentagem de respondentes que ganham uma diária de Cr$ 255,00 ou mais; (b) a probabilidade de sortear um respondente cuja diária seja de Cr$ 255,00 ou mais; (c) a porcentagem de respondentes que ganham de Cr$ 170,00 a Cr$ 178,50; (d) a probabilidade de sortear um respondente cuja diária caia entre Cr$ 170,00 e Cr$ 178,50; (e) a probabilidade de sortear um respondente cuja diária seja de Cr$ 170,00 ou menos; (f) a probabilidade de sortear um respondente cuja diária seja de "Cr$ 170,00 ou menos" *ou* de "Cr$ 187,00 ou mais"; (g) a probabilidade de sortear dois respondentes cujas diárias sejam de Cr$ 170,00 ou menos.
5. Dado um conjunto de escores brutos, distribuídos normalmente, onde $\bar{X}$ = 80 e DP = 7,5, determinar (a) a porcentagem de respondentes que obtiveram 60 ou menos; (b) a probabilidade de sortear um respondente que tenha obtido 60 ou menos; (c) a porcentagem de respondentes que obtiveram entre 80 e 90; (d) a probabilidade de sortear um respondente que tenha obtido entre 80 e 90; (e) a porcentagem de respondentes que obtiveram 85 ou mais; (f) a probabilidade de sortear um respondente que tenha obtido 85 ou mais; (g) a probabilidade de sortear um respondente que tenha tirado "70 ou menos" *ou* "90 ou mais"; (h) a probabilidade de sortear três respondentes que tenham tirado 90 ou mais.

† EXERCÍCIOS PROPOSTOS

1. X é $N(20; 49)$.
   Calcular $P(X < 30)$.

2. X é $N(10; 100)$.
   Calcular $P(12 \leqslant X \leqslant 20)$.
3. X é $N(30; 16)$.
   Calcular $P(19 \geqslant X)$.
4. X é $N(20; 25)$.
   Calcular $P(30 \geqslant X)$.
5. X é $N(50; 81)$.
   Calcular $P(40 \leqslant X \leqslant 60)$.
6. X é $N(10; 16)$.
   Calcular $P(X \geqslant 5)$.
7. X é $N(10; 25)$.
   Calcular $P(X < 5)$.
8. X é $N(10; 25)$.
   Calcular $P(5 \leqslant X \leqslant 15)$.
9. Uma moeda eqüiprovável vai ser jogada 18 vezes. Calcular, *sem* recorrer às tabelas, o seguinte:
   a) $P(X = 12)$
   b) $P(8 < X < 10)$ Obs.: X = número de caras (C).
   c) $P(X > 15)$
   d) $P(3 \leqslant X < 8)$
   Respostas com 5 casas decimais.
10. Um dado eqüiprovável (6 faces) vai ser jogado 9 vezes. Calcular, *sem* recorrer às tabelas, o seguinte:
    a) $P(X < 3)$ Obs.: X = número de ocorrências do tipo "face 3".
    b) $P(2 < X \leqslant 6)$
    Respostas com 5 casas decimais.
11. Quanto vale o 7º termo do desenvolvimento de $(q + p)^{13}$?
12. Se $q = 0{,}28$, quanto vale a probabilidade expressa pelo 7º termo do desenvolvimento de $(q + p)^{13}$? Resposta com 3 casas decimais.
13. Uma prova tem 40 questões e cada questão, 4 alternativas — sendo uma só correta. Se os alunos resolverem "chutar" as respostas, qual o número mais provável de acertos?

14. Uma moeda honesta foi lançada 100 vezes. Qual a probabilidade de terem saído exatamente 64 caras? Resposta com 4 casas decimais.

15. Uma moeda honesta foi lançada 100 vezes. Qual a probabilidade de terem saído 80 caras ou mais?

16. Uma moeda honesta foi lançada 60 vezes. Se X = número de coroas, qual a probabilidade de $(20 \leqslant X \leqslant 40)$? Resposta com 4 casas decimais.

17. Sabe-se que a variável X tem distribuição binomial com os parâmetros $n = 40$ e $p = 0,6$. Calcular a média aritmética (μ) e a variância (σ^2) dessa distribuição.

18. Com os dados do problema 9, calcule $P(X \leqslant 16)$. Resposta com 3 casas decimais.

7

Amostras e Populações

O pesquisador social procura geralmente tirar conclusões a respeito de um grande número de sujeitos. Por exemplo, ele poderia desejar estudar os 220.000.000 de cidadãos que constituem a população dos Estados Unidos da América do Norte, os 1.000 membros de um sindicato particular, os 10.000 norte-americanos negros que moram numa cidade do sul ou os 45.000 estudantes habitualmente matriculados numa universidade do leste.

Até aqui, temos admitido que o pesquisador social investiga todo o grupo que ele tenta compreender. Conhecido por *população* ou *universo*, tal grupo consiste em um conjunto de indivíduos que partilham de, pelo menos, uma característica comum, seja ela cidadania, filiação a uma associação de voluntários, etnia, matrícula na universidade etc. Desse modo, poderíamos falar a respeito da população dos Estados Unidos, da população dos membros do sindicato, da população dos norte-americanos negros residentes numa cidade do sul ou da população de alunos universitários.

Posto que o pesquisador trabalha com tempo, energia e recursos econômicos limitados, ele raras vezes estuda individualmente todos os sujeitos da população na qual está interessado. Em lugar disso, o pesquisador estuda apenas uma *amostra*— que se constitui de um número menor de sujeitos tirados de uma determinada população. Através do processo de amostragem, o pesquisador busca generalizar (conclusões) de sua amostra (o grupo pequeno) para a população toda (o grupo maior), da qual essa mesma amostra foi extraída.

O processo de amostragem faz parte do dia-a-dia. De que outra forma conseguiríamos informação acerca de pessoas diversas, não fora amostrando as que estão à nossa volta? Por exemplo, poderíamos casualmente discutir assuntos políticos com alguns estudantes a fim de determinar a opinião geral que a classe estudantil tem com relação a esse assunto; poderíamos tentar descobrir como nossos colegas de classe estão-se preparando para um exame particular através de conversas antecipadas com apenas alguns membros da classe; poderíamos mesmo investir na bolsa depois de descobrir que uma pequena amostra de nossos sócios ganhou dinheiro dessa forma.

MÉTODOS DE AMOSTRAGEM

Os métodos de amostragem do pesquisador são, em geral, mais elaborados e sistemáticos do que aqueles utilizados no dia-a-dia. Ele preocupa-se basicamente em saber se sua amostra de sujeitos é bem representativa da população, a fim de ser-lhe possível fazer generalizações de uma para a outra. Para fazer tais inferências, o pesquisador seleciona um método apropriado de amostragem que leva em conta a possibilidade de todos os membros da população fazerem parte da amostra ou, então, de apenas alguns membros fazerem parte dela. Se todos os componentes de tal população tiverem igual oportunidade (probabilidade) de participar da amostra, diz-se que o método usado é o da amostragem *casual*; se este não for o caso, fala-se em amostragem *não-casual.*

Amostras Não-casuais

O método de amostragem não-casual mais conhecido–amostragem *acidental*–difere muito pouco dos nossos procedimentos diários de amostragem, uma vez que ele se baseia com exclusividade no que convém ao pesquisador. Em outras palavras, o pesquisador simplesmente inclui os sujeitos convenientes na amostra, dela excluindo os inconvenientes. A maioria dos estudantes recorda-se de, pelo menos, algumas ocasiões em que um instrutor, interessado numa pesquisa, pediu-lhes que participassem de um experimento ou preenchessem um questionário. A popularidade dessa forma de amostragem acidental em Psicologia fez com que alguns críticos olhassem para essa disciplina como sendo "a ciência dos calouros universitários", dada a grande freqüência com que eles figuram como sujeitos experimentais em pesquisas.

Outro tipo de amostragem não-casual é o de *quotas.* Neste procedimento, diversas características de uma população, tais como idade, sexo, classe social ou etnia são amostradas nas mesmas proporções em que figuram na população. Suponham, por exemplo, que tivéssemos de extrair uma amostra, por quota, da população de alunos de uma dada universidade, onde 42% fossem mulheres e 58%, homens. Usando esse método, os entrevistadores recebem a incumbência de localizar uma quota de estudantes de tal forma que somente 42% da amostra consista de mulheres e 58%, de homens. As mesmas porcentagens que figuram na população são reproduzidas na amostra. Se o tamanho global da amostra fosse 200, então 84 moças e 116 rapazes deveriam ser selecionados.

Uma terceira variedade de amostra não-casual é a de *julgamento* ou *conveniência.** A idéia básica aqui envolvida é a de que a lógica, o senso comum ou um julgamento equilibrado podem ser usados na seleção de uma amostra que seja representativa de um grupo maior (população). Por exemplo, para extrair uma amostra de revistas que reflitam valores da classe média norte-americana, poderíamos, levados apenas pela intuição, selecionar *Reader's Digest, People* ou *Parade*, uma vez que os artigos dessas revistas *parecem* refletir aquilo que a maioria dos norte-americanos de classe média deseja (por exemplo, a satisfação do sonho americano, sucesso econômico, e coisas semelhantes). De modo similar, certos bairros que tradicionalmente elegiam o candidato oficial, poderiam ser selecionados num esforço de predizer o resultado de uma eleição oficial.

*N.T.: Uma tradução mais literal obrigaria ao uso de "proposital".

Amostras Casuais*

Como já foi mencionado anteriormente, a amostragem casual dá a cada membro da população igual oportunidade de fazer parte da amostra. Essa característica da amostragem casual implica que todos os sujeitos da população devem ser identificados *antes* da extração da amostra, exigência geralmente preenchida mediante a obtenção (elaboração) de uma lista que contenha todos os sujeitos da população. Um pouco de atenção sugere logo que o fato de conseguir um arrolamento completo de todos os membros da população nem sempre é tarefa fácil, mormente se uma população muito grande ou muito diversificada estiver sendo estudada. Para dar um exemplo simples, onde poderíamos conseguir uma lista *completa* dos alunos regularmente matriculados numa universidade grande? Aqueles pesquisadores que já tentaram fazê-lo poderão dar testemunho da dificuldade encontrada. Para exemplificar uma tarefa um pouco mais laboriosa, tente-se encontrar uma lista de todos os residentes numa cidade grande. De que modo podemos estar seguros de ter identificado todo mundo–mesmo aqueles residentes que não desejam ser identificados?

O tipo básico de amostra aleatória (casual), *amostra casual simples*, pode ser obtido por um processo que não é diferente daquela tão conhecida técnica de escrever o nome de cada um em pedacinhos separados de papel e, de olhos vendados, extrair apenas alguns nomes de um chapéu. Este procedimento dá, idealmente, a cada membro da população uma igual oportunidade de ser selecionado para a amostra, uma vez que um e somente um pedacinho de papel, *por pessoa*, é colocado no chapéu. Por várias razões (inclusive o fato de que o pesquisador necessitaria de um chapéu extremamente grande), o pesquisador interessado numa amostra casual via de regra não tira nomes de chapéus. Em vez disso, ele usa uma *tábua de números aleatórios* semelhante à Tabua H localizada no fim deste livro. Parte de uma tábua de números aleatórios foi reproduzida abaixo.

	Coluna																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Linha 1	2	3	1	5	7	5	4	8	5	9	0	1	8	3	7	2	5	9	9	3
2	6	2	4	9	7	0	8	8	6	9	5	2	3	0	3	6	7	4	4	0
3	*0*	*4*	*5*	5	5	0	*4*	*3*	*1*	0	5	3	7	4	3	5	0	8	9	0
4	1	1	8	3	7	4	4	1	0	9	6	2	2	1	3	4	3	1	4	8
5	1	6	0	3	5	0	3	2	4	0	4	3	6	2	2	2	3	5	0	0

Uma tábua de números aleatórios é construída com a finalidade de gerar uma série de números totalmente desprovida de "lei de formação". Como resultado, o processo de utilização de uma tábua de números aleatórios produz uma amostra não-viesada semelhante à produzida pela retirada de nomes, com os olhos vendados, de um chapéu.

Para extrair uma amostra casual simples por meio de uma tábua de números

* N.T.: Em alguns textos já aparece a expressão *amostra randômica* onde "randômica" é um aportuguesamento da palavra inglesa *random* (casual, aleatório).

aleatórios, o pesquisador arranja, em primeiro lugar, uma lista de todos os componentes da população, e atribui a cada um deles um (e um só) número de identificação. Por exemplo, se ele estivesse fazendo uma pesquisa que envolvesse os 500 alunos matriculados na disciplina "Introdução à Sociologia", ele poderia conseguir, com a secretaria, uma lista de todos os nomes, e atribuir a cada um, sem repetições, um número da série 001, 002, etc. até 500. Concluída a preparação da lista, ele inicia a montagem de sua amostra extraindo números de uma tabela de números aleatórios. Digamos que o pesquisador queira extrair uma amostra de 50 alunos para representar os 500 membros dessa população. Ele poderia iniciar a consulta à tabela de números aleatórios a partir de qualquer número (de olhos fechados, por exemplo) e, movendo-se em qualquer sentido, pegar números adequados até que os 50 membros fossem selecionados. Ao olhar para uma porção da tábua de números aleatórios acima, poderíamos arbitrariamente começar na intersecção da coluna 1 com a linha 3, caminhar da esquerda para a direita e tomar todos os números que surgissem entre 001 e 500. Os primeiros números que aparecem no cruzamento da coluna 1 com a linha 3 são 0, 4 e 5. Desse modo, o aluno número 045 é o primeiro membro da população a ser selecionado para a amostra. Continuando da esquerda para a direita, vemos que 4, 3 e 1 são os três números que vêm a seguir, de modo que o aluno 431 (da lista original) é selecionado. Esse processo continua até que os 50 sujeitos da amostra tenham sido escolhidos. Nota ao estudante: ao usar uma tábua de números aleatórios, não considere números que se repetem ou que sejam superiores aos necessários.

Todos os métodos de amostragem casual são, na verdade, variações do procedimento de amostragem aleatória simples que acabou de ser ilustrado. Por exemplo, no caso de amostragem *sistemática*, não há necessidade de usar uma tábua de números aleatórios, uma vez que os membros da população que participam da amostra são determinados a partir de intervalos fixos. Ao empregar amostragem sistemática, então, cada n-ésimo membro de uma população é incluído numa amostra daquela população. Ilustrando: na extração de uma amostra de tamanho 1.000 de uma população composta de 10.000 inquilinos, poderíamos, depois de proceder ao arrolamento inicial, selecionar cada décimo sujeito da lista.

A vantagem da amostragem sistemática está no fato de que não é necessário usar tábuas de números aleatórios. Como resultado, esse método leva menos tempo do que o da amostragem casual simples, especialmente quando as populações são muito grandes. Do ponto de vista das desvantagens, extrair uma amostra sistemática implica admitir que o fator "posição" na lista dos componentes da população não introduz viés nos resultados, isto é, não torna a seleção menos aleatória. Se esse aspecto não for levado a sério, na amostra final podem aparecer *mais* sujeitos de um tipo do que de outro. Isso pode ocorrer, por exemplo, quando se extrai uma amostra sistemática de uma população de casas, onde as de esquina (por serem geralmente as mais caras do quarteirão) ocupam uma posição fixa; ou, então, quando os nomes de uma lista telefônica são amostrados por intervalos fixos, pode ocorrer que certos nomes, vinculados a problemas étnicos, sejam subestimados.

Outra variação da amostra casual simples, a amostra *estratificada*, consiste em dividir a população em subgrupos mais homogêneos—estratos—dos quais amostras aleatórias, simples, são extraídas. Suponham, por exemplo, que desejemos estudar a aceitação de vários dispositivos para o controle da natalidade entre os membros de determinada cidade. Como as atitudes relativas a controle da natalidade variam com a religião e com a situação sócio-econômica, poderíamos estratificar nossa população com base nessas variáveis, e obter, assim, subgrupos cuja homogeneidade é maior. Mais especificamente, suponhamos que fosse possível identificar, na população, católicos, protestantes e judeus, bem como classe alta, classe média e classe baixa. Nosso procedimento de estratificação poderia produzir os seguintes subgrupos ou estratos:

protestantes de classe alta
protestantes de classe média
protestantes de classe baixa
católicos de classe alta
católicos de classe média
católicos de classe baixa
judeus de classe alta
judeus de classe média
judeus de classe baixa

Tendo identificado nossos estratos, começamos a colher amostras casuais simples de cada subgrupo ou estrato (por exemplo, de protestantes da classe baixa, de católicos da classe média, e assim por diante), até que toda a população tenha sido amostrada. Em outras palavras, cada estrato é tratado, para fins de coleta de amostra, como se fosse uma população completa, e o procedimento de amostragem casual simples é aplicado. Especificamente, cada membro de um estrato recebe um número de identificação, é arrolado e, a seguir, amostrado por meio de uma tábua de números aleatórios. O passo final do procedimento consiste em juntar, num só grupo, os sujeitos selecionados de cada subgrupo ou estrato, a fim de obter-se uma amostra da população toda.

A estratificação baseia-se na idéia de que um grupo homogêneo requer uma amostra menor do que um grupo heterogêneo. Por exemplo, estudar os sujeitos que passam por uma particular esquina do centro da cidade, provavelmente requer uma amostra maior do que a necessária para estudar sujeitos que vivem num subúrbio de classe média. Em geral, encontram-se, no centro da cidade, sujeitos que podem apresentar as mais variadas combinações de características. Em contraste, as pessoas que moram num subúrbio de classe média são, via de regra, mais semelhantes no que respeita a educação, renda, orientação política, tamanho da família, atitude para com o trabalho—para mencionar apenas algumas características.

Superficialmente, as amostras estratificadas casuais apresentam uma enorme semelhança com o método de quotas não-casuais, tal como já discutido, uma vez que ambos os procedimentos requerem, de hábito, a inclusão de características de amostra nas proporções exatas com que figuram na população. Portanto, se 32% de nossa população compõem-se de protestantes da classe média, então exatamente 32% da nossa amostra devem ser colhidos dentre protestantes da classe média; da mesma forma, se 11% de nossa população consiste em judeus da classe baixa, então 11% de nossa amostra devem ser constituídos de forma semelhante, e assim por diante. No terreno da amostragem estratificada, surge uma exceção quando um estrato particular está desproporcionalmente bem representado na amostra, tornando-se possível, assim, uma subanálise mais minuciosa daquele grupo. Tal situação poderia surgir, por exemplo, quando negros norte-americanos, que constituem uma pequena proporção de uma dada população, são "superamostrados" num esforço de examinar suas características de maneira mais pormenorizada.

Apesar de suas aparentes similaridades, amostras por quota e por estratos são essencialmente diferentes. Enquanto que os membros de amostras por quotas são escolhidos através do método pelo qual o pesquisador optou, os membros de amostras estratificadas são sempre selecionados de *forma casual*, de hábito por meio de uma tábua de números aleatórios aplicados à lista completa (rol) dos componentes da população.

Antes de encerrarmos o tópico de métodos de amostragem, vamos examinar a natureza de uma forma particularmente popular de amostragem casual, conhecida por método de *conglomerados*. Tais amostras são bastante usadas para reduzir os custos de grandes pesquisas, nas quais os entrevistadores devem ser enviados a locais muito esparsos, fato do qual decorrem muitas viagens. No emprego do método de conglomerados pelos menos dois níveis de amostragem são empregados:

1. *a unidade primária de amostragem* ou conglomerado, que corresponde a alguma área bem delineada onde se concentram características encontradas na população toda (por exemplo, estado, comarca, quarteirão, e assim por diante) e
2. os sujeitos amostrados dentro de cada conglomerado.

Para fins ilustrativos, imaginem que quiséssemos entrevistar uma amostra representativa de sujeitos que vivem numa área grande da cidade. Extrair uma amostra casual simples, ou sistemática ou estratificada de respondentes espalhados numa grande área, implicaria muitas viagens, sem mencionar tempo e dinheiro. Se recorrêssemos a amostragem por conglomerados, entretanto, limitaríamos nossas entrevistas àqueles sujeitos localizados dentro de relativamente poucos conglomerados. Poderíamos começar, por exemplo, tomando os quarteirões da cidade como nossa unidade primária de amostragem ou conglomerado. Seria possível continuar, então, arranjando uma lista de todos os quarteirões abrangidos pela área, a partir da qual sortearíamos uma amostra casual e simples de quarteirões. Depois de ter extraído nossa amostra de quarteirões da cidade, selecionaríamos os respondentes (ou lares) de cada quarteirão pelo emprego do mesmo método casual. Mais especificamente, todos os sujeitos (ou lares) de cada um dos quarteirões escolhidos são listados (arrolados) e, com o auxílio de uma tábua de números aleatórios, seleciona-se uma amostra de respondentes de cada quarteirão. Através do método de conglomerados, cada entrevistador opta por um dos quarteirões já sorteados e contacta alguns respondentes que ali residem.

Numa escala mais ampla, o mesmo procedimento de conglomerados pode ser aplicado a pesquisas de âmbito nacional; para tanto, basta (1) que regiões, estados, paróquias sejam tomados como unidades amostrais primárias, sobre as quais incide o processo inicial de escolha e (2) que de cada região, estado ou paróquia seja extraída uma amostra aleatória simples–objeto das entrevistas. Desse modo, os entrevistado-

res não precisam percorrer (cobrir) todas as regiões, estados ou paróquias, mas somente uma quantidade muito menor de áreas–que tenham sido aleatoriamente selecionadas para a pesquisa.

ERRO AMOSTRAL

Daqui para a frente, tomaremos um cuidado especial em distinguir as características das amostras (que de fato estudamos) das características das populações–que constituem objeto de nossas generalizações. A fim de marcar essa distinção em nossos procedimentos estatísticos, não podemos mais usar os mesmos símbolos para representar a média e o desvio padrão tanto de uma amostra como de uma população. Ao invés, devemos empregar diferentes símbolos, conforme estejamos tratando de características amostrais ou populacionais. Com referência à média, vamos sempre simbolizá-la por $\bar{X}$ quando se tratar de uma *amostra* e por M, quando se tratar de uma *população*. Com referência ao desvio padrão, notaremos por s quando se tratar de *amostra* e por σ, quando se tratar de *população*.

O comportamento típico do pesquisador é tentar obter uma amostra que seja representativa da população (grupo maior) na qual ele tem algum interesse. Uma vez que amostras casuais simples dão a todos os membros da população a mesma oportunidade de seleção, elas acabam sendo, no final das contas, mais representativas das características populacionais do que as amostras não-casuais. Tal como foi discutido de forma sucinta no Capítulo 1, entretanto, por obra do acaso, sozinho, podemos *sempre* esperar alguma diferença entre uma amostra, aleatória ou não, e a população da qual ela foi extraída. $\bar{X}$ quase nunca será exatamente igual a M; s será idêntico a σ em raríssimas ocasiões. Conhecida por *erro amostral*, essa diferença aparece independentemente do fato de o plano amostral ter sido bem elaborado e executado, apesar das melhores intenções do pesquisador, e em casos onde não tenha havido burla ou erro.

Para ilustrar a ocorrência do erro amostral, vamos observar a Tabela 7.1, que contém uma população composta de 20 notas finais de exame e 3 amostras, A, B e C, extraídas de forma casual dessa população (mediante o uso de uma tábua de números aleatórios). Como seria de esperar-se, a média populacional (M = 71,55) não é

TABELA 7.1 População e Três Amostras Casuais de Notas Finais de Exame

População			*Amostra A*	*Amostra B*	*Amostra C*
70	80	93	96	40	72
86	85	90	99	86	96
56	52	67	56	56	49
40	78	57	52	67	56
89	49	48	303	249	273
99	72	30	$\bar{X}$ = 75,75	$\bar{X}$ = 62,25	$\bar{X}$ = 68,25
96	94	1.431			
	M = 71,55				

aritmeticamente idêntica a qualquer das médias amostrais; de modo análogo, há diferenças entre as próprias médias amostrais.

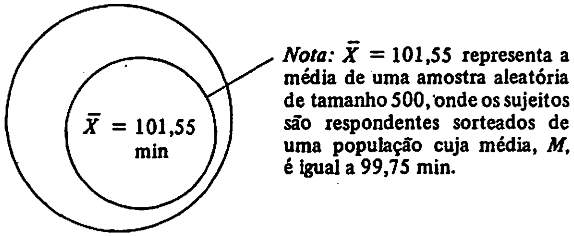
DISTRIBUIÇÃO AMOSTRAL DE MÉDIAS

Dada a existência de erro amostral, o aluno pode perguntar-se como é possível *generalizar* de uma amostra para uma população. Para dar uma resposta sensata, vamos considerar o trabalho de um pesquisador hipotético que esteja estudando o tempo gasto em audição de rádio pela população de um milhão de residentes numa cidade do centro-oeste. Para economizar tempo e dinheiro, ele entrevista somente uma amostra colhida ao acaso da população de residentes. Ele escolhe 500 residentes por meio de uma tábua de números aleatórios e pergunta a cada membro—"Quantos minutos você ouve rádio diariamente?"—e descobre que o tempo gasto varia de 0 a 240 min. Como demonstra a Figura 7.1, o tempo médio gasto em audição de rádio, numa amostra de 500 residentes, é de 101,55 min.

Acontece que o nosso pesquisador hipotético é um tanto excêntrico, e possui uma inclinação notável para a extração de amostras de populações. Tão intenso é o seu entusiasmo por amostragem que ele continua a extração de amostras adicionais, cada uma composta de 500 residentes, e calcula, a seguir, o tempo médio de audição para cada amostra. Esse procedimento continua até que o nosso excêntrico pesquisador tenha conseguido extrair 98 *amostras, cada uma* com 500 residentes. Nesse processo de extração de 98 amostras casuais, ele, de fato, estuda 49.000 respondentes (500 X 98 = 49.000).

Vamos admitir, conforme demonstra a Figura 7.2, que a população inteira de nossa cidade do centro-oeste tenha uma média de tempo de audição de 99,75 min. Como também ilustra a Figura 7.2, vamos supor que as amostras colhidas pelo nosso excêntrico pesquisador comportem médias que variem de 89 a 111 min. De acordo com a discussão prévia que foi feita, isso poderia simples e facilmente ocorrer por causa do erro amostral.

Distribuições de freqüências de *escores brutos* podem obter-se tanto de amostras quanto de populações. De modo análogo, podemos construir uma *distribuição de médias amostrais*, ou seja, uma distribuição de freqüências de um grande número de



FGURA 7.1 Média do Tempo de Audição de uma Amostra Aleatória Tirada de uma População Hipotética

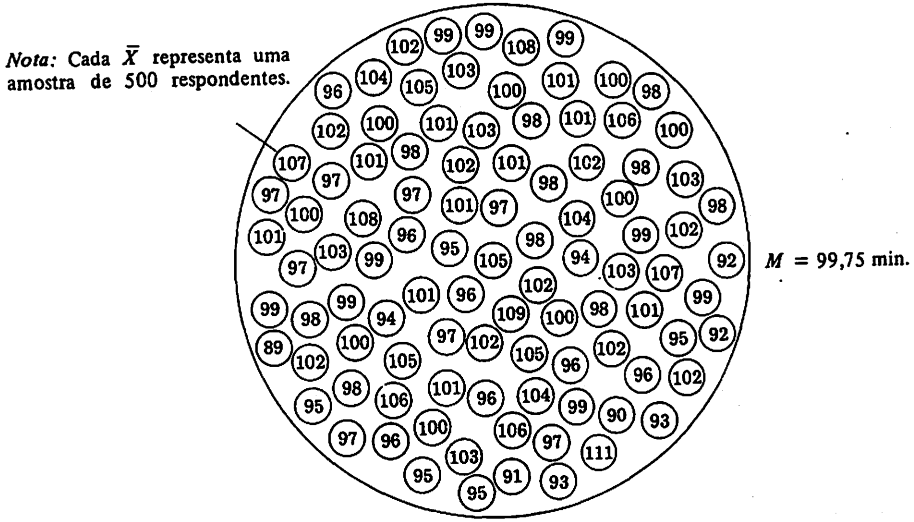


FIGURA 7.2 Média do Tempo de Audição, Relativa a 98 Amostras Casuais Sorteadas de uma População Hipotética em que $M = 99{,}75$ min.

médias amostrais aleatórias que tenham sido extraídas da mesma população. A Tabela 7.2 apresenta, sob a forma de distribuição, as 98 médias amostrais coletadas pelo nosso pesquisador excêntrico. Tal como quando se trabalha com uma distribuição de escores brutos, as médias da Tabela 7.2 foram dispostas consecutivamente, da mais alta à mais baixa, e suas respectivas freqüências de ocorrência indicadas numa coluna adjacente.

Características de uma Distribuição de Médias Amostrais

Até este ponto, não atacamos diretamente o problema da generalização de amostras para populações. O modelo teórico conhecido por distribuição de médias amostrais (tal como foi ilustrado pelas 98 médias amostrais obtidas pelo nosso excêntrico pesquisador) tem certas propriedades que lhe conferem um papel importante no processo de amostragem. Antes de prosseguirmos com o processo de generalizações de amostras para populações, precisamos examinar primeiro as características de uma distribuição amostral de médias:

1. *A distribuição de médias amostrais aproxima-se de uma curva normal.* Tal como graficamente ilustrado na Figura 7.3(a), dispondo as médias amostrais da Tabela 7.2 num polígono de freqüência, acabamos obtendo uma distribuição normal. Isto é verdade para todas as distribuições de médias amostrais independentemente da forma da distribuição dos escores brutos que constituem a população da qual as médias foram extraídas.[1]

[1] Isso pressupõe que extraímos grandes amostras casuais, de igual tamanho, de uma dada população de escores brutos.

2. *A média de uma distribuição de médias amostrais ("a média das médias") é igual à verdadeira média populacional.* Se tomarmos um grande número de médias amostrais casuais da mesma população e calcularmos a média de todas elas, obteremos o valor da verdadeira média populacional. Portanto, como demonstra a Figura 7.3, a média da distribuição amostral de médias (a) é igual à média da população original; (b) essas médias podem ser consideradas permutáveis, isto é, podem ser tomadas umas pelas outras.
3. *O desvio padrão de uma distribuição de médias amostrais é menor do que o desvio padrão da população.*

TABELA 7.2 **Distribuição de Médias Amostrais (Audição de Rádio) Resultantes de 98 Amostras Casuais**

Médias	*f*
111 min	1
110	1
109	1
108	2
107	2
106	3
105	4
104	5
103	6
102	8
101	9
100	9
99	9
98	8
97	7
96	6
95	5
94	4
93	3
92	2
91	1
90	1
89 min	1
	$N = 98$

Como ilustra a Figura 7.3, a variabilidade da distribuição amostral é sempre menor que a variabilidade da população toda. Isso é verdade porque tomamos escores médios (em lugar da totalidade dos escores brutos que entram na composição dessas médias), eliminando, desta forma, valores extremos. Por exemplo, o escore médio 100 pode ser obtido a partir dos escores 60, 90, 110 e 140. Assim: $(60 + 90 + 110 + 140 = \frac{400}{4} = 100)$. Ao levar esses escores brutos a um gráfico, incluímos valores que se estendem de 60 a 140; entretanto, ao colocar o escore médio em gráfico, obviamente reduzimos a ocor-

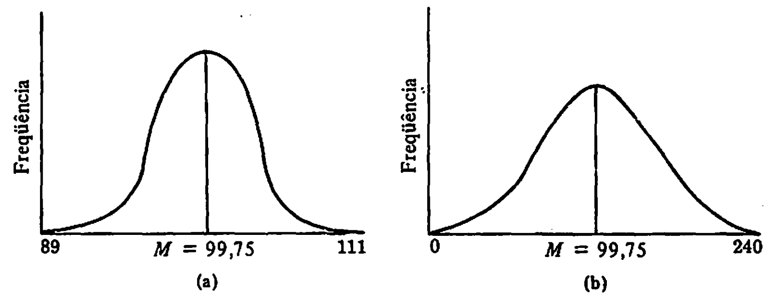


FIGURA 7.3 **Polígonos de Freqüências Relativos a (a) Distribuição de Médias Amostrais da Tabela 7.2 e (b) Populações das Quais essas Médias Foram Extraídas**

rência de valores extremos a um único valor, 100. Como resultado, esperamos obter um desvio padrão menor sempre que certa quantidade de escores médios for agrupada e levada a gráfico.

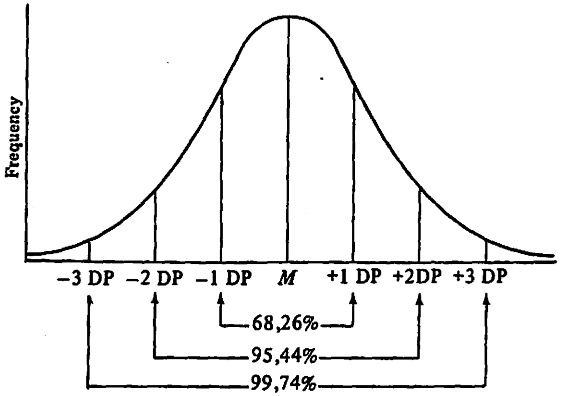
A Distribuição de Médias Amostrais Vista como uma Curva Normal

Se definirmos probabilidade em termos de freqüência de ocorrência, como indicado no Capítulo 6, então a curva normal poderá ser considerada uma distribuição de probabilidades (e pode-se dizer que a probabilidade diminui à medida que caminhamos ao longo da linha de base, a partir da média, em ambos os sentidos).

Com essa noção, podemos encontrar a probabilidade de obter diferentes escores brutos numa distribuição, desde que sejam dados a média e o desvio padrão. Por exemplo, para encontrar a probabilidade associada à localização de um indivíduo cuja renda anual está entre Cr\$85.000,00 e Cr\$119.000,00 numa população de renda média igual a Cr\$85.000,00, e desvio padrão igual a Cr\$25.500,00, traduzimos o escore bruto Cr\$119.000,00 em escore z (+ 1,33) e entramos na Tabela B do fim do livro para obter a porcentagem da freqüência total que cai entre o escore z (1,33) e a média. Essa área contém 40,82% dos escores brutos. Portanto, $P = 0{,}41$ (arredondando) de encontrarmos um sujeito cuja renda anual situa-se entre Cr\$85.000,00 e Cr\$119.000,00. Se desejarmos a probabilidade de encontrar alguém cuja renda é de Cr\$119.000,00 ou mais, precisamos ir um passo adiante, subtraindo de 50% o porcentual obtido na Tabela B. (Subtraímos de 50% porque essa é a porcentagem que se localiza de ambos os lados da média.) Subtraindo 40,82% de 50%, verificamos que 9,18% caem exatamente em ou além de Cr\$119.000,00. Portanto, movendo a vírgula duas casas para a esquerda, podemos dizer que $P = 0{,}09$ (9 casos favoráveis em 100) de encontrarmos um sujeito de renda igual a ou maior que Cr\$119.000,00.

Na presente situação, não estamos interessados em obter probabilidades associadas à distribuição de *escores brutos*. Ao contrário, encontramo-nos trabalhando com uma distribuição de *médias amostrais* que foram extraídas de uma população total de escores, e desejamos fazer afirmações probabilísticas acerca dessas mesmas médias.

Como ilustra a Figura 7.4, já que a distribuição de médias amostrais toma a forma



DP = desvio padrão

FIGURA 7.4 A Distribuição de Médias Amostrais Vista Como uma Distribuição de Probabilidades

de uma curva normal, podemos dizer que as probabilidades decrescem à medida que nos distanciamos da média das médias (isto é, da verdadeira média populacional). Isto faz sentido porque, como o aluno poderá recordar, a distribuição amostral é um resultado que decorre de diferenças aleatórias entre médias amostrais (erro amostral). Por essa razão, podemos esperar que, por acaso e puro acaso, a maioria das médias amostrais cairá perto do valor que representa a verdadeira média populacional, enquanto que relativamente poucas médias amostrais cairão longe dela.

A Figura 7.4 indica que cerca de 68% das médias amostrais de uma distribuição amostral caem entre −1 DP e +1 DP, a partir da média das médias (a verdadeira média populacional). Em termos probabilísticos, podemos dizer que $P = 0{,}68$ para qualquer média amostral que caia nesse intervalo. Da mesma forma, podemos dizer que a probabilidade de que qualquer média amostral caia entre −2 DP e +2 DP é de 0,95 (95 casos favoráveis em 100), e assim por diante.

Uma vez que a distribuição amostral toma a forma da curva normal, podemos também usar os escores z e a Tabela B para calcular a probabilidade de obter *qualquer média amostral*–e não apenas aquelas que são múltiplos exatos do desvio padrão. Dados a média das médias e o desvio padrão da distribuição amostral, o processo é idêntico ao usado no capítulo anterior relativamente a uma distribuição de escores brutos. A única coisa que muda são os nomes.

Imaginem, por exemplo, que certa universidade divulgue que seus alunos recebem uma renda anual média (M) de Cr$34.000,00. Supondo que temos boas razões para questionar a legitimidade dessa afirmação, decidimos testá-la com uma amostra aleatória de 100 alunos. No processo, obtemos uma média amostral de apenas Cr$23.800,00. Perguntamos agora: com que probabilidade obteríamos uma média de Cr$23.800,00 ou menos se a verdadeira média populacional fosse, de fato, Cr$34.000,00? Será que a universidade divulgou a verdade? Ou será que isso é apenas um golpe publicitário na tentativa de aumentar as matrículas ou as doações? A Figura 7.5 ilustra a área para a qual buscamos uma solução.

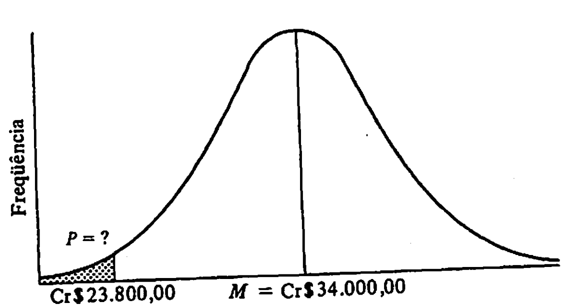


FIGURA 7.5 Probabilidade Associada à Obtenção de uma Média Amostral de Cr$23.800,00 ou Menos, no Caso de a Verdadeira Média Populacional ser Cr$34.000,00 com um Desvio Padrão de Cr$4.420,00

Suponhamos que o desvio padrão da distribuição amostral seja de Cr$4.420,00. Seguindo o procedimento estabelecido, "traduzimos" a média amostral de Cr$23.800,00 em escore z, da seguinte forma:

$$z = \frac{\bar{X} - M}{\sigma_{\bar{X}}} = \frac{23.800{,}00 - 34.000{,}00}{4.420{,}00} = -2{,}31,$$

onde

$\bar{X}$ = média de uma amostra da população

M = média das médias (que corresponderia à verdadeira média populacional–segundo a declaração da universidade)

$\sigma_{\bar{X}}$ = desvio padrão da distribuição de médias amostrais.

O resultado do procedimento acima informa-nos que uma média amostral de Cr$23.800,00 cai a exatamente 2,31 desvios padrões abaixo da média populacional de Cr$34.000,00–argüída como verdadeira. Consultando a Tabela B no fim do livro, verificamos que 48,96% das médias amostrais situam-se entre Cr$23.800,00 e Cr$34.000,00. Se subtrairmos essa porcentagem de 50%, obteremos o porcentual da distribuição que corresponde às médias amostrais de Cr$23.800,00 ou menos–na hipótese de a média populacional ser *mesmo* Cr$34.000,00. Tal porcentagem corresponde a 1,04% (50% − 48,96% = 1,04%). Portanto, a probabilidade, por arredondamento, é de 0,01 (1 caso em 100) de obter-se uma média amostral de Cr$23.800,00 ou menos, caso a verdadeira média populacional seja Cr$34.000,00. Com uma probabilidade tão pequena de erro, podemos dizer com certa confiança que a verdadeira média populacional *não* é realmente Cr$34.000,00. Não é de estranhar se o relatório da universidade com relação à renda anual de seus alunos representar apenas propaganda de baixa categoria.

ERRO PADRÃO DA MÉDIA

Até agora, estivemos admitindo que o pesquisador possuía, de fato, informação de primeira qualidade com referência à distribuição de médias amostrais. Estivemos trabalhando como se ele, da mesma forma como o pesquisador excêntrico, realmente tivesse coligido dados de um número bem grande de médias amostrais, todas colhidas de forma aleatória de alguma população. Se tal fato tivesse ocorrido como se acabou de expor, seria uma tarefa bastante simples fazer generalizações a respeito da população, uma vez que a média das médias assume um valor que é igual ao da verdadeira média populacional.

Na prática, o pesquisador raramente coleta dados de mais do que uma ou duas amostras, a partir das quais ele ainda deseja generalizar para a população toda. Extrair uma distribuição de médias amostrais requer o mesmo esforço que o de estudar individualmente cada membro da população. Como resultado, o pesquisador não tem conhecimento real nem da média das médias nem do desvio padrão da distribuição amostral. Todavia, ele possui um bom método para *estimar* o desvio padrão da distribuição de médias amostrais a partir dos dados que lhe fornecem uma única amostra. Essa estimativa é conhecida por *erro padrão da média* e é simbolizada por $\sigma_{\bar{X}}$.* Em símbolos:

$$\sigma_{\bar{X}} = \frac{s}{\sqrt{N-1}}\text{, onde}$$

$\sigma_{\bar{X}}$ = erro padrão da média (estimativa do desvio padrão de uma distribuição de médias amostrais)
s = desvio padrão de uma *amostra*
N = número total de escores da *mesma amostra*.

Para ilustrar, se o desvio padrão de uma amostra de dez respondentes for 2,5, então

$$\sigma_{\bar{X}} = \frac{2{,}5}{\sqrt{10-1}} = \frac{2{,}5}{3{,}0} = 0{,}83$$

Como salientamos acima, o pesquisador que examina apenas uma ou duas amostras não pode conhecer a média das médias, ou seja, o valor que corresponde à verdadeira média da população. A única coisa que ele tem é a média amostral, a qual difere da média real (da população) por força do chamado erro de amostragem. Mas, com isso, não acabamos voltando ao ponto de partida? Como é possível estimar a verdadeira média populacional a partir de uma simples média amostral, especialmente diante de tais diferenças inevitáveis entre amostras e populações?

* N.T.: Em muitos textos, o erro padrão da média, baseado no desvio padrão da população e simbolizado por $\sigma_{\bar{X}}$, distingue-se do erro padrão da média estimado, cujo cálculo baseia-se no desvio padrão da amostra e cujo símbolo é $S_{\bar{X}}$. Sem medir toda a população, entretanto, o valor do seu desvio padrão não é conhecido e deve, por isso, ser estimado. A título de simplificação, decidimos fechar os olhos a essa distinção, introduzindo uma única fórmula para o erro padrão da média, simbolizado por $\sigma_{\bar{X}}$, cuja base são os dados amostrais.

Na verdade, já nos distanciamos bastante de nossa posição original. Depois de ter discutido a natureza da distribuição de médias amostrais, estamos agora preparados para estimar o valor de uma média populacional. Com a ajuda do erro padrão da média, podemos encontrar *o intervalo de valores dentro do qual a verdadeira média populacional pode cair. Podemos, também, estimar a probabilidade com que nossa média populacional realmente cairá dentro desse mesmo intervalo (de valores de médias)*. Este é o conceito de *intervalo de confiança*.

INTERVALOS DE CONFIANÇA

A fim de podermos explorar o procedimento que leva à obtenção de um intervalo de confiança, vamos aprofundar o exame de um exemplo anterior. Suponham que uma amostra aleatória de 100 alunos de certa universidade apresente uma renda anual média de Cr$ 23.800,00. Como esses dados resultam de apenas uma amostra aleatória, e não de toda a população de alunos, não há uma maneira segura de garantir que essa renda média reflita, de fato, a renda média da população de alunos da universidade. Como já vimos anteriormente, o erro amostral é, no fim das contas, a conseqüência inevitável de amostrar a partir de populações.

Sabemos, entretanto, que 68,26% de todas as médias amostrais aleatórias, numa distribuição de médias amostrais, situar-se-ão entre −1 DP e +1 DP, a contar da verdadeira média populacional. Se estimarmos o desvio padrão da distribuição amostral ($\sigma_{\bar{X}}$ = Cr$ 3.400,00), e usarmos a nossa média amostral de Cr$ 23.800,00 como uma estimativa da média populacional, poderemos estabelecer o intervalo dentro do qual, com 68 possibilidades em 100 (por arredondamento), a verdadeira média populacional cairá. Conhecido por *intervalo de confiança de 68%*, esse intervalo de rendas médias está ilustrado graficamente na Figura 7.6.

O intervalo de confiança de 68% pode ser obtido da seguinte maneira:

$$\text{intervalo de confiança de 68\%} = \bar{X} \pm \sigma_{\bar{X}}\text{, onde}$$

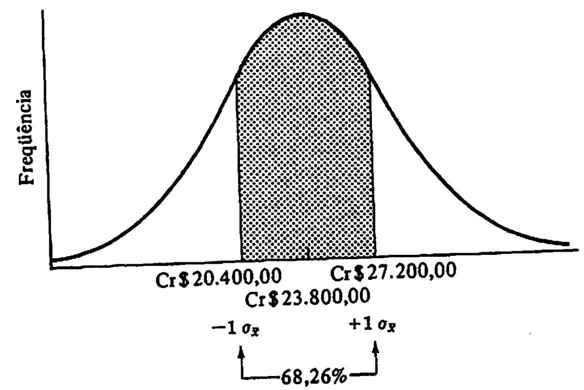


FIGURA 7.6 Um Intervalo de Confiança de 68%, Quando $\sigma_{\bar{X}}$ = Cr$ 3.400,00 e $\bar{X}$ = Cr$ 23.800,00

$\bar{X}$ = média amostral
$\sigma_{\bar{X}}$ = erro padrão da média

Aplicando a fórmula acima ao problema em foco, resulta:

intervalo de confiança de 68% = Cr$ 23.800,00 ± Cr$ 3.400,00
= Cr$ 20.400,00 ⊢⊣ Cr$ 27.200,00

O pesquisador, portanto, afirma que ele tem *68% de confiança* de que a renda média da população composta por esses alunos universitários é de Cr$ 23.800,00 ± Cr$ 3.400,00. Em outras palavras, há 68 possibilidades em 100 ($P = 0,68$) de que a verdadeira média populacional caia realmente no intervalo de limites Cr$ 20.400,00 e Cr$ 27.200,00 (Cr$ 23.800,00 – Cr$ 3.400,00 = Cr$ 20.400,00; Cr$ 23.800,00 + Cr$ 3.400,00 = = Cr$ 27.200,00). Essa estimativa é feita a despeito do erro amostral, embora dentro de uma margem de erro (de mais ou menos Cr$ 3.400,00) e a um nível de confiança especificado (68%).

Os intervalos de confiança podem ser construídos para qualquer nível de probabilidade. A maioria dos pesquisadores não se sente suficientemente confiante em estimar uma média populacional sabendo que só há 68 possibilidades em 100 da mesma estar correta (de cada 100 médias amostrais 68 cairão dentro do intervalo Cr$ 20.400,00 e Cr$ 27.200,00). Como resultado, tornou-se convencional usar um intervalo de confiança *mais amplo*, embora *menos preciso*, mas que assegura com *melhor probabilidade* o cálculo da estimativa da média populacional. O uso do *intervalo de confiança de 95%* permite obter esse cálculo através do qual a média populacional é estimada com a certeza de que há 95 possibilidades em 100 de estar-se correto; e há 5 possibilidades em 100 de estar-se errado (de cada 100 médias amostrais, 95 cairão dentro do intervalo). Entretanto, mesmo usando o intervalo de confiança de 95%, deve-se ter sempre em mente que a média amostral do pesquisador poderia ser uma daquelas cinco médias amostrais que caíram fora do intervalo estabelecido. Na tomada de decisões estatísticas nunca há possibilidade de ter certeza absoluta.

Como proceder para encontrar o intervalo de confiança de 95%? Já sabemos que 95,44% das médias amostrais numa distribuição amostral situam-se entre –2 DP e +2 DP, contados a partir da média das médias. Consultando a Tabela B, podemos afirmar que 1,96 desvios padrões, em ambos os sentidos, cobrem exatamente 95% das médias amostrais (47,50% de cada lado da média das médias). A fim de encontrar o intervalo de confiança de 95%, devemos primeiro multiplicar o erro padrão da média por 1,96 (o intervalo corresponde a 1,96 unidades de $\sigma_{\bar{X}}$, contadas para a esquerda e para a direita a partir da média). Portanto,

intervalo de confiança de 95% = $\bar{X} \pm (1{,}96)\,\sigma_{\bar{X}}$, onde

$\bar{X}$ = média amostral
$\sigma_{\bar{X}}$ = erro padrão da média.

Se aplicarmos o intervalo de confiança de 95% à estimativa da renda média dos alunos universitários, veremos que

intervalo de confiança de 95% = Cr$ 23.800,00 ± (1,96) Cr$ 3.400,00
= Cr$ 23.800,00 ± Cr$ 6.664,00
= Cr$ 17.136,00 ⊢⊣ Cr$ 30.464,00

Conclusão: temos 95% de certeza de que a verdadeira média populacional cai entre Cr$ 17.136,00 e Cr$ 30.464,00.

Vamos resumir, passo a passo, o procedimento que leva à obtenção do intervalo de confiança de 95% na seguinte amostra casual de escores brutos:

X
1
5
2
3
4
1
2
2
4
3

PASSO 1: Calcular a média da amostra.

X
1
5
2
3
4
1
2
2
4
3
$\Sigma X = 27$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{27}{10} = 2{,}7$$

PASSO 2: Calcular o desvio padrão da amostra.

X	X^2
1	1
5	25
2	4
3	9
4	16

$$s = \sqrt{\frac{\Sigma X^2}{N} - \bar{X}^2}$$

$$= \sqrt{\frac{89}{10} - (2{,}7)^2}$$

X	X^2	
1	1	
2	4	$=\sqrt{8,9-7,29}$
2	4	$=\sqrt{1,61}$
4	16	$= 1,27$
3	9	
	$\Sigma X^2 = 89$	

PASSO 3: Calcular o erro padrão da média.

$$\sigma_{\bar{X}} = \frac{s}{\sqrt{N-1}} = \frac{1,27}{\sqrt{10-1}} = \frac{1,27}{3} = 0,42$$

PASSO 4: Multiplicar o erro padrão da média por 1,96.

$$\text{Intervalo de confiança de 95\%} = \bar{X} \pm (1,96)\,\sigma_{\bar{X}}$$
$$= 2,7 \pm (1,96)(0,42) = 2,7 \pm 0,82$$

PASSO 5: Somar e subtrair esse produto da média amostral, a fim de obter o intervalo dentro do qual cai a média populacional.

$$\text{Intervalo de confiança de 95\%} = 2,7 \pm 0,82 = 1,88 \longmapsto 3,52$$

Podemos estar 95% confiantes de que a verdadeira média populacional caia entre 1,88 e 3,52.*

Um intervalo de confiança ainda mais rigoroso é o *intervalo de confiança de 99%.* Ao consultar a Tabela B no fim do livro, vemos que o escore z de 2,58 representa 49,50% da área situada de cada lado da curva, a partir do centro (média). Dobrando essa porcentagem, obtemos 99% da área sob a curva; isso significa que 99% das médias amostrais cairão nesse intervalo. Em termos probabilísticos, de cada 100 médias amostrais, 99 cairão entre −2,58 DP e +2,58 DP a contar da média. Reciprocamente, somente 1 de cada 100 médias cairá fora do intervalo. Em símbolos:

$$\text{intervalo de confiança de 99\%} = \bar{X} \pm (2,58)\,\sigma_{\bar{X}}, \text{ onde}$$

$\bar{X}$ = média amostral
$\sigma_{\bar{X}}$ = erro padrão da média.

* N.T.: A título ilustrativo foi usada uma amostra pequena. Na prática, entretanto, um pesquisador que use o procedimento acima para calcular um intervalo de confiança deve trabalhar com, pelo menos, 30 dados, a fim de garantir a suposição de normalidade na distribuição de médias amostrais. (Ver a discussão da estatística t no Capítulo 8.)

Com referência à nossa estimativa da renda média dos alunos universitários, temos:

intervalo de confiança de 99% = Cr\$ 23.800,00 ± (2,28) Cr\$ 3.400,00
= Cr\$ 23.800,00 ± Cr\$ 8.772,00
= Cr\$ 15.028,00 ⟼ Cr\$ 32.572,00

Acabamos de estabelecer que, com 99% de certeza, a verdadeira média populacional cairá em algum ponto entre Cr\$ 15.028,00 e Cr\$ 32.572,00.

O aluno deveria observar que o intervalo de confiança de 99% consiste em uma faixa mais ampla (de Cr\$ 15.028,00 a Cr\$ 32.572,00) do que o intervalo de confiança de 95% (de Cr\$ 17.136,00 a Cr\$ 30.464,00). O intervalo de confiança de 99% abrange a maior parte da área total sob a curva normal e, portanto, um número maior de médias amostrais. Essa faixa mais ampla de escores médios dá-nos maior confiança no sentido de termos estimado bem a verdadeira média populacional. Apenas uma única média amostral em cada 100 ficará fora do intervalo. Por outro lado, se aumentarmos a nossa confiança de 95% para 99%, também sacrificaremos o grau de precisão no estabelecimento da média populacional. Mantendo o tamanho da amostra constante, o pesquisador deve escolher entre estar correto com maior precisão ou estar correto com maior confiança.

A fim de resumir passo a passo o procedimento para o cálculo do intervalo de confiança de 99%, vamos reexaminar a mesma amostra aleatória de escores:

X
1
5
2
3
4
1
2
2
4
3

PASSO 1: Calcular a média da amostra.

X	
1	
5	
2	
3	$\bar{X} = \frac{\Sigma X}{N}$
4	
1	
2	$= \frac{27}{10}$
2	
4	$= 2,7$
3	
$\Sigma X = 27$	

PASSO 2: Calcular o desvio padrão da amostra.

X	X^2
1	1
5	25
2	4
3	9
4	16
1	1
2	4
2	4
4	16
3	9
	$\Sigma X^2 = 89$

$$s = \sqrt{\frac{\Sigma X^2}{N} - \bar{X}^2} = \sqrt{\frac{89}{10} - (2{,}7)^2} = \sqrt{8{,}9 - 7{,}29} = \sqrt{1{,}61} = 1{,}27$$

PASSO 3: Calcular o erro padrão da média.

$$\sigma_{\bar{X}} = \frac{s}{\sqrt{N-1}} = \frac{1{,}27}{\sqrt{10-1}} = \frac{1{,}27}{3} = 0{,}42$$

PASSO 4: Multiplicar o erro padrão da média por 2,58.

$$\text{Intervalo de confiança de 99\%} = \bar{X} \pm (2{,}8)\sigma_{\bar{X}} = 2{,}7 \pm (2{,}58)(0{,}42) = 2{,}7 \pm 1{,}08$$

PASSO 5: Somar e subtrair esse resultado (produto) da média amostral, a fim de ober o intervalo dentro do qual cai a média populacional.

$$\text{Intervalo de confiança de 99\%} = 2{,}7 \pm 1{,}08 = 1{,}62 \longmapsto 3{,}78.$$

Temos, assim, 99% de certeza de que a verdadeira média populacional cai entre 1,62 e 3,78.

ESTIMATIVA DE PROPORÇÕES

Até aqui, concentramo-nos em procedimentos para estimar médias populacionais. Não raro, entretanto, o pesquisador procura obter uma estimativa duma *proporção* populacional a partir de uma outra proporção resultante do estudo de uma amostra casual. Uma circunstância em que isso ocorre com freqüência é a chamada prévia eleitoral, cujos dados sugerem que certa proporção de votos irá para determinado partido ou determinado candidato. Quando um pesquisador de opinião pública anuncia que 45% dos votos serão a favor de certo candidato, ele o faz com a convicção de que não está 100% correto. Em geral, ele tem 95% ou 99% de confiança de que sua proporção estimada cai dentro de certa faixa de proporções (por exemplo, entre 40% e 50%).

Estimamos proporções da mesma maneira utilizada para a estimação de médias. Todas as estatísticas–inclusive médias e proporções–possuem distribuições amostrais próprias. Da mesma forma como calculamos há pouco o erro padrão da média, podemos agora calcular o *erro padrão da proporção*. Em símbolos:

$$\sigma_P = \sqrt{\frac{P(1-P)}{N}} \quad \text{onde}$$

σ_P = erro padrão da proporção (estimativa do desvio padrão da distribuição amostral de proporções)
P = proporção amostral
N = tamanho da amostra.

A título ilustrativo, vamos admitir que 45% de uma amostra aleatória de 100 alunos universitários declare-se favorável à legalização do uso de drogas. O erro padrão da proporção seria

$$\sigma_P = \sqrt{\frac{0{,}45(0{,}55)}{100}} = \sqrt{\frac{0{,}2475}{100}} = \sqrt{0{,}0025} = 0{,}05.$$

Para encontrar o intervalo de confiança de 95%, multiplicamos o erro padrão da proporção por 1,96, e o resultado obtido (a) somamos à proporção amostral e (b) subtraímos da proporção amostral. Assim:

$$\text{Intervalo de confiança de 95\%} = P \pm (1{,}96)\sigma_P \text{, onde}$$

P = proporção amostral
σ_P = erro padrão da proporção.

Se quisermos estimar a proporção de universitários que são favoráveis à legalização do uso de drogas, procedemos da seguinte forma:

$$\text{Intervalo de confiança de 95\%} = 0{,}45 \pm (1{,}96)(0{,}05) = 0{,}45 \pm 0{,}098 = 0{,}35 \longmapsto 0{,}55$$

Temos, dessa maneira, 95% de certeza de que a verdadeira proporção populacional nem é menor que 0,35, nem maior que 0,55. Mais especificamente, em algum ponto entre 35% e 55% situa-se a porcentagem de alunos universitários que são favoráveis a essa legalização. Há um risco de 5% de estarmos errados; por outras palavras, 5 vezes em cada 100 tais intervalos de confiança não conterão a verdadeira proporção populacional.

Vamos resumir o procedimento que permite estimar uma proporção mediante o uso do intervalo de confiança de 95%. Admitamos que a proporção amostral a partir da qual nossa estimativa deve ser feita seja da ordem de 0,40 (40% dos 100 sujeitos cairão nessa categoria).

PASSO 1: Obter o erro padrão da proporção.

$$\sigma_P = \sqrt{\frac{P(1-P)}{N}} = \sqrt{\frac{0{,}40(0{,}60)}{100}} = \sqrt{\frac{0{,}24}{100}} = \sqrt{0{,}0024} = 0{,}049$$

PASSO 2: Multiplicar o erro padrão da proporção por 1,96.

$$\text{Intervalo de confiança de } 95\% = P \pm (1{,}96)\sigma_P = 0{,}40 \pm (1{,}96)(0{,}049)$$
$$= 0{,}40 \pm 0{,}096$$

PASSO 3: Somar esse resultado (produto) à proporção amostral e depois subtrair esse resultado da mesma proporção amostral, a fim de obter o intervalo dentro do qual se situa a proporção populacional.

$$\text{Intervalo de confiança de } 95\% = 0{,}40 \pm 0{,}096 = 0{,}30 \vdash\!\dashv 0{,}50$$

Podemos assim dizer, com 95% de certeza, que a verdadeira proporção populacional cai entre 0,30 e 0,50.

RESUMO

Este capítulo explorou os conceitos e os procedimentos básicos relacionados com o problema de fazer generalizações de amostras para populações. Foram apresentados dois métodos de amostragem: casual e não-casual. Salientou-se que o erro amostral–aquela diferença inevitável entre amostras e populações–ocorre sempre, a despeito de o plano de amostragem ter sido bem concebido e bem executado. Como conseqüência do erro amostral, há possibilidade de discutir as características da distribuição de médias amostrais, distribuição essa que, graficamente, assume a forma de uma curva normal, e cujo desvio padrão pode ser estimado com o auxílio do erro padrão da média. Munidos de tais informações, podemos calcular intervalos de confiança para médias (ou proporções), dentro dos quais ser-nos-á possível ter uma dada certeza (95% ou 99%) de onde a verdadeira média (ou proporção) populacional cai de fato. Dessa forma, teremos condições de fazer generalizações de uma amostra para uma população.

PROBLEMAS

1. Calcule o erro padrão da média com a seguinte amostra composta de 30 escores amostrais.

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

2. Com a média amostral do problema 1, calcule (a) o intervalo de confiança de 95%; (b) o intervalo de confiança de 99%.
3. Calcule o erro padrão da média com a seguinte amostra composta de 34 escores:

10	1
4	8
10	7
5	5
5	6
6	10
7	6
3	8
5	7
4	7
4	6
5	5
6	5
6	4
7	3
5	4
8	5

4. Com a média amostral do problema 3, calcule (a) o intervalo de confiança de 95%; (b) o intervalo de confiança de 99%.
5. Calcule o erro padrão da média com a seguinte amostra composta de 32 escores:

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

6. Com a média amostral do problema 5, calcule (a) o intervalo de confiança de 95%; (b) o intervalo de confiança de 99%.
7. Tendo em mente estimar a proporção de alunos de um determinado "campus" universitário que eram favoráveis à extinção dos centros acadêmicos, um pesquisador social entrevistou uma amostra aleatória de 50 estudantes e descobriu que 57% deles era, de fato, favorável a essa extinção (proporção amostral = 0,57). Com essa

informação, (a) calcule o erro padrão da proporção e (b) construa um intervalo de confiança de 95%.

8. Dadas uma amostra de tamanho 150 e uma proporção amostral de 0,32, (a) calcule o erro padrão da proporção e (b) construa um intervalo de confiança de 95%.
9. Dadas uma amostra de tamanho 200 e uma proporção amostral de 0,25, (a) calcule o erro padrão da proporção e (b) construa um intervalo de confiança de 95%.

PARTE III
Tomada de Decisões

8
Testes de Diferenças entre Médias

No Capítulo 7, vimos que uma média populacional ou uma proporção podem ser estimadas a partir da informação que obtemos de uma única amostra. Por exemplo, poderíamos estimar o grau de anomia em determinada cidade, a proporção de pessoas idosas economicamente prejudicadas ou a atitude média relativa ao problema da segregação racial numa população de norte-americanos negros.

Muito embora tenha o enfoque descritivo da estimação de médias e proporções importância óbvia, ele *não* constitui a atividade principal nem o objetivo último do processo de tomada de decisões exercitado pelos pesquisadores. A maioria deles, contrariamente, preocupa-se com o problema denominado *teste de hipótese*, onde estão envolvidas diferenças entre duas ou mais amostras.

Ao testarem diferenças entre amostras, os pesquisadores fazem perguntas tais como: (a) Será que os alemães diferem dos norte-americanos com respeito à obediência à autoridade? (b) O maior índice de suicídios ocorre entre protestantes ou católicos? (c) Será que o fato de um entrevistador ser branco ou preto interfere na "honestidade" da resposta de um negro? (d) Será que políticos conservadores dão a seus filhos uma disciplina mais severa que políticos liberais? (Ver Capítulo 1.) Observe-se que cada uma dessas indagações implica fazer uma *comparação* entre dois grupos: conservadores versus liberais, entrevistadores negros versus entrevistadores brancos, protestantes versus católicos, alemães versus norte-americanos.

HIPÓTESE NULA: NÃO HÁ DIFERENÇA ENTRE AS MÉDIAS

Tornou-se habitual em análise estatística começar pelo teste da chamada *hipótese nula*—que, em termos bem simples, *afirma terem duas amostras sido extraídas da mesma população.* De acordo com a hipótese nula, qualquer diferença observada entre as amostras é considerada como uma ocorrência casual, mero resultado de erro amostral. Portanto, uma diferença entre duas *médias amostrais* não representa, à luz da hipótese nula, uma *verdadeira diferença* entre as *médias populacionais.*

No contexto presente, a hipótese nula pode ser simbolizada da seguinte forma:

$M_1 = M_2$, onde

M_1 = média da primeira população
M_2 = média da segunda população

Vamos examinar as hipóteses nulas relativas às indagações há pouco formuladas:

1. Os alemães são nem mais nem menos obedientes à autoridade que os norte-americanos.
2. Protestantes apresentam a mesma taxa de suicídio que católicos.
3. Respondentes negros são igualmente honestos em suas respostas, quer sejam entrevistados por negros, que por brancos.
4. As crianças tanto de políticos conservadores quanto de políticos liberais recebem a mesma "dose" de sentido disciplinar.

Deve-se notar que a hipótese nula *não nega* a possibilidade de ocorrerem diferenças entre médias *amostrais*. Ao contrário, ela procura explicar tais diferenças através da inevitável presença do erro amostral. De acordo com a hipótese nula, por exemplo, se descobrirmos que uma *amostra aleatória* de dentistas do sexo *feminino* ganha menos dinheiro ($\bar{X}$ = Cr$ 204.000,00) do que uma *amostra aleatória* de dentistas do sexo *masculino* ($\bar{X}$ = Cr$ 255.000,00), não concluímos, *só por isso*, que a *população das dentistas* ganha menos que a *população dos dentistas*. Em lugar disso, tratamos nossa diferença amostral (Cr$ 255.000,00 − Cr$ 204.000,00 = Cr$ 51.000,00) como um produto de erro amostral–aquela diferença que inevitavelmente surge quanto extraímos uma amostra de uma população. Como veremos mais tarde, esse aspecto da hipótese nula estabelece um elo importante com a teoria da amostragem.

HIPÓTESE EXPERIMENTAL:* HÁ DIFERENÇA ENTRE AS MÉDIAS

A hipótese nula é em geral (embora não necessariamente) estabelecida com a esperança de que possa ser rejeitada. Tal afirmação faz muito sentido se atentarmos para o fato de que a maioria dos pesquisadores vive buscando relações entre variáveis. Em outras palavras, eles estão sempre muito mais interessados na busca de diferenças do que na prova de que elas não existem. Para ilustrar: quem se daria o trabalho de estudar católicos e protestantes, quanto à sua taxa de suicídio, na esperança de concluir que eles *não* diferem? As diferenças entre grupos–sejam elas baseadas em fundamentos teóricos ou empíricos–freqüentemente fornecem a estrutura lógica para a realização de um estudo.

Se rejeitarmos a hipótese nula, ou seja, se concluirmos que a hipótese que antecipa a inexistência de diferença entre as médias é *frágil*, automaticamente aceitaremos a

* N.T.: A hipótese experimental, também chamada *hipótese alternativa*, costuma ser indicada por H_a (ler "agá-a"). A hipótese nula, também chamada *hipótese probanda*, designa-se por H_0 (ler "agá-zero").

hipótese experimental que afirma *existir* uma verdadeira diferença populacional. Em pesquisa, esse é o resultado final sempre esperado. A hipótese experimental estabelece que as duas amostras foram colhidas de populações portadoras de médias diferentes. Diz mais: que a diferença obtida do confronto entre as médias amostrais é *grande demais* para ser explicada *apenas* por erro de amostragem.

A hipótese experimental relativa à diferença entre duas médias é simbolizada como se segue:

$M_1 \neq M_2$, onde

M_1 = média da primeira população
M_2 = média da segunda população ($\neq$ significa "não é igual")

Com referência às indagações há pouco formuladas, eis como poderiam ser as hipóteses experimentais:

1. Os alemães diferem dos norte-americanos com relação à obediência à autoridade.
2. Protestantes e católicos apresentam diferentes taxas de suicídio.
3. Respondentes negros diferem em suas respostas conforme sejam entrevistados por negros ou brancos.
4. Políticos liberais diferem de políticos conservadores quanto à "dose" de sentido disciplinar imposta aos filhos.

DISTRIBUIÇÃO AMOSTRAL DE DIFERENÇAS ENTRE MÉDIAS

No capítulo precedente, vimos que as 98 médias resultantes das 98 amostras colhidas pelo nosso excêntrico pesquisador puderam ser levadas a gráfico, dando origem a uma distribuição amostral de médias. De modo análogo, vamos agora imaginar que o mesmo pesquisador excêntrico colha *não uma*, mas *duas* amostras aleatórias *de cada vez*, de um dado universo de pessoas. Suponha-se, por exemplo, que ele colha uma amostra de 500 políticos liberais *e* uma amostra de 500 políticos conservadores. Para testar a hipótese experimental de que os liberais são pais *mais* ou *menos* permissivos que os conservadores, ele formula perguntas a todos os indivíduos componentes das amostras acerca de seus métodos educacionais (por exemplo: Você costuma punir seus filhos? Você já os espancou? Em caso afirmativo, com que freqüência?). Das respostas a tais perguntas, o pesquisador obtém uma medida* da permissividade relacionada com a educação

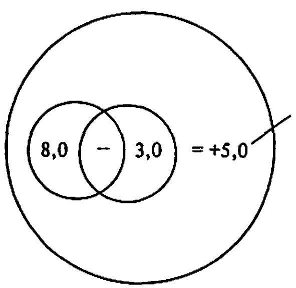
* N.T.: Seria ingenuidade imaginar que o conjunto de tais perguntas constitui um instrumento válido, capaz de produzir uma medida consistente da variável "permissividade educacional". É fácil de perceber que nada impede–aliás tudo favorece!–que esses políticos dêem respostas que, embora contrariando sua postura pessoal, convenham à imagem pública que desejam projetar. Não se trata, aqui, de "criticar" o instrumento "desse excêntrico pesquisador", mas, sim, de alertar o leitor para o fato de que a construção de um instrumento de medida (questionário ou qualquer outro) implica utilizar tecnologia bastante especial.

de crianças, medida essa que pode ser usada na comparação entre a amostra de liberais e a de conservadores. Os escores que resultam desse questionário variam de 1 (não permissivo) a 10 (muito permissivo). Como graficamente ilustrado na Figura 8.1, nosso pesquisador excêntrico descobre que a amostra proveniente de liberais é mais permissiva ($\bar{X} = 8{,}0$) do que aquela procedente de conservadores ($\bar{X} = 3{,}0$).

Poderíamos perguntar: à luz do erro amostral, será que a diferença entre 8,0 e 3,0 ($8{,}0 - 3{,}0 = +5{,}0$) pode ser estritamente explicada *apenas* pela ação do acaso? Será que devemos aceitar a hipótese nula (H_0) de que não há diferença populacional?* Ou será que a diferença amostral obtida (+5,0) é grande o suficiente para indicar (sugerir) a existência de uma verdadeira diferença populacional entre conservadores e liberais, diferença relacionada com suas práticas de educar crianças?

Aprendemos, no Capítulo 2, a interpretar distribuições de freqüências de dados brutos colhidos numa particular população. No Capítulo 7, vimos que era possível construir uma distribuição amostral de escores médios, bem como uma distribuição de freqüências de médias amostrais. Focalizando nossa atenção no problema específico desse capítulo, precisamos ampliar a noção de distribuição de freqüências e examinar a natureza de uma *distribuição amostral de diferenças*, isto é, uma distribuição de freqüências de um número bem grande de *diferenças* entre médias amostrais aleatórias que tenham sido extraídas de uma dada população.

Para ilustrar a distribuição amostral de diferenças, vamos voltar ao trabalho de nosso pesquisador excêntrico cujo amor pela amostragem levou-o, mais uma vez, a extrair um número tão grande de amostras que ultrapassa os limites habituais. Em lugar de extrair uma única amostra de 500 liberais e outra única amostra de 500 conservadores, ele decide extrair 70 *pares* de tais amostras (70 amostras com, *cada uma*, 500 conservadores, e 70 *outras* amostras com, *também*, 500 liberais *cada uma*). Em outras palavras, cada vez que ele extrai 500 conservadores aleatoriamente ele extrai, também de forma casual, 500 liberais.



Nota: 5,0 representa a diferença entre as médias de duas amostras aleatórias, cada uma composta de 500 sujeitos.

FIGURA 8.1 Diferença Entre as Médias de Permissividade Resultantes de Amostras de Liberais e de Conservadores (População Hipotética)

* N.T.: Essa é uma forma abreviada de perguntar o seguinte: "Será que devemos aceitar a hipótese nula de que, sendo diferentes as médias das amostras, são *também* diferentes as médias das populações?

Após ter colhido suas amostras, nosso excêntrico pesquisador interroga cada sujeito de cada amostra (isto é, 1.000 X 70 = 70.000 pessoas) relativamente ao método que adota na educação dos filhos; desse levantamento, resulta um escore médio da variável "permissividade" para cada uma das amostras de liberais e *outro* escore médio (da *mesma* variável) para cada uma das amostras de conservadores. Além disso, ele calcula, *para cada par* de amostras, o escore-diferença das médias. Por exemplo, se a média de permissividade para os liberais for 7,0 e a média de permissividade para os conservadores, 6,0, então o escore-diferença será +1,0; por igual raciocínio, se a média para os liberais for 5,0 e a média para os conservadores, 8,0, o escore-diferença será −3,0. Obviamente, quanto maior o escore-diferença, mais as duas amostras diferem com respeito à característica pesquisada (isto é, permissividade na educação dos filhos). Note-se que sempre *subtraímos*, para cada amostra, *o escore médio dos conservadores do escore médio dos liberais.* Os 70 escores-diferenças (resultantes das diferenças entre as médias dos 70 pares de amostras) obtidos pelo nosso pesquisador foram ilustrados na Figura 8.2.

Vamos supor que saibamos que a população de conservadores e a de liberais na verdade não diferem absolutamente com respeito à permissividade nos métodos de criar filhos. Digamos que $M = 5{,}0$ em ambas as populações. Se admitirmos que a hipótese nula está correta e que liberais e conservadores são idênticos nesse aspecto, poderemos usar as 70 diferenças de médias que o nosso pesquisador obteve para ilustrar a distribui-

Nota: cada escore representa a diferença entre (a média de) uma amostra de 500 liberais e (a média de) outra amostra de 500 conservadores.

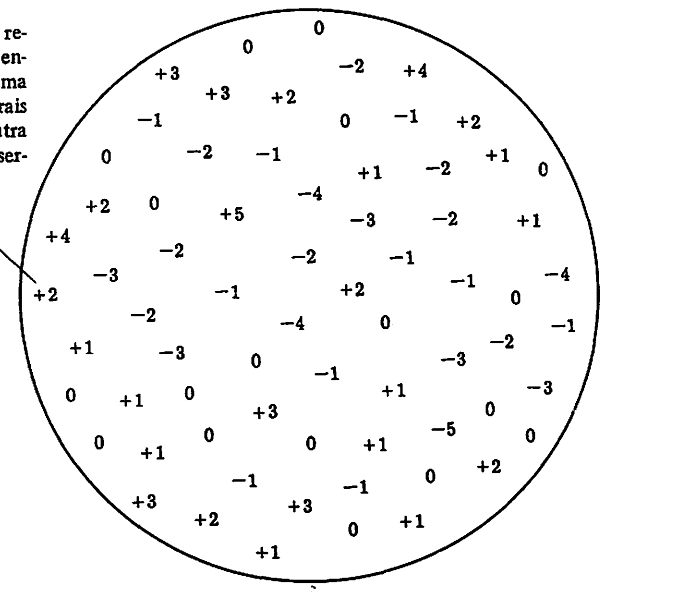


FIGURA 8.2 Setenta Escores-diferenças Representativos das Diferenças Quanto à Permissividade Entre Amostras Aleatórias de Liberais e Conservadores. (População Hipotética)

ção amostral das diferenças. Esta afirmação procede porque a distribuição amostral de diferenças parte do pressuposto de que todos os pares de amostras *só* diferem por força do *erro amostral* e não por causa de *reais* diferenças entre as populações.

As 70 diferenças entre as médias, que figuram na Figura 8.2, foram reagrupadas sob a forma de distribuição de freqüências na Tabela 8.1. Da mesma forma como tem ocorrido com outros tipos de distribuições de freqüências, essas diferenças foram colocadas em ordem decrescente, com as respectivas freqüências de ocorrência na coluna adjacente.

Com vistas a salientar as propriedades mais importantes de uma distribuição amostral de diferenças, os dados da Tabela 8.1 foram graficamente apresentados na Figura 8.3. Vemos que, como ilustra a tabela, a *distribuição amostral de diferenças entre médias amostrais tende para uma curva normal cuja média ("média das diferenças") é zero.*[1] Isso faz sentido porque as diferenças entre médias, sendo algumas positivas e outras negativas, tendem a compensar-se na distribuição (já que para cada valor negativo existe um valor positivo eqüidistante da média.

TABELA 8.1 **Distribuição Amostral das Diferenças Resultantes de 70 Pares de Amostras Aleatórias**

*Diferenças Entre Médias**	*f*
+5	1
+4	2
+3	5
+2	7
+1	10
0	18
−1	10
−2	8
−3	5
−4	3
−5	1
	N = 70

* Esses escores-diferenças incluem os valores fracionários. (Por exemplo, −5 inclui os valores que vão desde −5,0 até 5,9.)

Por ser uma curva normal, a maioria das diferenças médias amostrais desta distribuição cai nos arredores de zero—que é o ponto da distribuição que tem freqüência máxima; há relativamente poucas diferenças médias em quaisquer das caudas da distribuição. Tal fato é de se esperar, uma vez que o conjunto de todas as diferenças da distribuição resulta muito mais de erro amostral do que de reais diferenças populacionais entre conservadores e liberais. Em outras palavras, se a diferença média real entre a

[1] O que pressupõe que extraímos aleatoriamente amostras grandes de uma dada população de escores brutos.

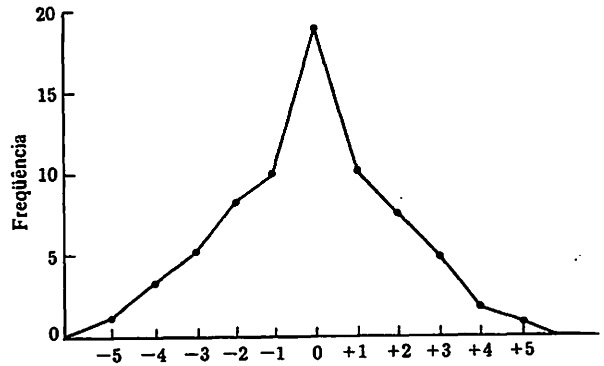


FIGURA 8.3 **Polígono de Freqüência Relativo à Distribuição Amostral de Diferenças (Ver Tabela 8.1)**

população de conservadores e a de liberais for zero, também esperamos que a média de uma distribuição amostral de diferenças seja zero.

TESTES DE HIPÓTESES RELACIONADOS COM A DISTRIBUIÇÃO DE DIFERENÇAS

Em capítulos anteriores, aprendemos a fazer afirmações probabilísticas acerca da ocorrência de escores brutos e médias amostrais. No presente caso, vamos tentar fazer afirmações de probabilidade com relação aos escores-diferenças da distribuição amostral das diferenças médias. Como já foi salientado anteriormente, essa distribuição amostral assume a forma de uma curva normal e, portanto, pode ser considerada como uma *distribuição de probabilidade*. Podemos dizer que a probabilidade decresce à medida

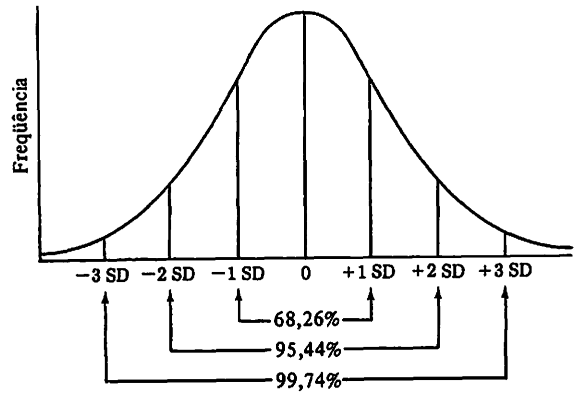


FIGURA 8.4 **Distribuição Amostral de Diferenças Entendida Como uma Distribuição de Probabilidades**

que nos afastamos da média das diferenças (zero). Mais especificamente, como ilustra a Figura 8.4, vemos que 68,26% das diferenças médias caem entre −1 DP e +1 DP, a contar de zero. Em termos probabilísticos, isso indica que $P = 0{,}68$ de que qualquer diferença entre médias amostrais cairá dentro desse intervalo. De forma análoga, podemos dizer que a probabilidade é de aproximadamente 0,95 (95 casos em 100) de que qualquer diferença de médias amostrais caia entre −2 DP e +2 DP a contar da média da distribuição, que é zero, e assim por diante.

A distribuição amostral de diferenças oferece uma sólida base para testarmos hipóteses acerca da diferença média entre duas amostras aleatórias. Suponhamos, por exemplo, que uma amostra de 100 liberais tenha um escore médio de permissividade igual a 7 e que outra amostra, também de 100 sujeitos, porém conservadores, apresente média 2. O raciocínio é o seguinte: se a diferença média que obtivemos, 5, resultante de (7 − 2 = 5), situar-se tão longe da diferença "zero" de maneira a conferir-lhe somente uma pequena *probabilidade* de ocorrência na distribuição amostral de diferenças, *rejeitamos a hipótese nula*, ou seja, a hipótese que afirma ser a diferença obtida *apenas* o resultado da *ação do acaso.** Se, por outro lado, nossa diferença média amostral cai tão perto de zero de forma que sua *probabilidade* de ocorrência é grande, devemos *aceitar a hipótese nula* e tratar a diferença obtida como resultante de erro de amostragem.

Portanto, devemos procurar determinar a que distância a diferença média obtida (neste caso, 5) fica da diferença média de zero. Ao fazer isso, devemos primeiro transformar nossa diferença obtida em unidades de desvio padrão.

Recorde-se que para traduzir *escores brutos* em unidades de desvio padrão usamos a fórmula

$$z = \frac{X - \bar{X}}{\sigma}, \text{ onde}$$

X = escore bruto
$\bar{X}$ = média da distribuição de escores brutos
σ = desvio padrão da distribuição de escores brutos.

Analogamente, podemos traduzir os escores médios de uma distribuição de médias amostrais em unidades de desvio padrão através da fórmula

$$z = \frac{\bar{X} - M}{\sigma_{\bar{X}}}, \text{ onde}$$

$\bar{X}$ = média amostral
M = média da população (média das médias)
$\sigma_{\bar{X}}$ = erro padrão da média (estimativa do desvio padrão da distribuição de médias).

* N.T.: É comum usar-se com o mesmo sentido *erro amostral* e *diferença resultante da ação do acaso.*

No contexto presente procuramos, de forma semelhante, traduzir a nossa diferença média amostral (+5) em unidades de desvio padrão, mediante a seguinte fórmula:

$$z = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{\sigma_{dif}}, \text{ onde}$$

$\bar{X}_1$ = média da primeira amostra
$\bar{X}_2$ = média da segunda amostra
"0" = "zero", isto é, o valor da média da distribuição amostral de diferenças (admitimos que $M_1 - M_2 = 0$)
σ_{dif} = desvio padrão da distribuição amostral de diferenças.

Já que o valor da média da distribuição de diferenças é sempre admitido como sendo *zero*, podemos eliminar esse valor (0) da fórmula acima (escala z) sem alterar o resultado. Portanto:

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{dif}}$$

Com relação à permissividade entre liberais e conservadores, precisamos, primeiro, traduzir nossa diferença média (obtida, calculada) no eqüivalente z. Se o desvio padrão da distribuição amostral de diferenças (σ_{dif}) for 2, obteremos o seguinte escore z:

$$z = \frac{7 - 2}{2} = \frac{5}{2} = +2{,}5$$

Portanto, a diferença média, 5, entre liberais e conservadores situa-se a 2,5 desvios padrões, a contar da média zero, na distribuição de diferenças.

Perguntamos: *qual a probabilidade de que uma diferença de 5 ou mais entre médias amostrais ocorra única e exclusivamente em função do erro amostral*? Recorrendo à Tábua B no fim do livro, verificamos que $z = 2{,}5$ representa 49,38% da distribuição *quer para a direita, quer para a esquerda* da média zero. Em outras palavras, 98,76% (isto é, 49,38% + 49,38% = 98,76%) das diferenças médias amostrais caem entre zero e uma diferença média de 5 para *ambos os lados* (a partir da média), donde resultam, como ilustra a Figura 8.5, valores positivos *e* valores negativos. Em termos probabilísticos, isto significa que existe uma $P = 0{,}99$ (99 possibilidades em 100) de que uma diferença média caia entre −5 e +5. Se subtrairmos de 100% (100% − 98,76% = 1,24%), encontraremos, por arredondamento, $P = 0{,}01$; essa é a probabilidade de que uma diferença média de 5 (ou maior que 5) entre amostras ocorra única e exclusivamente por causa do erro amostral. Em outras palavras, uma diferença média de 5 ou mais ocorre *por força do erro amostral* (e, portanto, aparece na distribuição amostral) *apenas uma vez* em cada 100 diferenças médias. Sabendo disso, não poderíamos considerar a rejeição da hipótese nula e a *aceitação da hipótese experimental*? (Lembremo-nos de que a hipótese experimental—H_a—estabelece que realmente existe uma diferença populacional

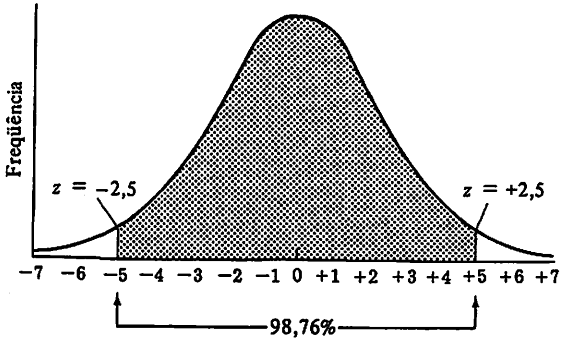


FIGURA 8.5 **Representação Gráfica da Porcentagem da Área Total na Distribuição de Diferenças Entre $z = -2{,}5$ e $z = +2{,}5$**

entre conservadores e liberais–no caso de a *variável observacional* ser a permissividade na criação de filhos.) Afinal, 1 possibilidade em 100 representa uma boa margem de risco, não é mesmo?

Dada a situação acima, a maioria de nós decidiria rejeitar a hipótese nula, *ainda que pudéssemos estar errados* ao fazê-lo. (Não esqueça de que sempre existe 1 possibilidade em 100.) Entretanto, a decisão não é sempre tão fácil. Suponha-se, por exemplo, que seja do nosso conhecimento o fato de a diferença média ocorrer, por erro amostral, 10 (isto é, $P = 0{,}10$), 15 (isto é, $P = 0{,}15$) ou 20 (isto é, $P = 0{,}20$) vezes em cada 100. Será que *ainda assim* vamos rejeitar a hipótese nula? Ou será melhor e mais "prudente" atribuir nossa diferença ao erro amostral?

Precisamos, pois, de um *ponto de referência consistente** para decidir se uma diferença entre duas médias amostrais é *grande demais* para ser explicada *apenas* pela ação do acaso. Em outras palavras, precisamos de um método que nos possibilite decidir quando um determinado resultado é *estatisticamente significante.* **

NÍVEIS DE SIGNIFICÂNCIA

Para decidir se a diferença amostral obtida é estatisticamente significante–resultado de uma *real* diferença entre as populações e não apenas produto de erro amostral–é habitual estabelecer um *nível de confiança* (também chamado *nível de significância*), nível esse que representa a probabilidade com que a hipótese nula pode ser *rejeitada* com confiança (segurança)*** ou, dizendo de outro modo, a probabilidade com que a hipótese experimental pode ser *aceita* (com confiança). Em conseqüência, decidimos pela

* N.T.: Esse *ponto de referência* é comumente designado *valor crítico.*

** N.T.: Um resultado é *estatisticamente significante* quando é *igual a* ou *maior que* o *valor crítico* (isto é, *o ponto de referência*), o qual, via de regra, encontra-se *tabelado.*

*** N.T.: Essa *confiança* ou *segurança* é sempre relativa e não constitui um problema do *estatístico*, senão do *especialista* na área em que determinada pesquisa esteja sendo feita.

rejeição da hipótese nula sempre que for muito pequena a probabilidade de que a diferença amostral tenha sua origem no erro de amostragem (por exemplo, 5 casos em 100).*

Por uma questão de convenção, usamos o *nível de confiança* (*significância*) de 0,05 (= 5%). Em termos mais simples, estamos dispostos a rejeitar a hipótese nula se a diferença amostral obtida ocorrer *por acaso* somente 5 vezes ou menos em 100 (isto é, 5% no máximo). O nível de confiança de 0,05 foi indicado graficamente na Figura 8.6. Como ela bem ilustra, este nível de confiança encontra-se nas áreas pequenas das caudas da distribuição de diferenças entre médias. Essas são as áreas sob a curva que representam uma distância de ± (mais ou menos) 1,96 desvios padrões contados a partir de zero, que é a média das diferenças.

Com vistas a entender melhor por que esse ponto particular na distribuição amostral representa o nível de confiança (significância) de 0,05, podemos recorrer à Tabela B (fim do livro) para determinar a porcentagem da freqüência total que está associada aos 1,96 desvios padrões contados a partir da média. Vemos que 1,96 desvios padrões para *ambos os lados* demarcam 2,5% das diferenças médias amostrais. (Assim: 50% − 47,5% = 2,5%.) Em outras palavras, 95% das diferenças amostrais caem entre −1,96 DPs e +1,96 DPs contados da média das diferenças (zero); somente 5% caem a partir desse ponto (1,96) (2,5% + 2,5% = 5%).

Níveis de significância podem ser estabelecidos para qualquer grau de probabilidade. Por exemplo, o nível de 0,01 é mais rigoroso, uma vez que sua adoção implica rejeitar a hipótese nula somente no caso de haver 1 possibilidade em 100 de que a dife-

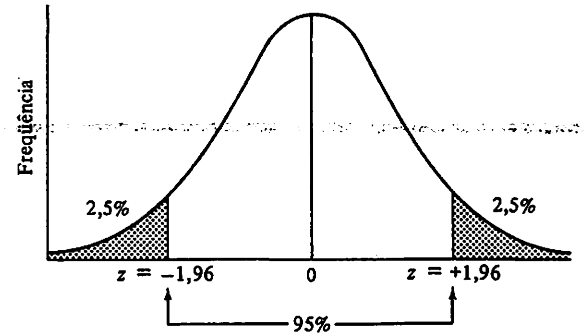


FIGURA 8.6 **Representação Gráfica do Nível de Significância (Confiança) de 0,05 (isto é, 5%)**

* N.T.: Sendo de importância capital esse aspecto da teoria da decisão estatística, não é demais recolocar a questão nos seguintes termos: uma *diferença entre médias amostrais* ocorre (1) porque existe uma *real diferença entre as médias das populações* de onde foram extraídas as amostras ou (2) porque houve *interferência do erro de amostragem.* Se *houve* erro de amostragem, tudo se passa *como se* não existisse diferença entre as médias amostrais. Esse fenômeno, porém, é relativamente *raro* e, por isso, está associada a ele uma *probabilidade muito pequena.* Em outras palavras, as reais diferenças não ocorrem *por acaso*, mas, sim, por força de um *agente específico.*

rença amostral encontrada deva-se *apenas* à ação do acaso, isto é, do erro de amostragem. O nível de 0,01 é representado pela área que se estende de −2,58 a +2,58 desvios padrões contados a partir de zero.

Os níveis de confiança não nos oferecem garantia *absoluta* quanto à correção da hipótese nula. Sempre que decidimos rejeitá-la a um certo nível de significância, expomo-nos ao risco de estar tomando a decisão errada. Rejeitar a hipótese nula quando deveríamos tê-la aceitado ocasiona o chamado *erro alfa* (ou *erro tipo I*). A probabilidade de cometer um erro alfa só existe quando rejeitamos* a hipótese nula e varia de conformidade com o nível de confiança adotado. Por exemplo, se rejeitarmos a hipótese nula, ao nível de 5%, e concluirmos que conservadores *de fato* diferem de liberais quanto aos métodos de criação de filhos, haverá 5 possibilidades em 100 de estarmos errados. Em outras palavras, $P = 0{,}05$ de que cometamos um erro alfa—e conservadores *não* diferem realmente de liberais. De modo análogo, se escolhermos o nível de confiança (significância) de 1%, só há 1 possibilidade em 100 ($P = 0{,}01$) de tomarmos a decisão errada quanto à diferença entre liberais e conservadores. É óbvio que, quanto mais *rigoroso* o nosso nível de confiança, menor a probabilidade de cometermos um erro alfa. (Note-se, também, que, quanto mais rigoroso o nível, mais na "ponta" das caudas ele se localiza.) Para exemplificar com um caso extremo, se adotarmos o nível de 0,001, o risco de ocorrência de um erro alfa fica limitado a uma vez em cada mil.

Entretanto, quanto mais nas pontas das caudas localizar-se o nível de significância, maior o risco de estarmos cometendo outro tipo de erro, conhecido por *erro beta* (ou *erro tipo II*). É assim que se designa o erro de aceitar a hipótese nula quando, na verdade, ela deveria ter sido rejeitada. O erro beta indica que a nossa hipótese experimental ainda pode estar correta, apesar da decisão de rejeitá-la em favor da hipótese nula. Uma forma de reduzir o risco de cometer este tipo de erro beta é aumentar o tamanho das amostras, de sorte que, havendo uma real diferença entre as populações, tal fato fique mais evidenciado pelo procedimento.**

Nunca podemos estar certos de que não tomamos uma decisão errada relativamente à hipótese nula, já que examinamos apenas uma amostra e não a população toda. Como não temos conhecimento dos verdadeiros valores populacionais,*** corremos o risco de cometer ou um erro tipo I ou um erro tipo II, dependendo da nossa decisão. Tal risco, na tomada de decisões estatísticas, depende da disposição do pesquisador.

ERRO PADRÃO DA DIFERENÇA

Nunca temos a possibilidade de conhecer, de imediato, o desvio padrão da distribuição de diferenças entre médias. E, tal como é o caso da distribuição amostral de médias (Capítulo 7), seria um esforço descomunal tentar extrair um número muito grande de

* N.T.: Quando rejeitamos a hipótese nula *indevidamente*.

** N.T.: Uma analogia pode ajudar a entender melhor o procedimento: quando certas partículas não são facilmente visíveis numa superfície, uma lente de aumento, porque "aumenta" toda a região por ela abrangida, realça, também, aquilo que antes era imperceptível. Em outras palavras, um teste estatístico é uma "lente" de aumento que o pesquisador aplica sobre os dados.

*** N.T.: Verdadeiros valores populacionais é o mesmo que *parâmetros*.

amostras pareadas, a fim de tornar possível o seu cálculo. Contudo, esse desvio padrão desempenha um importante papel nos testes de hipóteses relacionados com diferenças entre médias e, portanto, não pode ser ignorado.

Felizmente, existe um método simples através do qual o desvio padrão da distribuição de diferenças pode ser estimado de forma acurada com base nas duas amostras extraídas. Essa estimativa do desvio padrão da distribuição amostral de diferenças costuma ser designada *erro padrão da diferença* e simbolizada por σ_{dif}. Em símbolos:

$$\sigma_{\text{dif}} = \sqrt{\sigma_{\bar{X}_1}^2 + \sigma_{\bar{X}_2}^2}\ \text{, onde}$$

σ_{dif} = erro padrão da diferença
$\sigma_{\bar{X}_1}$ = erro padrão da primeira média amostral
$\sigma_{\bar{X}_2}$ = erro padrão da segunda média amostral.

Com o propósito de ilustrar, vamos supor que obtivemos os seguintes dados para uma amostra de 50 liberais e outra amostra de 50 conservadores:

Liberais ($N = 50$)	*Conservadores* ($N = 50$)
$\bar{X} = 7{,}0$	$\bar{X} = 6{,}0$
$s = 2{,}0$	$s = 1{,}5$

A fim de calcular o erro padrão da diferença, precisamos, primeiro, encontrar o erro padrão para cada média amostral. Recorde-se que isso se faz a partir do desvio padrão de cada amostra (ver Capítulo 7):

$$\sigma_{\bar{X}_1} = \frac{s_1}{\sqrt{N_1 - 1}} = \frac{2{,}0}{\sqrt{50 - 1}} = \frac{2{,}0}{7{,}0} = 0{,}29$$

$$\sigma_{\bar{X}_2} = \frac{s_2}{\sqrt{N_2 - 1}} = \frac{1{,}5}{\sqrt{50 - 1}} = \frac{1{,}5}{7{,}0} = 0{,}21$$

Uma vez que conheçamos o $\sigma_{\bar{X}}$ para cada média amostral, podemos obter o σ_{dif} como se segue:

$$\sigma_{\text{dif}} = \sqrt{\sigma_{\bar{X}_1}^2 + \sigma_{\bar{X}_2}^2} = \sqrt{0{,}29^2 + 0{,}21^2} = \sqrt{0{,}08 + 0{,}04} = \sqrt{0{,}12}$$

$$= 0{,}35$$

O erro padrão da diferença (nossa estimativa do desvio padrão da distribuição de diferenças) é, pois, 0,35. Se estivéssemos testando a diferença entre liberais ($\bar{X} = 7{,}0$) e conservadores ($\bar{X} = 6{,}0$) com respeito à variável "permissividade", poderíamos usar nosso resultado (0,35) a fim de traduzir a diferença média amostral (obtida) no seu equivalente z (isto é, em escala z):

$$z = \frac{\overline{X}_1 - \overline{X}_2}{\sigma_{\text{dif}}} = \frac{7-6}{0{,}35} = \frac{1}{0{,}35} = 2{,}86$$

Voltando à Tabela B (fim do livro), verificamos que um escore z igual a 2,86 representa exatamente 49,79% das diferenças médias para *cada* lado, ou 99,58% das diferenças médias para *ambos* os lados da curva a partir da média das diferenças, que é zero. (Em outras palavras, 49,79% + 49,79% = 99,58%.) Se subtrairmos essa soma de 100%, veremos que menos de 1% (0,42%) dos escores-diferenças entre médias possui um valor igual ou superior a 1. Portanto, P é menor que 0,01 de obtermos uma diferença média igual a 1 por puro acaso (= erro de amostragem). Podemos rejeitar a hipótese nula tanto ao nível de significância de 0,05 quanto ao de 0,01 dependendo, apenas, do critério estabelecido para o nosso estudo.*

Uma Ilustração

Com vistas a dar ao leitor uma ilustração passo a passo do procedimento (acima) que permite testar a significância da diferença entre duas médias amostrais, vamos imaginar que, ao nível de 0,05, quiséssemos pôr à prova a seguinte hipótese: do ponto de vista etnocêntrico, não há diferença entre homens e mulheres ($M_1 = M_2$). Nossa hipótese experimental é que as mulheres diferem dos homens com respeito ao etnocentrismo[2] ($M_1 \neq M_2$). Para testar essa hipótese, vamos imaginar que "aplicamos" um instrumento destinado a avaliar a variável "etnocentrismo" a uma amostra aleatória de 35 homens e a *outra* amostra, também aleatória, de 35 mulheres, obtendo os seguintes escores para cada uma delas (X = escores etnocêntricos. A escala vai de 1 a 5 e quanto maior o valor maior o grau de etnocentrismo.)

Homens ($N = 35$)		*Mulheres* ($N = 35$)	
X_1	X^2	X_2	X^2
1	1	1	1
1	1	1	1
1	1	2	4
2	4	1	1
1	1	1	1
1	1	1	1
3	9	3	9
3	9	1	1

[2] "Etnocentrismo" refere-se à tendência de avaliar todos os grupos de pessoas pela utilização de nossos próprios padrões culturais.

* N.T.: É prática não só usual como também "cientificamente saudável" fixar o nível de significância *antes* da realização do experimento e da coleta de dados. Tal prática *impede* o pesquisador de fazer-se "concessões", caso os resultados finais não abonem a sua hipótese experimental.

Homens ($N = 35$)		*Mulheres* ($N = 35$)	
X_1	X^2	X_2	X^2
1	1	1	1
1	1	2	4
2	4	4	16
1	1	1	1
2	4	1	1
1	1	1	1
1	1	1	1
1	1	5	25
1	1	1	1
2	4	2	4
4	16	2	4
5	25	1	1
1	1	1	1
1	1	1	1
2	4	1	1
1	1	2	4
2	4	3	9
1	1	1	1
2	4	1	1
1	1	1	1
1	1	2	4
1	1	2	4
1	1	2	4
3	9	1	1
3	9	1	1
1	1	1	1
4	16	1	1
$\Sigma X = 60$	$\Sigma X^2 = 142$	$\Sigma X = 54$	$\Sigma X^2 = 114$

PASSO 1: Achar a média aritmética de cada amostra.

$$\overline{X}_1 = \frac{\Sigma X_1}{N} = \frac{60}{35} = 1{,}71$$

$$\overline{X}_2 = \frac{\Sigma X_2}{N} = \frac{54}{35} = 1{,}54$$

PASSO 2: Achar o desvio padrão de cada amostra.

$$s_1 = \sqrt{\frac{\Sigma X^2}{N} - \overline{X}^2} = \sqrt{\frac{142}{35} - 2{,}92} = \sqrt{4{,}06 - 2{,}92} = \sqrt{1{,}14} = 1{,}07$$

$$s_2 = \sqrt{\frac{\Sigma X^2}{N} - \overline{X}^2} = \sqrt{\frac{114}{35} - 2{,}37} = \sqrt{3{,}26 - 2{,}37} = \sqrt{0{,}89} = 0{,}94$$

PASSO 3: Achar o erro padrão de cada média.

$$\sigma_{\bar{X}_1} = \frac{s_1}{\sqrt{N-1}} = \frac{1{,}07}{\sqrt{34}} = \frac{1{,}07}{5{,}83} = 0{,}18$$

$$\sigma_{\bar{X}_2} = \frac{s_2}{\sqrt{N-1}} = \frac{0{,}94}{\sqrt{34}} = \frac{0{,}94}{5{,}83} = 0{,}16$$

PASSO 4: Achar o erro padrão da diferença.

$$\sigma_{\text{dif}} = \sqrt{\sigma_{\bar{X}_1}^2 + \sigma_{\bar{X}_2}^2} = \sqrt{(0{,}18)^2 + (0{,}16)^2} = \sqrt{0{,}03 + 0{,}03} = \sqrt{0{,}06}$$

$$= 0{,}25$$

PASSO 5: Traduzir a diferença média amostral em unidades de erro padrão da diferença.

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\text{dif}}} = \frac{1{,}71 - 1{,}54}{0{,}25} = \frac{0{,}17}{0{,}25} = 0{,}68$$

PASSO 6: Achar a porcentagem da área total sob a curva normal entre z e a média das diferenças (0). (Ver Tabela B.)

$$\begin{array}{r} 25{,}17\% \\ +25{,}17\% \\ \hline 50{,}34\% \end{array}$$

PASSO 7: Subtrair de 100% para achar a porcentagem da área total associada à diferença média amostral obtida.

$$\begin{array}{r} 100{,}00\% \\ -50{,}34\% \\ \hline 49{,}66\% \end{array}$$

Do resultado do Passo 7, acima, vemos que, por arredondamento, há $P = 0{,}50$ de obtermos uma diferença média de 0,17 (isto é, $1{,}71 - 1{,}54 = 0{,}17$) por mero erro amostral. Como conseqüência, devemos aceitar a hipótese nula e rejeitar a hipótese experimental ao nível de significância de 0,05. A probabilidade de ocorrência da nossa diferença média obtida entre homens e mulheres é *maior* que 5% (ou seja, 5 em 100). Para sermos exatos, a probabilidade é de 50 em 100! Conclusão: nossos dados amostrais não indicam que, relativamente à variável "etnocentrismo", haja diferença significativa entre homens e mulheres.

COMPARAÇÕES ENTRE AMOSTRAS PEQUENAS

É comum que pesquisadores, principalmente no campo social, trabalhem com amostras

que contêm pequeno número de respondentes (ou de dados). Do ponto de vista prático, uma amostra pode ser considerada *pequena* se contiver, por exemplo, *menos* de 30 sujeitos. Embora os resultados baseados em amostras pequenas possam ser convenientes, se não necessários, dada a dificuldade de colher amostras maiores, eles podem comprometer de forma séria as conclusões, se sua interpretação for associada à área sob a curva normal (Tabela B). Esta afirmação é verdadeira porque a distribuição amostral de diferenças só assume a forma da curva normal se as amostras que a compõem forem grandes. Um pesquisador que esteja trabalhando com 5, 10 ou 20 respondentes por amostra não tem condições de satisfazer tal exigência. Como resultado, ele não pode usar escores z baseados na distribuição normal.

Esse "afastamento da normalidade" na distribuição de diferenças pode ser compensado estatisticamente mediante o uso do que se convencionou chamar *razão t** (ou *estatística t*). Tal como ocorre com o escore z, a razão t pode ser usada para traduzir uma diferença média amostral em unidades de erro padrão da diferença. Ainda com a escala z como pano de fundo, a razão t pode ser analogamente obtida da seguinte forma: calcular a diferença entre as médias das amostras e dividi-la pelo erro padrão da diferença. Em símbolos:

$$t = \frac{X_1 - X_2}{\sigma_{\text{dif}}}, \text{ onde}$$

$\bar{X}_1$ = média da primeira amostra
$\bar{X}_2$ = média da segunda amostra
σ_{dif} = erro padrão da diferença.

Como pode ser observado acima, a fórmula da razão t é idêntica à usada para o cálculo dos escores z, aprendida anteriormente. Porém, ao contrário do que ocorre com a estatística z, a estatística t precisa ser interpretada com referência a *graus de liberdade*[3] (gl), graus esses que não só variam com o tamanho da amostra mas que são responsáveis diretos pelo formato da distribuição amostral de diferenças. Quanto maior a amostra, maior o número de graus de liberdade. Quanto maior o número de graus de liberdade, maior a aproximação da distribuição de diferenças à curva normal. Com infinitos graus de liberdade, a razão t transforma-se num escore z e, aí, podemos empregar a Tabela B para interpretar nossos resultados.

Mas que ocorre quando trabalhamos com amostras pequenas? Como é que resolvemos o problema de encontrar os graus de liberdade e interpretar nossa razão t? Quando esta última estiver sendo usada na comparação de duas médias amostrais, o número de graus de liberdade pode ser calculado pela aplicação da fórmula

$$\text{gl} = N_1 + N_2 - 2, \text{ onde}$$

[3] Graus de liberdade, tecnicamente, referem-se à liberdade de variação num conjunto de escores. Se tivermos uma amostra de 6 escores, então 5 deles têm "liberdade" de variar, enquanto que 1 e apenas 1 é fixo (quanto ao valor). Portanto, numa amostra simples de tamanho 6 (por exemplo, 6 respondentes), gl $= N - 1$ ou 5.

* Fala-se também em t de Student.

N_1 = tamanho da primeira amostra
N_2 = tamanho da segunda amostra.

Portanto, se estivermos comparando uma amostra de 6 liberais com outra de 8 conservadores, nossos graus de liberdade serão: $6 + 8 - 2 = 12$.

Com o auxílio da Tabela C (fim do livro) e do número (calculado) de graus de liberdade, podemos interpretar qualquer valor obtido de t. A Tabela C registra os valores de t necessários para a rejeição da hipótese nula aos níveis de significância de 0,05 e 0,01, associados a diferentes graus de liberdade. Ao examinar a Tabela C, vemos uma coluna encimada por gl (graus de liberdade) e uma lista de valores t para cada grau de liberdade (p. ex., 1, 2, 3, ..., 30, ..., 40, ... 100, ..., ∞), aos níveis de significância de 0,05 e 0,01. Como veremos adiante, esses valores t podem ser usados na interpretação do nosso t calculado (também chamado observado).

Comparação entre Amostras Pequenas: Uma Ilustração

Para ilustrar o uso da estatística t, dos graus de liberdade e da Tabela C no teste da diferença entre médias de duas amostras pequenas, considere-se a seguinte situação: certo pesquisador decide testar se o "comportamento caridoso" varia em função do "anonimato" ou "não-anonimato" do doador. Daí:

Hipótese Nula: ($M_1 = M_2$) *O grau de "comportamento caridoso" não varia em função do "anonimato" ou "não-anonimato" do doador.*

Hipótese Experimental: ($M_1 \neq M_2$) *O grau de "comportamento caridoso" varia conforme seja o doador mantido no "anonimato" ou no "não-anonimato".*

Para testar essa hipótese, o pesquisador estabelece 5% como nível de significância; por outras palavras, ele decide, desde o início, rejeitar a hipótese nula *somente se* a probabilidade for de 5 em 100* de que a diferença média amostral resulte de erro de amostragem (isto é, por ação do acaso). Depois de ter estabelecido esse critério de significância, ele colhe duas amostras aleatórias de doadores potenciais. A todos os membros de ambas as amostras ele solicita donativos, em dinheiro, para distribuição aos sobreviventes de um terremoto de grandes proporções. Aos 6 membros da primeira amostra ele dá a garantia do anonimato; aos 6 membros da segunda amostra ele promete tornar públicos os seus nomes. Temos, aí, as condições experimentais de *anonimato* versus *identidade conhecida.*

As quantias** doadas pelos membros de ambas as amostras figuram no quadro abaixo:

* N.T.: Cumpre salientar que o nível de significância de 5%, tal como foi estabelecido, é a ***probabilidade máxima tolerada***, mas ***nada impede*** que esse nível seja ***efetivamente menor*** que o estabelecido. ***Maior é que não pode!***

** N.T.: No texto original, as quantias figuram em dólares (US$). Tendo em mente aproveitar o exemplo, que foi muito bem concebido, a unidade monetária (US$) foi ***eliminada***, de sorte que 1, 2, 3, 4 e 5 podem significar ***qualquer quantia***, em ***qualquer moeda***, desde que seja respeitada a paridade cambial.

Anonimato ($N = 6$)		*Identidade Conhecida* ($N = 6$)	
X_1	X^2	X_2	X^2
1	1	3	9
2	4	5	25
1	1	5	25
1	1	5	25
2	4	4	16
1	1	5	25
$\Sigma X = 8$	$\Sigma X^2 = 12$	$\Sigma X = 27$	$\Sigma X^2 = 125$

Observamos que os 6 membros da amostra "anônima" deram, no total, 8, enquanto que os 6 membros da amostra "não-anônima", 27. O procedimento passo a passo que se segue pode ser usado para testarmos se a diferença obtida é estatisticamente significante.

PASSO 1: Achar a média de cada amostra.

$$\overline{X}_1 = \frac{\Sigma X_1}{N} = \frac{8}{6} = 1{,}33$$

$$\overline{X}_2 = \frac{\Sigma X_2}{N} = \frac{27}{6} = 4{,}50$$

PASSO 2: Achar o desvio padrão de cada amostra.

$$s_1 = \sqrt{\frac{\Sigma X^2}{N} - \overline{X}^2} = \sqrt{\frac{12}{6} - (1{,}33)^2} = \sqrt{2{,}00 - 1{,}77} = \sqrt{0{,}23} = 0{,}48$$

$$s_2 = \sqrt{\frac{\Sigma X^2}{N} - \overline{X}^2} = \sqrt{\frac{125}{6} - (4{,}50)^2} = \sqrt{20{,}83 - 20{,}25} = \sqrt{58} = 0{,}76$$

PASSO 3: Achar o erro padrão de cada média.

$$\sigma_{\overline{X}_1} = \frac{s}{\sqrt{N-1}} = \frac{0{,}48}{\sqrt{5}} = \frac{0{,}48}{2{,}24} = 0{,}21$$

$$\sigma_{\overline{X}_2} = \frac{s}{\sqrt{N-1}} = \frac{0{,}76}{\sqrt{5}} = \frac{0{,}76}{2{,}24} = 0{,}34$$

PASSO 4: Achar o erro padrão da diferença.

$$\sigma_{\text{dif}} = \sqrt{\sigma_{\overline{X}_1}^2 + \sigma_{\overline{X}_2}^2} = \sqrt{(0{,}21)^2 + (0{,}34)^2} = \sqrt{0{,}04 + 0{,}12} = \sqrt{0{,}16} = 0{,}40$$

PASSO 5: Traduzir a diferença média amostral em unidades de erro padrão da diferença.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\text{dif}}} = \frac{1,33 - 4,50}{0,40} = -\frac{3,17}{0,40} = -7,93$$

PASSO 6: Achar o número de graus de liberdade.

$$gl = N_1 + N_2 - 2 = 6 + 6 - 2 = 10$$

PASSO 7: Comparar a razão t (obtida, calculada) com o t adequado (que figura na Tabela C).

$$\begin{aligned} t_0 &= t \text{ observado (obtido)} &= 7,93 \\ tc &= t \text{ crítico (tabelado)} &= 2,228 \\ gl &&= 10 \\ P &&= 0,05 \end{aligned}$$

Como bem evidencia o passo 7, para rejeitarmos a hipótese nula ao nível de confiança (significância) de 0,05, com 10 graus de liberdade, nosso t observado (t_0) deve ser igual a 2,228 ou maior. No caso em foco, obtivemos um $t = 7,93$. Portanto, rejeitamos a hipótese nula e aceitamos a hipótese experimental. O grau de "comportamento caridoso" varia, na realidade, em função do anonimato e do não-anonimato. Para ser mais específico, a condição "identidade conhecida" produz, significantemente "mais caridade" ($\bar{X} = 4,50$) do que a condição "anonimato" ($\bar{X} = 1,33$).

COMPARAÇÕES ENTRE AMOSTRAS DE TAMANHOS DIFERENTES

Até agora, temos trabalhado com amostras que contêm exatamente o mesmo número de respondentes ou de dados. Na ilustração anterior, por exemplo, cada amostra continha 6 respondentes. Porém, quando, de fato, estamos realizando uma pesquisa, não raro ocorre que nossas amostras diferem quanto ao tamanho. Poderíamos, assim, ter uma amostra de 50 liberais e outra de 64 conservadores ou uma amostra de 15 homens e outra de 22 mulheres. Para que seja possível fazer comparações entre amostras de tamanhos diferentes, precisamos encontrar uma forma de atribuir *peso* adequado à influência de cada amostra. No caso de $\bar{X}$, isso ocorre automaticamente, uma vez que sempre dividimos ΣX por N. Tal, porém, não se aplica ao erro padrão da diferença: cada desvio padrão amostral—em que se baseia o σ_{dif}—"contribui" da mesma maneira na fórmula aprendida páginas atrás, muito embora diferenças grandes e importantes, ligadas ao tamanho da amostra, possam estar presentes.

Esse problema pode ser contornado mediante o emprego da fórmula do erro padrão da diferença, em que a influência relativa de cada desvio padrão pode ser ponderada em termos do respectivo tamanho da amostra. Essa fórmula figura a seguir:

$$\sigma_{\text{dif}} = \sqrt{\left(\frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2 - 2}\right)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}$$

onde

s_1 = desvio padrão da primeira amostra
s_2 = desvio padrão da segunda amostra
N_1 = tamanho da primeira amostra
N_2 = tamanho da segunda amostra.

Para ilustrar passo a passo o procedimento que permite comparar amostras de tamanhos diferentes, considere-se a hipótese de que crianças negras e crianças brancas de certa região urbana diferem quanto à tendência para a criminalidade. Nesse caso, temos:

Hipótese Nula: ($M_1 = M_2$) *Crianças negras não diferem de crianças brancas com relação à variável observacional "tendência para a criminalidade".*

Hipótese Experimental: ($M_1 \neq M_2$) *Crianças negras diferem de crianças brancas com relação à variável observacional "tendência para a criminalidade".*

Para testar essa hipótese ao nível (de confiança/significância) de 0,05, imagine-se que certo pesquisador tenha submetido duas amostras aleatórias, de tamanhos diferentes, a um instrumento de medida capaz de fornecer informações numéricas (dados) relacionadas com a "tendência para a criminalidade". A amostra de crianças brancas era de tamanho 4 e a de negras, 7. Nessa escala, o valor 1 representa *baixa* tendência para a criminalidade e o valor 5, *forte* tendência. Os escores finais constam do quadro abaixo:

Brancas ($N = 4$)		*Negras* ($N = 7$)	
X_1	X^2	X_2	X^2
1	1	4	16
2	4	1	1
1	1	1	1
3	9	1	1
$\Sigma X_1 = 7$	$\Sigma X^2 = 15$	2	4
		2	4
		1	1
		$\Sigma X_2 = 12$	$\Sigma X^2 = 28$

O procedimento para testar essa hipótese pode ser ilustrado como se segue:

PASSO 1: Calcular a média de cada amostra.

$$\bar{X}_1 = \frac{\Sigma X_1}{N} = \frac{7}{4} = 1,75$$

$$\bar{X}_2 = \frac{\Sigma X_2}{N} = \frac{12}{7} = 1,71$$

PASSO 2: Calcular o desvio padrão de cada amostra.

$$s_1 = \sqrt{\frac{\Sigma X^2}{N} - \overline{X}^2} = \sqrt{\frac{15}{4} - 3{,}06} = \sqrt{3{,}75 - 3{,}06} = \sqrt{0{,}69} = 0{,}83$$

$$s_2 = \sqrt{\frac{\Sigma X^2}{N} - \overline{X}^2} = \sqrt{\frac{28}{7} - 2{,}92} = \sqrt{4{,}00 - 2{,}92} = \sqrt{1{,}08} = 1{,}04$$

PASSO 3: Calcular o erro padrão da diferença.

$$\sigma_{\text{dif}} = \sqrt{\left(\frac{Ns_1^2 + Ns_2^2}{N_1 + N_2 - 2}\right)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)} = \sqrt{\left(\frac{4(0{,}83)^2 + 7(1{,}04)^2}{4 + 7 - 2}\right)\left(\frac{1}{4} + \frac{1}{7}\right)}$$

$$= \sqrt{\left(\frac{2{,}76 + 7{,}56}{9}\right)(0{,}25 + 0{,}14)} = \sqrt{\left(\frac{10{,}32}{9}\right)(0{,}39)} = \sqrt{(1{,}15)(0{,}39)}$$

$$= \sqrt{0{,}45} = 0{,}67$$

PASSO 4: Traduzir a diferença média amostral em unidades de erro padrão da diferença.

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sigma_{\text{dif}}} = \frac{1{,}75 - 1{,}71}{0{,}67} = \frac{0{,}04}{0{,}67} = 0{,}06$$

PASSO 5: Achar o número de graus de liberdade.

$$gl = N_1 + N_2 - 2 = 4 + 7 - 2 = 9$$

PASSO 6: Comparar a razão t obtida com a razão t tabelada (Tabela C).

razão t obtida (t observado)	=	0,06
razão t tabelada (t crítico)	=	2,262
gl	=	9
P	=	0,05

Como indica o passo 6, para rejeitarmos a hipótese nula ao nível de 0,05, com 9 graus de liberdade, nosso t calculado deve ser igual a 2,262 ou maior. Posto que obtivemos um t calculado de apenas 0,06, devemos aceitar a hipótese nula e rejeitar a hipótese experimental. Nossos resultados não sustentam a opinião de que crianças negras e crianças brancas diferem significativamente com respeito à variável "tendência para a criminalidade".

COMPARAÇÃO DE DADOS RESULTANTES DE DUAS MENSURAÇÕES TEMPORALMENTE DISTINTAS DA MESMA AMOSTRA

Até aqui, discutimos comparações entre duas amostras independentes, tais como "homens vs. mulheres", "negros vs. brancos", "liberais vs. conservadores". Antes de passarmos para outro tópico, precisamos introduzir uma última variação de procedimento estatístico para a comparação de duas médias. Estamos falando do procedimento chamado *painel* ou *antes-e-depois*: trata-se de mensurar a *mesma* amostra em dois momentos distintos (momento 1 vs. momento 2). Por exemplo, um pesquisador pode querer medir a variável "hostilidade" numa mesma amostra de crianças em dois momentos distintos: *antes* e *depois* de terem assistido a um particular programa de televisão. Da mesma forma, poderíamos querer medir "diferenças atitudinais" com referência a um determinado candidato em dois momentos também distintos: *antes* e *depois* da campanha eleitoral.

Para ilustrar, passo a passo, esse procedimento comparativo (antes-e-depois), vamos supor que diversos sujeitos tenham sido desapropriados e forçados a mudar de residência em virtude da construção de um novo elevado. Como pesquisadores sociais, temos interesse em determinar o impacto da mobilidade residencial forçada sobre os sentimentos de "vizinhança" (isto é, se havia, por parte dos *mesmos* sujeitos, predisposição favorável para com os vizinhos *antes* da mudança e se essa predisposição sofreu alteração depois dela). Nesse caso, então, M_1 é o escore médio de "aceitação dos vizinhos" no momento 1 (*antes* da mudança) e M_2 o escore médio de "aceitação dos vizinhos" (outros, naturalmente) no momento 2 (*depois* da mudança). Portanto:

Hipótese Nula: ($M_1 = M_2$) *O grau de "aceitação dos vizinhos" não sofre influência da variável "mudança forçada".*

Hipótese Experimental: ($M_1 \neq M_2$) *O grau de "aceitação dos vizinhos" sofre influência da variável "mudança forçada".*

Para testar o impacto da mudança forçada sobre os sentimentos de "vizinhança", entrevistamos uma amostra aleatória* de 6 sujeitos, *antes* e *depois* da mudança. De nossas entrevistas resultam os seguintes escores de "aceitação dos vizinhos" (escores mais altos, numa escala de 1 a 4, indicam maior aceitação):

Respondente	*Antes da Mudança* X_1	*Depois da Mudança* X_2	*Diferença* $X_1 - X_2 = D$	*(Diferença)²* D^2
Estêvão	2	1	1	1
Mírian	1	2	−1	1
Carolina	3	1	2	4
Linda	3	1	2	4
Alan	1	2	−1	1
Davi	4	1	3	9
	$\Sigma X_1 = 14$	$\Sigma X_2 = 8$		$\Sigma D^2 = 20$

* N.T.: A casualidade da amostra refere-se *apenas* ao momento 1; no momento 2, os sujeitos entrevistados vão ser *obrigatoriamente os mesmos.*

O quadro acima demonstra que o foco da comparação do tipo antes-e-depois está na *diferença* entre os momentos 1 e 2, como bem reflete a fórmula para o cálculo do desvio padrão (da distribuição dos escores-diferenças antes/depois):

$$s = \sqrt{\frac{\Sigma D^2}{N} - (\overline{X}_1 - \overline{X}_2)^2}$$

Nessa fórmula,

s = desvio padrão da distribuição de escores-diferenças antes/depois
D = diferença resultante da subtração do "escore depois" do "escore antes"
N = tamanho da amostra (número de respondentes).

PASSO 1: Calcular a média de cada momento.

$$\overline{X}_1 = \frac{\Sigma X_1}{N} = \frac{14}{6} = 2{,}33$$

$$\overline{X}_2 = \frac{\Sigma X_2}{N} = \frac{8}{6} = 1{,}33$$

PASSO 2: Calcular o desvio padrão da diferença entre os escores colhidos nos momentos 1 e 2.

$$s = \sqrt{\frac{\Sigma D^2}{N} - (\overline{X}_1 - \overline{X}_2)^2} = \sqrt{\frac{20}{6} - (2{,}33 - 1{,}33)^2} = \sqrt{\frac{20}{6} - 1{,}00}$$

$$= \sqrt{3{,}33 - 1{,}00} = \sqrt{2{,}33} = 1{,}53$$

PASSO 3: Calcular o erro padrão da diferença.

$$\sigma_{\text{dif}} = \frac{s}{\sqrt{N-1}} = \frac{1{,}53}{\sqrt{6-1}} \quad \frac{1{,}53}{2{,}24} = 0{,}68$$

PASSO 4: Transformar a diferença média amostral em unidades de erro padrão da diferença.

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sigma_{\text{dif}}} = \frac{2{,}33 - 1{,}33}{0{,}68} = \frac{1{,}00}{0{,}68} = 1{,}47$$

PASSO 5: Achar o número de graus de liberdade.

$$\text{gl} = N - 1 = 6 - 1 = 5$$

Nota: N refere-se ao *tamanho da amostra* e não ao número de escores já que há 2 por respondente.

PASSO 6: Comparar a razão t calculada com a razão t tabelada (Tabela C).

razão t calculada (t observado) = 1,47
razão t tabelada (t crítico) = 2,571
gl = 5
P = 0,05

A fim de que seja possível rejeitar a hipótese nula ao nível de 0,05, com 5 graus de liberdade, devemos obter um t observado de 2,571. Como o nosso t observado é igual a 1,47–menos do que o t crítico– aceitamos a hipótese nula e rejeitamos a hipótese experimental. A diferença amostral obtida relativamente a "sentimentos de vizinhança" antes e depois da mudança é fruto, na verdade, de puro erro amostral.

REQUISITOS PARA O USO DO ESCORE z E DA RAZÃO t

Como teremos a oportunidade de ver ao longo de todo o restante do livro, todos os testes estatísticos, para poderem ser usados, devem preencher certos requisitos. Quando mais não for, pelo menos deve ser possível *admitir* certas condições. O emprego de um teste inadequado pode confundir e levar o pesquisador a falsas conclusões. Por isso, os seguintes requisitos devem ser rigorosamente observados sempre que cogitemos de usar os escores z ou a razão t como testes de significância:

1. *Comparação entre duas médias.* O escore z e a razão t são empregados a fim de possibilitar: (a) comparação entre duas médias de amostras (aleatórias) independentes; (b) comparação entre as médias de dois conjuntos de escores relativos à mesma amostra, porém colhidos em momentos distintos.
2. *Dados Intervalares.* Admitimos, aqui, que os nossos dados pertencem ao nível intervalar de mensuração. Portanto, não podemos usar escores z ou razão t com dados ordinais ou dados que só possam ser categorizados, como é o que ocorre no nível nominal de mensuração (ver Capítulo 1).
3. *Amostragem aleatória (casual).* As nossas amostras devem ser sempre colhidas aleatoriamente.
4. *Distribuição normal.* O emprego da razão t em amostras pequenas pressupõe que a variável observacional (na amostra) tenha distribuição normal na população. O escore z, em amostras grandes, não é muito afetado se esse requisito não for respeitado. Geralmente, não temos 100% de certeza de que existe normalidade de distribuição. Não havendo razão para supor o contrário, muitos pesquisadores admitem pragmaticamente que a variável observacional tenha distribuição normal. Entretanto, se o pesquisador tiver boas razões para suspeitar de possível não-normalidade, aconselhamo-lo a não dar preferência ao teste t (ver Capítulo 6).

RESUMO

Este capítulo ocupou-se dos testes de hipóteses relacionadas com diferenças entre médias amostrais. Com esse propósito em mente, descrevemos e ilustramos uma distribuição de diferenças entre médias amostrais como uma distribuição de probabilidades. Com o auxílio dessa distribuição e do erro padrão da diferença, foi possível fazer afirmações de probabilidade acerca da diferença média; a partir daí, foi também possível rejeitar ou aceitar uma hipótese nula a um específico nível de confiança (ou nível de significância). Além disso, vimos que uma estatística t (razão t), associada a graus de liberdade, podia ser usada para testar hipóteses relacionadas com diferenças entre pequenas amostras e entre amostras de tamanhos diferentes e, ainda, para testar a mesma amostra quando mensurada em dois momentos distintos. A adequação da estatística t depende de certos requisitos, tais como: (1) comparação entre duas médias; (2) dados intervalares;* (3) amostragem aleatória e (4) distribuição normal dos dados.

PROBLEMAS

1. Pesquisadores sociais procuraram testar a hipótese de que os jornais lidos pela classe baixa são igualmente "dirigidos" para temas sexuais quanto os lidos pela classe média. Empregando um "índice de sexualidade", eles coligiram dados de uma amostra aleatória composta de 40 artigos publicados em revistas de classe média e *outra* amostra, também aleatória, com *outros* 40 artigos publicados em revistas de classe baixa. Enquanto que a amostra da classe média apresentou uma média de sexualidade igual a 3,0, com um desvio padrão de 1,5, a amostra da classe mais baixa apresentou média 4,0 e desvio padrão igual a 2,0. (Critério: maiores escores médios indicam maior grau de sexualidade.) Com os dados acima, testar a hipótese de que não há diferença (significativa), relativamente à sexualidade, entre os conteúdos das publicações dirigidas a essas duas classes sócio-econômicas. O que é que indicam os seus resultados?
2. Dois grupos de alunos fizeram exames finais de Estatística. Um desses grupos recebeu preparação formal para esse exame, enquanto que o outro apenas leu os textos básicos, sem, entretanto, comparecer às aulas. O primeiro grupo (o que assistiu às aulas) conseguiu as seguintes notas: 2, 2, 3 e 4; o segundo grupo (o que não assistiu às aulas) obteve as seguintes notas: 1, 1, 2 e 3. Testar a hipótese nula de que não há diferença, com respeito aos escores de exame, entre o primeiro grupo de alunos e o segundo. Que é que seus resultados indicam? (*Observação:* a escala de notas ia de 1 a 10, sendo que os maiores valores atribuía-se maior conhecimento de Estatística.)
3. Dadas as duas seguintes amostras aleatórias de escores, testar a significância** da diferença entre suas médias:

* N.T.: Melhor seria dizer *no mínimo* "dados intervalares". Com efeito, nada impede que, com "dados racionais", faça-se uso da estatística t.

** N.T.: *Testar a significância* significa verificar se o *t calculado* ou *observado* (t_0) é *igual a* ou *maior que* o *t crítico* (t_c). Lembrar que se $t_0 \geqslant t_c \rightarrow H_0$ rejeitada
e se $t_0 < t_c \rightarrow H_0$ não rejeitada.

Amostra 1	*Amostra 2*
8	1
3	5
1	8
7	3
7	2
6	1
8	2

4. Dadas as duas seguintes amostras aleatórias de escores, testar a significância da diferença entre suas médias:

Amostra 1	*Amostra 2*
6	6
6	5
8	7
7	7
5	3
4	3
8	5
7	6
7	3

5. Testar a significância da diferença entre as médias das seguintes amostras aleatórias de escores:

Amostra 1	*Amostra 2*
15	10
18	11
12	12
17	10
19	10

6. Testar a significância da diferença entre as médias das seguintes amostras aleatórias de escores:

Amostra 1	*Amostra 2*
1	2
1	2
2	4
3	2
3	2

7. Testar a significância da diferença entre as médias das seguintes amostras aleatórias de escores:

Amostra 1	*Amostra 2*
5	10
7	7
7	9
3	9
6	7
5	8
4	
6	
7	

8. Testar a significância da diferença entre as médias das seguintes amostras aleatórias de escores:

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

9. Testar se a diferença entre as médias dos escores que compõem as duas seguintes amostras aleatórias é significante:

Amostra 1	*Amostra 2*
10	10
4	10
1	8
2	7
4	
8	
3	
5	

10. Antes e depois de assistirem a um filme, cujo objetivo era aliviar a discriminação contra grupos minoritários, seis estudantes foram testados. A variável observacional era "atitude para com judeus", e quanto maiores fossem os escores, mais favoráveis seriam as atitudes. Com os dados abaixo, testar a hipótese de que, com relação às "atitudes", o filme não exerceu a menor influência sobre os estudantes.

Estudantes	*Antes*	*Depois*
A	2	4
B	2	5
C	4	3
D	6	8
E	7	9
F	5	8

11. Testar a significância da diferença entre médias, numa situação do tipo "antes-depois", com os seguintes escores componentes de amostras aleatórias:

Respondente	*Antes*	*Depois*
A	7	3
B	6	4
C	5	2
D	4	3

12. Testar a significância da diferença entre médias, numa situação do tipo "antes-depois", com os seguintes escores componentes de amostras aleatórias:

Respondente	*Antes*	*Depois*
A	6	3
B	7	4
C	10	9
D	9	7
E	8	5

9

Análise de Variância

Negros e brancos, machos e fêmeas, liberais e conservadores representam as espécies de comparações de duas amostras que ocuparam nossa atenção no capítulo anterior. Entretanto, a realidade social não pode–sempre–ser convenientemente dividida em dois grupos; os respondentes nem sempre se dividem de modo tão simplificado.

Como resultado, o pesquisador quase sempre procura fazer comparações entre três, quatro, cinco ou mais amostras ou grupos. Para ilustrar, ele pode querer estudar a influência da identidade racial (negros, brancos ou orientais) sobre a discriminação profissional, do grau de privação econômica (severo, moderado ou brando) sobre a delinqüência juvenil, da classe social (alta, média, baixa ou operária) sobre a motivação subjacente às realizações.

O aluno pode perguntar a si mesmo se devemos usar uma *série* de razões t para fazer comparações entre três ou mais médias. Suponhamos, por exemplo, que desejemos testar a influência da classe social sobre a motivação subjacente a realizações. Por que não podemos comparar duas a duas todas as possíveis combinações de classe social e obter um t (t observado) para cada comparação? Pelo uso deste método, quatro amostras geram seis combinações pareadas para as quais seis valores (razões) t devem ser calculados:

1. classe alta versus classe média;
2. classe alta versus classe operária;
3. classe alta versus classe baixa;
4. classe média versus classe operária;
5. classe média versus classe baixa;
6. classe operária versus classe baixa.

O procedimento de calcular uma série de razões t envolve não somente uma boa dose de trabalho como também uma limitação estatística. Isto deve-se ao fato de ele aumentar a probabilidade de cometer-se um erro alfa–o erro de rejeitar-se a hipótese nula quando, na verdade, ela deveria ser aceita. Recordem que o pesquisador está geralmente disposto a aceitar um risco de 5% no que se refere a cometer um erro alfa (o chamado nível de confiança da ordem de 0,05). Ele espera, portanto, que, *por mero acaso*, 5 de cada 100 diferenças de médias amostrais sejam grandes o suficiente para

serem consideradas significantes. Quanto mais testes estatísticos realizarmos, entretanto, maiores as possibilidades de encontrarmos valores estatisticamente significantes em virtude de erro amostral (mais do que em virtude de uma real diferença populacional) e, daí, cometer um erro alfa. Quando realizamos um grande número de tais testes, a interpretação do resultado torna-se problemática. Um exemplo extremo: como interpretaríamos um t significante no meio de 1.000 comparações pareadas feitas num estudo particular? Sabemos que pelo menos algumas diferenças entre médias podem ocorrer simplesmente como resultado de erro amostral.

Para superar esse problema e esclarecer a interpretação de nosso resultado, necessitamos de um teste estatístico que, mantendo o erro alfa num nível constante, permita tomar uma *única* decisão–geral–quanto à presença de uma diferença significativa entre as três ou mais médias que buscamos comparar. Tal teste é conhecido por *análise de variância*.

A LÓGICA DA ANÁLISE DE VARIÂNCIA

Para fazer uma análise de variância, a *variação total* de um conjunto de escores é tratada como sendo divisível em dois componentes: a distância* dos escores brutos com relação às médias dos grupos a que pertencem–e a isto se chama *variação dentro dos grupos*–, e a própria distância existente entre as médias dos vários grupos–o que recebe o nome de *variação entre grupos*.

A fim de tornar possível o exame da variação dentro dos grupos, os escores de motivação-realização de sujeitos pertencentes a quatro classes sociais–(1) baixa, (2) operária, (3) média e (4) alta–foram representados graficamente na Figura 9.1, onde X_1, X_2, X_3 e X_4 correspondem a escores brutos em seus grupos de referência e $\bar{X}_1$, $\bar{X}_2$, $\bar{X}_3$ e $\bar{X}_4$, às respectivas médias grupais. Em termos simbólicos, vemos que tal variação "dentro" dos grupos refere-se às distâncias entre X_1 e $\bar{X}_1$, entre X_2 e $\bar{X}_2$, entre X_3 e $\bar{X}_3$ e entre X_4 e $\bar{X}_4$.

Podemos visualizar também a variação "entre" grupos. Com o auxílio da Figura 9.2, vemos que o grau de motivação subjacente a realizações varia com a classe social: o grupo pertencente à classe alta ($\bar{X}_4$) tem maior motivação do que o grupo pertencente à classe média ($\bar{X}_3$) o qual, por sua vez, tem maior motivação do que o grupo perten-

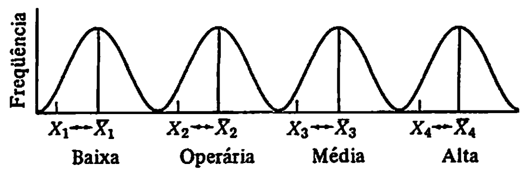


FIGURA 9.1 Representação Gráfica da Variação Dentro de Quatro Grupos Pertencentes a Classes Sociais Distintas

* N.T.: *Distância*, aqui, corresponde a *discrepância*.

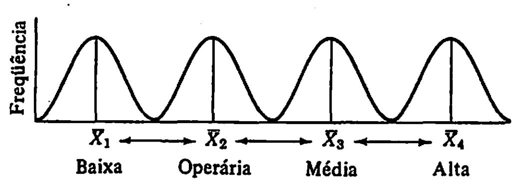


FIGURA 9.2 Representação Gráfica da Variação Entre Quatro Grupos Pertencentes a Classes Sociais Distintas

cente à classe operária ($\bar{X}_2$) e este, fechando o círculo, maior motivação do que o grupo pertencente à classe baixa ($\bar{X}_1$).

A distinção entre a variação *dentro* dos grupos e a variação *entre* grupos não é peculiar à análise de variância. Muito embora não figurasse com essa denominação, encontramos uma distinção semelhante na definição da estatística t, onde a diferença *entre* $\bar{X}_1$ e $\bar{X}_2$ foi comparada com o erro padrão da diferença (σ_{dif}), que corresponde a uma estimativa combinada de diferenças *dentro* de cada grupo. Portanto,

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{dif}} \quad \begin{array}{l} \leftarrow \text{variação entre grupos} \\ \leftarrow \text{variação dentro dos grupos} \end{array}$$

De modo similar, a análise de variância comporta uma *razão F*, cujo numerador representa a variação entre os grupos comparados, e cujo denominador contém uma estimativa da variação dentro desses grupos. Como veremos, a estatística F (razão F) indica o tamanho da diferença entre os grupos *em função* do tamanho da variação dentro de cada grupo. Da mesma forma como era verdade para a estatística t, quanto maior a estatística F (quanto maior a variação entre os grupos com relação à variação dentro dos grupos), maior a probabilidade de rejeitar-se a hipótese nula e aceitar-se a hipótese experimental.

SOMAS DE QUADRADOS

No âmago da análise de variância está o conceito de *soma de quadrados*, que representa o passo inicial para medir a variação total e a variação entre e dentro dos grupos. Pode ser até uma agradável surpresa descobrir que somente o título "soma de quadrados" é novidade para nós. O conceito em si foi introduzido no Capítulo 5 como sendo um passo importante no procedimento para a obtenção do desvio padrão. Naquele contexto, aprendemos a encontrar a soma de quadrados multiplicando, numa dada distribuição, cada discrepância por si mesma, somando, em seguida, os valores resultantes (Σx^2). Esse procedimento eliminou os sinais negativos, ao mesmo tempo que forneceu uma base matemática sólida para o desvio padrão.

Quando aplicado a uma situação em que grupos devem ser comparados, há mais de um tipo de soma de quadrados, muito embora cada um deles represente *a soma de afastamentos, ao quadrado, a partir da média*. Conforme a distinção entre variação total e seus dois componentes, temos as seguintes notações: SS_t = soma total de quadrados, SS_e = soma de quadrados entre grupos, e SS_d = soma de quadrados dentro dos grupos.

Uma Ilustração de Pesquisa

Consideremos uma situação de pesquisa em que seja possível calcular cada tipo de soma de quadrados. Suponha-se que busquemos determinar a influência da orientação política sobre os métodos de educar as crianças. No capítulo precedente, estudamos esse tópico pela comparação entre liberais e conservadores. Em contraste, queremos agora fazer comparações que representam *vários* pontos ao longo da, por assim dizer, escala política (*continuum* político). Por exemplo, poderíamos comparar a permissividade na educação infantil demonstrada por conservadores, moderados, liberais e radicais. Em tal caso, teríamos:

Hipótese Nula: conservadores, moderados, liberais e radicais não diferem com respeito à permissividade na educação infantil. ($M_1 = M_2 = M_3 = M_4$)

Hipótese Experimental: conservadores, moderados, liberais e radicais diferem com respeito à permissividade na educação infantil. ($M_1 \neq M_2 \neq M_3 \neq M_4$)

Vamos imaginar que de fato entrevistamos amostras aleatórias de quatro conservadores, quatro moderados, quatro liberais e quatro radicais com vistas a determinar seus métodos de educar crianças. Imaginemos, ainda, que os valores obtidos para permissividade sejam os constantes da Tabela 9.1 (onde a variável permissividade assume valores que vão de 1 (pouca) até 5 (muita).

Soma de Quadrados Dentro dos Grupos

A *soma de quadrados dentro dos grupos dá-nos a soma dos quadrados dos afastamentos de cada valor bruto com relação à média da amostra a que ele pertence.* Portanto, a soma de quadrados dentro dos grupos pode ser obtida simplesmente pela *combinação* da soma de quadrados dentro de cada amostra. Em símbolos:

$$SS_d = \Sigma x_1^2 + \Sigma x_2^2 + \Sigma x_3^2 + \Sigma x_4^2 \text{, onde}$$

$$x = \text{afastamento (discrepância) obtido a partir de } (X - \bar{X}).$$

Aplicando a fórmula SS_d aos dados da Tabela 9.1, obtemos:

$$SS_d = 1{,}00 + 2{,}00 + 0{,}74 + 2{,}74 = 6{,}48.$$

Soma de Quadrados Entre Grupos

A *soma de quadrados entre grupos* representa a *soma dos quadrados dos afastamentos de cada média amostral com relação à media total.* De acordo com o que precede, devemos determinar a diferença entre cada média amostral e a média total ($\bar{X} - \bar{X}_t$), elevar essa diferença ao quadrado, multiplicá-la pelo número de valores da amostra e, por fim, somar tudo. A fórmula decorrente da definição de soma de quadrados entre grupos é

$$SS_e = \Sigma(\bar{X} - \bar{X}_t)^2 N \text{, onde}$$

$\bar{X}$ = qualquer média amostral
$\bar{X}_t$ = média total (média de todos os valores brutos originários de todas as amostras combinadas)
N = número de valores em qualquer das amostras
SS_e = soma de quadrados entre grupos.

O procedimento para encontrar a soma de quadrados entre grupos, relativamente aos dados da Tabela 9.1, figura logo abaixo do quadro.*

TABELA 9.1 Valores para Permissividade Relacionada com a Educação de Crianças, Obtidos de Amostras de Conservadores, Moderados, Liberais e Radicais.

Conservadores ($N = 4$)			*Moderados* ($N = 4$)		
X_1	x	x^2	X_2	x	x^2
1	– 0,50	0,25	1	– 1	1
2	0,50	0,25	3	1	1
1	– 0,50	0,25	2	0	0
2	0,50	0,25	2	0	0
$\Sigma X = 6$		$\Sigma x^2 = 1{,}00$	$\Sigma X = 8$		$\Sigma x^2 = 2{,}00$
$\bar{X}_1 = \frac{6}{4} = 1{,}5$			$\bar{X}_2 = \frac{8}{4} = 2{,}0$		

Liberais ($N = 4$)			*Radicais* ($N = 4$)		
X_3	x	x^2	X_4	x	x^2
1	– 0,75	0,56	3	1,25	1,56
2	0,25	0,06	2	0,25	0,06
2	0,25	0,06	1	– 0,75	0,56
2	0,25	0,06	1	– 0,75	0,56
$\Sigma X = 7$		$\Sigma x^2 = 0{,}74$	$\Sigma X = 7$		$\Sigma x^2 = 2{,}74$
$\bar{X}_3 = \frac{7}{4} = 1{,}75$			$\bar{X}_4 = \frac{7}{4} = 1{,}75$		

$\bar{X}_t = 1{,}75$

$$\begin{aligned} SS_e &= (1{,}50 - 1{,}75)^2 4 + (2{,}0 - 1{,}75)^2 4 + (1{,}75 - 1{,}75)^2 4 + (1{,}75 - 1{,}75)^2 4 \\ &= (-0{,}25)^2 4 + (0{,}25)^2 4 + (0)4 + (0)4 = (0{,}06)4 + (0{,}06)4 + (0)4 + (0)4 \\ &= 0{,}24 + 0{,}24 = 0{,}48 \end{aligned}$$

* N.T.: Por comodidade, N é o mesmo para todas as amostras, isto é, o número de elementos (sujeitos, dados) não varia de uma para outra. Isso não é absolutamente necessário, porém *muito conveniente*, e é o tipo de coisa bastante controlável pelo pesquisador.

Soma Total de Quadrados

Pode-se demonstrar que a *soma total de quadrados, a soma dos quadrados dos afastamentos de cada valor bruto com relação à média total do estudo em questão*, é igual a uma combinação (soma) dos seus componentes "dentro" e "entre". A soma total de quadrados para os dados da Tabela 9.1 pode ser calculada como se segue:

$$SS_t = SS_e + SS_d = 0{,}48 + 6{,}48 = 6{,}96.$$

A soma total de quadrados pode também ser definida em termos da equação

$$SS_t = \Sigma(X - \bar{X}_t)^2 \text{, onde}$$

X = um valor bruto de qualquer amostra
$\bar{X}_t$ = média total (média de todos os valores brutos originários de todas as amostras combinadas)
SS_t = soma total de quadrados.

Usando a fórmula anterior, subtraímos a média total ($\bar{X}_t$) de cada valor bruto do estudo (X), elevamos ao quadrado os desvios (afastamentos) resultantes, somando-os em seguida.

Para os dados da Tabela 9.1, vem:

$$\begin{aligned} SS_t = {} & (1 - 1{,}75)^2 + (2 - 1{,}75)^2 + (1 - 1{,}75)^2 + (2 - 1{,}75)^2 + (1 - 1{,}75)^2 \\ & + (3 - 1{,}75)^2 + (2 - 1{,}75)^2 + (2 - 1{,}75)^2 + (1 - 1{,}75)^2 + (2 - 1{,}75)^2 \\ & + (2 - 1{,}75)^2 + (2 - 1{,}75)^2 + (3 - 1{,}75)^2 \\ & + (2 - 1{,}75)^2 + (1 - 1{,}75)^2 + (1 - 1{,}75)^2 \\ = {} & (-0{,}75)^2 + (0{,}25)^2 + (-0{,}75)^2 + (0{,}25)^2 + (-0{,}75)^2 + (1{,}25)^2 \\ & + (0{,}25)^2 + (0{,}25)^2 + (-0{,}75)^2 + (0{,}25)^2 + (0{,}25)^2 + (0{,}25)^2 \\ & + (1{,}25)^2 + (0{,}25)^2 + (-0{,}75)^2 + (-0{,}75)^2 \\ = {} & 0{,}56 + 0{,}06 + 0{,}56 + 0{,}06 + 0{,}56 + 1{,}56 + 0{,}06 + 0{,}06 + 0{,}56 + 0{,}06 \\ & + 0{,}06 + 0{,}06 + 1{,}56 + 0{,}06 + 0{,}56 + 0{,}56 \\ = {} & 6{,}96 \end{aligned}$$

Cálculo das Somas de Quadrados

As fórmulas decorrentes das definições de somas de quadrados para valores intra-grupos, valores inter-grupos e valores totais baseiam-se na manipulação de afastamento, o que torna o processo trabalhoso e demorado. Felizmente, podemos, ao invés, empregar as fórmulas abaixo, muito mais simples, para obter um resultado na forma de estatística F, a qual é idêntica (excetuados os erros de arredondamento) à que se obtém com as fórmulas decorrentes das definições, de natureza mais complexa.

Os valores brutos da Tabela 9.1 foram transcritos na Tabela 9.2 com o propósito de ilustrar o uso das fórmulas abreviadas para a soma de quadrados.

A fórmula abreviada que permite calcular a soma total de quadrados é a seguinte:

$$SS_t = \Sigma X_t^2 - \frac{(\Sigma X_t)^2}{N_t} \text{, onde}$$

N_t = número total de valores, combinadas* todas as amostras.

Introduzindo os dados da Tabela 9.2 nessa fórmula resulta:

$$SS_t = (10 + 18 + 13 + 15) - \frac{(6 + 8 + 7 + 7)^2}{4 + 4 + 4 + 4}$$

$$= 56 - \frac{(28)^2}{16} = 56 - \frac{784}{16} = 56 - 49 = 7$$

TABELA 9.2 Valores para Permissividade Relacionada com a Educação de Crianças, Obtidos de Amostras de Conservadores, Moderados, Liberais e Radicais

Conservadores ($N = 4$)		*Moderados* ($N = 4$)	
X_1	X^2	X_2	X^2
1	1	1	1
2	4	3	9
1	1	2	4
2	4	2	4
$\Sigma X = 6$	$\Sigma X^2 = 10$	$\Sigma X = 8$	$\Sigma X^2 = 18$
$\bar{X} = \frac{6}{4} = 1{,}5$		$\bar{X} = \frac{8}{4} = 2{,}0$	

Liberais ($N = 4$)		*Radicais* ($N = 4$)	
X_3	X^2	X_1	X^2
1	1	3	9
2	4	2	4
2	4	1	1
2	4	1	1
$\Sigma X = 7$	$\Sigma X^2 = 13$	$\Sigma X = 7$	$\Sigma X^2 = 15$
$\bar{X} = \frac{7}{4} = 1{,}75$		$\bar{X} = \frac{7}{4} = 1{,}75$	

$\bar{X}_t = 1{,}75$

A soma de quadrados entre grupos pode ser obtida mediante a seguinte fórmula:

$$SS_e = \left[\sum \frac{(\Sigma X)^2}{N}\right] - \frac{(\Sigma X_{\text{total}})^2}{N_{\text{total}}} \text{, onde}$$

N = número total de escores (valores) em qualquer amostra
N_t = número total de escores, combinadas todas as amostras

* N.T.: "Combinada", aí, não alude a nenhum aspecto de análise combinatória; a palavra está sendo usada para indicar a junção de várias amostras numa só, que, por isso, torna-se maior.

Por exemplo, na Tabela 9.2:

$$SS_e = \frac{(6)^2}{4} + \frac{(8)^2}{4} + \frac{(7)^2}{4} + \frac{(7)^2}{4} - \frac{(28)^2}{16} = \frac{36}{4} + \frac{64}{4} + \frac{49}{4} + \frac{49}{4} - \frac{784}{16}$$

$$= 9{,}0 + 16 + 12{,}25 + 12{,}25 - 49{,}0 = 49{,}5 - 49{,}0 = 0{,}50$$

Posto que a soma de quadrados intra-grupos constitui a parte mais demorada para ser calculada, podemos tirar vantagem do fato de que a soma total de quadrados é igual à soma de seus dois componentes. Portanto,

$$SS_d = SS_t - SS_e$$

No presente caso,

$$SS_d = 7{,}00 - 0{,}50 = 6{,}50$$

A seguinte fórmula abreviada serve para encontrar eventuais erros no cálculo da soma de quadrados de valores intra-grupos:

$$SS_d = \sum\left[(\Sigma X^2) - \frac{(\Sigma X)^2}{N}\right]$$

onde

X = escore bruto de qualquer amostra
N = o número total de escores de qualquer amostra.

Introduzindo os dados da Tabela 9.2, resulta:

$$SS_d = \left[10 - \frac{(6)^2}{4}\right] + \left[18 - \frac{(8)^2}{4}\right] + \left[13 - \frac{(7)^2}{4}\right] + \left[15 - \frac{(7)^2}{4}\right]$$

$$= \left(10 - \frac{36}{4}\right) + \left(18 - \frac{64}{4}\right) + \left(13 - \frac{49}{4}\right) + \left(15 - \frac{49}{4}\right)$$

$$= (10 - 9{,}0) + (18 - 16{,}0) + (13 - 12{,}25) + (15 - 12{,}25)$$

$$= 1{,}0 + 2{,}0 + 0{,}75 + 2{,}75 = 6{,}50$$

QUADRADO MÉDIO

Como bem poderíamos esperar de uma medida de variação, o valor das somas de quadrados tende a tornar-se maior à medida que a variação aumenta. Por exemplo, é provável que SS = 10,9 designe uma variação maior do que SS = 1,3. Entretanto, a soma de quadrados também se torna maior com o aumento do tamanho da amostra, de sorte que $N = 200$ produzirá uma SS maior do que quando $N = 20$. Como resultado, a soma de quadrados não pode ser considerada como uma medida de variação "pura" inteiramente satisfatória, a menos que, é claro, possamos encontrar uma forma de controlar o número de escores envolvidos.

Felizmente, tal método existe numa medida de variação conhecida por *quadrado médio* (ou *variância*), que se obtém através da divisão da SS_e ou da SS_d pelo número adequado de graus de liberdade (no Capítulo 5, dividimos, de forma análoga, Σx^2 por N como um passo na obtenção do desvio padrão). Portanto,

$$MS_e = \frac{SS_e}{gl_e}, \text{ onde}$$

MS_e = quadrado médio entre grupos
SS_e = soma de quadrados entre grupos
gl_e = graus de liberdade entre grupos

e

$$MS_d = \frac{SS_d}{gl_d}, \text{ onde}$$

MS_d = quadrado médio dentro dos grupos
SS_d = soma de quadrados dentro dos grupos
gl_d = graus de liberdade dentro dos grupos.

Contudo, precisamos ainda obter os graus de liberdade apropriados. Para o quadrado médio entre grupos,

$$gl_e = R - 1, \text{ onde}$$

R = número de amostras.

Para encontrar o número de graus de liberdade no quadrado médio dentro dos grupos,

$$gl_d = N_{total} - R, \text{ onde}$$

N_{total} = número total de escores, somadas todas as amostras
R = número de amostras.

Ilustrando com os dados da Tabela 9.2, na qual $SS_e = 0{,}50$ e $SS_d = 6{,}50$, calculamos nossos graus de liberdade como se segue:

$$gl_e = 4 - 1 = 3$$

e

$$gl_d = 16 - 4 = 12.$$

Estamos agora preparados para obter os quadrados médios:

$$MS_e = \frac{0{,}50}{3} = 0{,}17$$

e

$$MS_d = \frac{6{,}50}{12} = 0{,}54.$$

A ESTATÍSTICA F (RAZÃO F)

Como observamos anteriormente, a análise de variância produz uma estatística F na qual se comparam a variação entre grupos e a variação dentro dos grupos. Estamos agora prontos para especificar o grau de cada tipo de variação—quando medida por meio de quadrados médios. Portanto, a estatística F pode ser considerada como indicativa do tamanho do quadrado médio entre grupos com relação ao tamanho do quadrado médio dentro dos grupos, ou seja:

$$F = \frac{MS_e}{MS_d}$$

Para a Tabela 9.2,

$$F = \frac{0{,}17}{0{,}54} = 0{,}31$$

Tendo obtido uma razão (estatística) F, devemos agora decidir se ela é grande o suficiente para rejeitar a hipótese nula e levar à aceitação da hipótese experimental. Será que conservadores, moderados, liberais e radicais realmente diferem quanto à permissividade na educação de crianças? Quanto maior a razão F calculada (isto é, quanto maior a MS_e e menor a MS_d), com maior probabilidade vamos obter um resultado significante em termos estatísticos.

Mas como reconhecer de maneira exata uma razão F significante? Recorde-se que, no Capítulo 8, a razão t que obtivemos foi comparada com diversos valores de uma tabela de razões t, ao nível de significância de 0,05, e com os graus de liberdade apropriados. De forma análoga, devemos agora interpretar nossa estatística F (calculada) com o auxílio da Tabela D no final do livro. Ela contém uma lista de razões F significantes, isto é, razões F que precisamos obter* para que seja possível rejeitar a hipótese nula aos níveis de significância de 0,05 e 0,01. Tal como ocorreu com a estatística t, o valor *específico* de F que devemos considerar (relativamente à Tabela D) depende do número de graus de liberdade. Portanto, ao consultar a Tabela D, procuramos duas vezes os graus de liberdade: *uma* para os grupos "entre" e *outra* para os grupos "dentro". Os graus de liberdade associados ao numerador (gl_e) foram escritos horizontalmente no topo da tabela (D); os graus de liberdade associados ao denominador (gl_d) foram colocados na vertical do lado esquerdo da mesma tabela. O corpo da Tabela D contém os F significantes aos níveis de 0,05 e 0,01.

Para os dados da Tabela 9.2, encontramos $gl_e = 3$ e $gl_d = 12$. Então, entramos na Tabela D localizando, em primeiro lugar, a coluna $gl_e = 3$ e a linha $gl_d = 12$. Com esse procedimento encontramos, no cruzamento, um F significante que, ao nível de 0,05, deve ser igual a 3,49, no mínimo; na tabela da página seguinte (*ainda* Tabela D), verificamos que, ao nível de 0,01, o F crítico deve ser igual a ou maior que 5,95. O nosso F calculado (F_o = F observado) é de apenas 0,31. Como resultado, não temos outra escolha senão "aceitar" a hipótese nula e atribuir as diferenças das médias amostrais (no tocante à permissividade ligada à criação de filhos) a *erro amostral*. Nada nos leva a crer que essas diferenças devam-se a uma *real* diferença nas populações constituídas por conservadores, moderados, liberais e radicais.

Os resultados de nossa análise de variância podem ser apresentados num quadro-resumo semelhante ao da Tabela 9.3. Tornou-se prática comum resumir uma análise de variância desse jeito.

TABELA 9.3 Quadro-resumo de Análise de Variância (Resumo Ligado aos Dados da Tabela 9.2)

Fonte de Variação	*gl*	*SS*	*MS*	*F*
Entre grupos	3	0,50	0,17	0,31
Dentro dos grupos	12	6,50	0,54	

Uma Ilustração

Com vistas a oferecer uma ilustração passo a passo de uma análise de variância, vamos supor que desejemos testar a hipótese de que o QI varia em função da classe social. Portanto:

Hipótese Nula: As classes alta, média e baixa não diferem significativamente com respeito ao QI ($M_1 = M_2 = M_3$).

* N.T.: a) A *razão F* ou a *estatística F* que se obtém a partir de um conjunto de dados, e que o Autor chama de *calculada*, costuma ser também denominada "F_o" (*F observado*). b) As razões F que figuram na Tabela D são também chamadas "F_c" (*F críticos*). Observe-se que a rejeição de uma hipótese nula em favor de uma hipótese experimental (ou alternativa) ocorre sempre que $F_o \geqslant F_c$.

Hipótese Experimental (ou Alternativa): As classes alta, média e baixa diferem significativamente com respeito ao QI ($M_1 \neq M_2 \neq M_3$).

Para testar essa hipótese, vamos convencionar o nível de significância de 5% como critério da confiança depositada nos resultados da pesquisa. Admitamos que seja possível medir o QI dos membros das três amostras de classes sociais colhidas: alta, média e baixa. Os resultados das mensurações figuram abaixo:

Classe Alta ($N=5$)		*Classe Média* ($N=5$)	
X_1	X^2	X_2	X^2
130	16900	120	14400
125	15625	115	13225
130	16900	115	13225
120	14400	110	12100
122	14884	112	12544
$\Sigma X = 627$	$\Sigma X^2 = 78709$	$\Sigma X = 572$	$\Sigma X^2 = 65494$
$\bar{X}_1 = 125{,}4$		$\bar{X}_2 = 114{,}4$	

Classe Baixa ($N=5$)	
X_3	X^2
110	12100
100	10000
90	8100
100	10000
85	7225
$\Sigma X = 485$	$\Sigma X^2 = 47425$
$\bar{X}_3 = 97{,}0$	

O procedimento passo a passo para testar a significância estatística das diferenças entre as médias obtidas é o seguinte:

PASSO 1: Calcular a média de cada amostra.

$$\bar{X}_1 = \frac{\Sigma X_1}{N} = \frac{627}{5} = 125{,}4 \qquad \bar{X}_2 = \frac{\Sigma X_2}{N} = \frac{572}{5} = 114{,}4 \qquad \bar{X}_3 = \frac{\Sigma X_3}{N} = \frac{485}{5} = 97{,}0$$

Note-se que diferenças entre as médias *existem de fato*, e que há uma tendência de aumento nos QIs à medida que a classe varia de baixa para alta.

PASSO 2: Calcular a soma total de quadrados

$$SS_{total} = \Sigma X^2_{total} - \frac{(\Sigma X_{total})^2}{N_{total}}$$

$$= (78709 + 65494 + 47425) - \frac{(627 + 572 + 485)^2}{15}$$

$$= 191628 - \frac{(1684)^2}{15} = 191628 - \frac{2835856}{15}$$

$$= 191628 - 189057{,}07 = 2570{,}93$$

PASSO 3: Calcular a soma de quadrados entre grupos.

$$SS_e = \left[\sum \frac{(\Sigma X)^2}{N}\right] - \frac{(\Sigma X_t)^2}{N_t}$$

$$= \frac{(627)^2}{5} + \frac{(572)^2}{5} + \frac{(485)^2}{5} - \frac{(1684)^2}{15}$$

$$= \frac{393129}{5} + \frac{327184}{5} + \frac{235225}{5} - \frac{2835856}{15}$$

$$= 78625{,}8 + 65436{,}8 + 47045{,}0 - 189057{,}07 = 191107{,}60 - 189057{,}07$$

$$= 2050{,}53$$

PASSO 4: Calcular a soma de quadrados dentro dos grupos.

$$SS_d = SS_t - SS_e = 2570{,}93 - 2050{,}53 = 520{,}40$$

ou

$$SS_d = \sum\left[(\Sigma X^2) - \frac{(\Sigma X)^2}{N}\right]$$

$$= \left[78709 - \frac{(627)^2}{5}\right] + \left[65494 - \frac{(572)^2}{5}\right] + \left[47425 - \frac{(485)^2}{5}\right]$$

$$= \left[78709 - \frac{393129}{5}\right] + \left[65494 - \frac{327184}{5}\right] + \left[47425 - \frac{235225}{5}\right]$$

$$= [78709 - 78625{,}8] + [65494 - 65436{,}8] + [47425 - 47045{,}0]$$

$$= 83{,}2 + 57{,}2 + 380{,}0 = 520{,}40$$

PASSO 5: Calcular os graus de liberdade entre grupos.

$$gl_e = k - 1 = 3 - 1 = 2$$

PASSO 6: Calcular os graus de liberdade dentro dos grupos.

$$gl_d = N_t - k = 15 - 3 = 12$$

PASSO 7: Calcular o quadrado médio entre grupos.

$$MS_e = \frac{SS_e}{gl_e} = \frac{2050{,}53}{2} = 1025{,}27$$

PASSO 8: Calcular o quadrado médio dentro dos grupos.

$$MS_d = \frac{SS_d}{gl_d} = \frac{520{,}40}{12} = 43{,}37$$

PASSO 9: Calcular a razão F (F_o, isto é, o F observado).

$$F = \frac{MS_e}{MS_d} = \frac{1025{,}27}{43{,}37} = 23{,}64$$

PASSO 10: Comparar o F observado com o F crítico constante da Tabela D.

F observado = 23,64
F crítico = 3,88

$$gl = \frac{2}{12}$$

$$P = 0{,}05$$

Como bem demonstra o passo 10, para rejeitar-se a hipótese nula ao nível de significância de 0,05, com $\frac{2}{12}$ graus de liberdade*, nosso F calculado (observado) deve ser de, no mínimo, igual a 3,88. Uma vez que obtivemos um F igual a 23,64, podemos rejeitar a hipótese nula e aceitar a hipótese experimental. Especificamente, é possível concluir que, com relação à variável QI, existe uma real diferença entre as classes baixa, média e alta.

* N.T.: Não existe grau de liberdade menor que 1. Por isso, não se vá confundir $gl = \frac{2}{12}$ com uma fração. $\frac{2}{12}$ significa: $\frac{2 \text{ gl no numerador}}{12 \text{ gl no denominador}}$.

COMPARAÇÃO MÚLTIPLA DE MÉDIAS

Um F significante dá-nos uma informação a respeito da *diferença global* existente entre os grupos (amostras) estudados. Se estivéssemos pesquisando a diferença entre apenas duas médias amostrais, nenhuma análise adicional seria necessária para a interpretação de nosso resultado: em tal situação, a diferença obtida ou *é* estatisticamente significante ou *não*, dependendo da magnitude de F. Entretanto, quando encontramos um F significante relacionado com diferenças entre três ou mais médias, pode ocorrer que seja importante determinar onde se situam essas diferenças. Por exemplo, na ilustração anterior, descobrimos que existem, com relação às três classes sociais pesquisadas, diferenças de QI significantes do ponto de vista estatístico. Considerem as possibilidades levantadas por esse F significante: $\bar{X}_1$ (alta) poderia diferir significantemente de $\bar{X}_2$ (média); $\bar{X}_1$ (alta) poderia diferir significantemente de $\bar{X}_3$ (baixa); $\bar{X}_2$ poderia diferir significativamente de $\bar{X}_3$ (baixa).

Como já foi explicado páginas atrás, tentar obter uma razão t para cada comparação ($\bar{X}_1$ versus $\bar{X}_2$; $\bar{X}_1$ versus $\bar{X}_3$; $\bar{X}_2$ versus $\bar{X}_3$) implicaria uma quantidade proibitiva de trabalho, além de aumentar a probabilidade de cometer-se um erro alfa. Felizmente, estatísticos desenvolveram vários outros testes que permitem fazer comparações múltiplas a partir de um F significante (isto é, $F_o \geqslant F_c$) e localizar *onde* se situam as diferenças significantes entre médias. Vamos introduzir, então, o teste DHS de Tukey, um dos mais úteis no terreno da comparação múltipla. (Obs.: DHS significa "*d*iferença *h*onestamente *s*ignificante".)

O DHS de Tukey é usado *apenas* quando um F significante já foi obtido**. Por este método, o que fazemos é comparar a diferença entre quaisquer duas médias com a DHS. *Uma diferença entre duas médias diz-se estatisticamente significante só se for igual a ou maior que a DHS.* Em símbolos:

$$\text{DHS} = q\alpha\sqrt{\frac{MS_d}{n}}$$

onde

$q\alpha$ = valor tabelado (isto é, valor *crítico*) para um dado nível de significância*, a partir do número máximo de médias que estejam sendo comparadas.

MS_d = quadrado médio intra-grupos (obtido na análise da variância)

n = número de respondentes em cada grupo (admite-se que em todos os grupos, ou amostras, haja o *mesmo* número de sujeitos).

Em oposição ao que ocorre com a razão t, a DHS leva em conta o fato de que a probabilidade de um erro alfa aumenta à medida que aumenta o número de médias

* N.T.: Recorde-se, mais uma vez, que o Autor "prefere" usar a expressão "nível de confiança" em lugar de "nível de significância". Na verdade, não há grande inconveniência, desde que essa "confiança" seja interpretada da seguinte forma: (100% – 5%) ou (100% – 1%).

** N.T.: Isto é, quando $F_o \geqslant F_c$.

comparadas. Dependendo do valor de $q\alpha$, quanto maior o número de médias, mais "conservadora" (= resistente) torna-se a DHS no tocante à rejeição da hipótese nula. Como resultado, obteremos menos diferenças significantes com a DHS do que com a razão t. Além disso, uma diferença entre médias tem maior probabilidade de ser *significante* numa comparação múltipla de três médias do que numa comparação múltipla de quatro ou cinco.

Para ilustrar o uso da DHS, vamos voltar ao exemplo anterior em que se verificou existirem diferenças nas classes sociais com respeito à variável QI. Mais especificamente, obtivemos um F significante (F = 23,64) para as seguintes médias colhidas de amostras das três classes sociais:

$\bar{X}_1$ (classe alta) = 125,4
$\bar{X}_2$ (classe média) = 114,4
$\bar{X}_3$ (classe baixa) = 97,0

PASSO 1: Construir uma tabela de diferenças entre médias ordenadas. Para os dados presentes, a ordem das médias (da menor para a maior) é 97,0, 114,4 e 125,4. Elas são dispostas em tabela de tal forma que a diferença entre cada par de médias apareça numa matriz. Assim a diferença entre $\bar{X}_1$ (alta) e $\bar{X}_3$ (baixa) é 28,40; a diferença entre $\bar{X}_1$ (alta) e $\bar{X}_2$ (média) é 11,0; e a diferença entre $\bar{X}_2$ (média) e $\bar{X}_3$ (baixa) é 17,4*.

	$\bar{X}_3$ = 97,0	$\bar{X}_2$ = 114,4	$\bar{X}_1$ = 125,4
$\bar{X}_3$	–	17,4	28,4
$\bar{X}_2$	–	–	11,0
$\bar{X}_1$	–	–	–

PASSO 2: Procurar $q\alpha$ na Tabela I. Para encontrar $q\alpha$ na Tabela I (fim do livro), precisamos ter: (a) os graus de liberdade (gl) relativos a MS_d; (b) o número máximo de médias (k) e (c) o nível de significância–0,01 ou 0,05. Já sabemos, graças à análise de variância realizada, que gl = 12. Portanto, entramos na Tabela I pela coluna do lado esquerdo, até localizar 12 graus de liberdade. Em seguida, como estamos comparando três pares de médias, consultamos a tabela horizontalmente, da esquerda para a direita, até localizar a coluna que contém o número máximo de médias (k = 3). Convencionando um nível de significância de 0,05, observamos que $q0{,}05 = 3{,}77$.

PASSO 3: Calcular a DHS.

$$\text{DHS} = q0{,}05\sqrt{\frac{MS_d}{n}} = 3{,}77\sqrt{\frac{43{,}37}{5}} = 3{,}77\sqrt{8{,}67} = 3{,}77(2{,}94)$$
$$= 11{,}08$$

PASSO 4: Comparar a DHS com a matriz de diferenças de médias. Para que qualquer diferença entre médias (ver matriz) seja considerada estatisticamente significante é

* N.T.: As diferenças são consideradas em valores absolutos, isto é, desprovidas de sinal.

preciso que ela se iguale a ou exceda a DHS. De volta à nossa matriz de diferenças de médias, verificamos que a diferença de 28,4 QI entre $\bar{X}_1$ (classe alta) e $\bar{X}_3$ (classe baixa), e a diferença de 17,4 QI entre $\bar{X}_2$ (classe média) e $\bar{X}_3$ (classe baixa) são maiores que DHS = 11,08. Como resultado, concluímos que essas diferenças entre médias são estatisticamente significantes ao nível de 0,05. Somente a diferença de 11,0 entre $\bar{X}_1$ e $\bar{X}_2$ não se iguala a nem excede DHS—e é, por isso, *não-significante* do ponto de vista estatístico.

EXIGÊNCIAS PARA O USO DA ESTATÍSTICA F (RAZÃO F)

A análise de variância só deve ser feita depois de o pesquisador ter levado em conta as seguintes exigências:

1. *Comparação entre três ou mais médias independentes*—A razão F é geralmente empregada para fazer-se uma comparação entre três ou mais médias extraídas de amostras independentes. A estatística F não se presta para testes em que o número de amostras é menor que dois. Porém, no caso específico de *duas* amostras, tanto faz usar F ou t, uma vez que $F = t^2$.
2. *Dados intervalares*—Ao fazer uma análise de variância, pressupomos ter atingido o nível intervalar de mensuração. Por igual raciocínio, dados categorizados ou ordenados não devem ser usados.
3. *Amostragem casual*—Nossas amostras deverão ter sido extraídas aleatoriamente, de uma dada população de escores.
4. *Distribuição normal*—Admitimos que a variável em foco possui, na população da qual se extraem as amostras, distribuição normal*.

RESUMO

A análise de variância pode ser usada para fazer comparações entre três ou mais médias amostrais. Esse teste produz uma estatística ou razão F cujo numerador representa a variação entre grupos, e cujo denominador contém uma estimativa da variação dentro dos grupos. A soma de quadrados representa o passo inicial para mensurar a variação. Entretanto, ele sofre grande influência do tamanho da amostra. Para superar esse problema, dividimos SS_e ou SS_d pelos graus de liberdade apropriados, a fim de obter o quadrado médio. A razão F indica o tamanho do quadrado médio entre os grupos em função do tamanho do quadrado médio dentro dos grupos. Interpretamos nosso F observado (calculado) mediante uma comparação com um F crítico adequado constante da Tabela D. Com base nessa comparação, decidimos se a hipótese nula deve ou não ser rejeitada. Após a obtenção de um F significante, podemos determinar onde se localizam as diferenças significantes aplicando o método de Tukey para a comparação múltipla de médias.

PROBLEMAS

1. Nas seguintes amostras aleatórias de classe social, testar a hipótese nula de que a "camaradagem entre vizinhos" não varia com a "classe social". (*Nota:* escores mais altos indicam maior grau de camaradagem.)

* N.T.: Não se pode admitir isso arbitrariamente. Havendo dúvida quanto à normalidade, faz-se um teste para garantir-se esse pré-requisito.

Classe			
Baixa	*Operária*	*Média*	*Alta*
8	7	6	5
4	3	5	2
7	2	5	1
8	8	4	3

2. Testar a significância das diferenças entre as médias das seguintes amostras aleatórias:

Amostra 1	*Amostra 2*	*Amostra 3*
2	5	8
1	4	9
3	3	7
3	4	8

3. Testar se as diferenças entre as médias das seguintes amostras aleatórias são significantes:

Amostra 1	*Amostra 2*	*Amostra 3*
12	6	3
6	5	2
8	7	5
7	5	3
6	1	1

4. Testar se as diferenças entre as médias das seguintes amostras aleatórias são significantes:

Amostra 1	*Amostra 2*	*Amostra 3*
5	4	3
5	3	5
4	2	1
3	2	3
6	1	3

5. Utilizando os mesmos dados do problema 4, determinar através do método de Tukey (comparação múltipla de médias), onde, exatamente, ocorrem as diferenças significantes.

6. Testar se as diferenças entre as médias das seguintes amostras aleatórias são significantes:

Amostra 1	*Amostra 2*	*Amostra 3*	*Amostra 4*
1	3	4	6
1	2	4	6
3	2	2	5
4	1	2	5
2	5	3	4
1	5	3	6

7. Utilizando os mesmos dados do problema 6, determinar, através do método de Tukey (comparação múltipla de médias), onde, exatamente, ocorrem as diferenças significantes.

10

Testes Não-paramétricos

Tal como indicamos nos Capítulos 8 e 9, precisamos exigir muito do pesquisador que emprega uma estatística t ou uma análise de variância para fazer comparações entre suas amostras. Cada um desses testes de significância apresenta uma lista de requisitos, que inclui, entre outros, o pressuposto de que a característica estudada (isto é, a variável observacional) tem distribuição normal numa particular população. Além disso, cada um desses testes supõe que a variável tenha sido mensurada a nível intervalar, no mínimo, garantindo-se, assim, a possibilidade de atribuir a cada membro da amostra o respectivo escore. Quando um teste de significância—tal como a estatística (razão) t ou a análise de variância—requer (1) normalidade e (2) nível intervalar de mensuração, fala-se em *teste paramétrico.*[1]

Como se arranja o pesquisador que não consegue empregar um teste paramétrico porque não pode, honestamente, admitir normalidade de distribuição ou porque seus dados foram colhidos num nível de mensuração inferior ao intervalar? Suponha-se, por exemplo, que ele esteja trabalhando com uma distribuição assimétrica, tal como renda anual, ou com dados que tenham sido categorizados ou contados (nível nominal) ou, ainda, com dados ordenados (nível ordinal).* De que modo esse pesquisador pode fazer comparações entre amostras sem violar os requisitos de um determinado teste?

Afortunadamente, os estatísticos desenvolveram uma boa quantidade de testes

---

[1] Tal designação baseia-se no termo "parâmetro", aplicável a qualquer característica de uma população.

*N.T.: Esse problema de níveis de mensuração deve ser entendido como uma "escada": qualquer nível, a partir do segundo, pressupõe que as características dos anteriores estejam sempre presentes. Então, por exemplo, suponha-se que X assuma os valores 160 cm, 170 cm e 185 cm, onde X representa a estatura de três sujeitos. Pelo que já foi visto anteriormente, "comprimento" pertence ao nível racional (escala de razão); entretanto, nada impede que se atribua o "posto" 1 ao sujeito de 160 cm, o 2 ao de 170 cm e o 3 ao de 185 cm. Ou, nessa mesma ordem, podem aplicar-se as denominações "baixo", "mediano" e "alto", onde (baixo < mediano < alto). Essa operação, como é fácil de observar, constitui um *abaixamento* de nível de mensuração (de nível racional para nível ordinal). O que é *importante* notar é que o abaixamento de nível é *quase sempre possível* enquanto que a *elevação* (isto é, melhora) *raramente se consegue*—com sucesso.

não-paramétricos de significância–testes cuja lista de requisitos não inclui normalidade de distribuição ou nível intervalar de mensuração. Para entendermos a posição importante dos testes não-paramétricos em pesquisa, precisamos entender também o conceito estatístico de poder. *Poder de um teste* é a probabilidade de rejeitarmos a hipótese nula quando ela é *realmente* falsa–e, por isso mesmo, deve ser rejeitada.

O poder varia de um teste para outro. Os testes mais poderosos–aqueles que, com maior probabilidade, levam à rejeição da hipótese nula (H_0) quando ela é mesmo falsa–são testes que possuem os pré-requisitos mais difíceis de satisfazer. Geralmente, esses testes são paramétricos, tais como o t ou o F, que pressupõem dados colhidos em nível intervalar (no mínimo) e distribuição normal, na população, das características estudadas (isto é, as variáveis) na amostra. Em contraposição, as alternativas não-paramétricas exigem muito menos em termos de pré-requisitos, e constituem-se testes de significância com poder bem menor que o dos correspondentes paramétricos. Em conseqüência, ao admitir a hipótese nula falsa (e ao manter inalterados fatores tais como o tamanho da amostra), é mais provável que o pesquisador a rejeite pelo uso adequado de F ou de t do que mediante uma alternativa não-paramétrica.

É óbvio que os pesquisadores querem, a todo custo, rejeitar a hipótese nula–quando ela é *falsa mesmo*. Por isso, a maioria deles preferiria, idealmente, empregar testes *paramétricos* de significância. Entretanto, como observamos há pouco, é comum não ser possível preencher as exigências mínimas dos testes paramétricos. Em primeiro lugar, muitos dos dados no campo psicossocial pertencem ou ao nível ordinal ou ao intervalar. Depois, nunca podemos ter certeza de que as características estudadas na amostra têm, de fato, distribuição normal na população da qual ela foi extraída.

Quando os requisitos de um teste (paramétrico) são violados, torna-se impossível conhecer o seu poder. Daí que os resultados de um teste paramétrico, cujos requisitos não foram satisfeitos, não têm interpretação significativa. Sob tais condições, a maioria dos pesquisadores opta inteligentemente por testes de significância não-paramétricos.

Este capítulo introduz alguns dos mais conhecidos testes de significância não-paramétricos: qui-quadrado, prova exata de Fisher, teste da mediana, prova de Mann-Whitney, prova de Kruskal-Wallis (análise de variância por postos), prova de Friedman (dupla análise de variância) e prova binomial.

QUI-QUADRADO: UM TESTE DE SIGNIFICÂNCIA

O mais "popular" teste não-paramétrico de significância utilizado em pesquisa chama-se *qui-quadrado* (χ^2). Como veremos adiante, o teste de χ^2 é usado na comparação* entre duas ou mais amostras.

À semelhança do que ocorreu com a estatística t e a análise de variância, existe uma distribuição amostral de qui-quadrado que pode ser usada para estimar a probabilidade da obtenção de um valor significante (de qui-quadrado) por mero *acaso* e não

* N.T.: O termo "comparação" precisa ser *interpretado*, já que ele pressupõe "maior que," menor que" ou "igual a". O que o teste de qui-quadrado geralmente permite fazer é um *estudo relacional* entre variáveis, ou seja, a determinação do *tipo de relação* existente entre elas: *independência* ou *dependência*.

porque na população matriz existem reais diferenças* entre as variáveis estudadas. Porém, ao contrário do que ocorre com os testes de significância anteriores, o qui-quadrado é empregado para fazer comparações entre *freqüências* e *não* entre escores médios. Como resultado, a hipótese nula para o teste de qui-quadrado estabelece que as populações *não diferem* relativamente à freqüência com que ocorre uma característica particular; por outro lado, a hipótese experimental estabelece que as diferenças amostrais *refletem diferenças reais* na população matriz–a partir da freqüência relativa de uma dada característica.

Para ilustrar o uso do qui-quadrado com dados freqüenciais (ou com proporções que *possam* ser reduzidas a freqüências), imagine-se que, ainda uma vez, tenhamos sido chamados a investigar a relação entre "orientação política" e "permissividade na educação de crianças". Em vez de *dar notas* (isto é, atribuir escores) a liberais e conservadores em termos de seu grau de permissividade, podemos *categorizar* nossos sujeitos amostrais numa base estritamente do tipo "ou ... ou";** em outras palavras, o problema é decidir *quais* são permissivos e *quais* são não-permissivos. Portanto:

Hipótese Nula: *A freqüência relativa de liberais permissivos é igual à freqüência relativa de conservadores permissivos.*

Hipótese Experimental: *A freqüência relativa de liberais permissivos é diferente da freqüência relativa de conservadores permissivos.*

CÁLCULO DO QUI-QUADRADO

O teste de significância denominado qui-quadrado ocupa-se essencialmente com a distinção entre freqüências esperadas e freqüências obtidas (observadas). As *freqüências esperadas* (f_e) referem-se aos termos da hipótese nula, de acordo com os quais se espera que a freqüência relativa (ou a proporção) seja a mesma para todos os grupos. Por exemplo, se a expectativa de que os liberais sejam permissivos é de 50%, então ela também corresponde a 50% para os conservadores. Em contraposição, as *freqüências observadas* (f_o) referem-se aos resultados obtidos de forma efetiva no momento da coleta de dados, donde decorre que, de um grupo para outro, elas podem–ou não–variar. *No caso de as diferenças entre as freqüências obtidas e as esperadas serem suficientemente grandes é que rejeitamos a hipótese nula e decidimos pela afirmação de que existe uma diferença real na população.*

Continuando com o presente exemplo, suponha-se que extraíssemos uma amostra aleatória de 20 liberais e *outra* amostra, também aleatória, de 20 conservadores; em seguida, esses 40 sujeitos poderiam ser dicotomizados em "permissivos" e "não-permissivos" relativamente à variável observacional "método de educar crianças". A Tabela 10.1 apresenta as freqüências obtidas que poderiam resultar.

* N.T.: Veja-se, a propósito, o que ficou dito na nota anterior.

** N.T.: Categorizar com base no critério "ou ... ou" é, ao cabo, *dicotomizar* (se forem duas categorias), *tricotomizar* (se forem três), *tetracotomizar* (se forem quatro) e assim por diante.

Os dados da Tabela 10.1 indicam que "métodos permissivos" foram usados por 5 dos 20 liberais e por 10 dos conservadores. Esses dados podem ser retabelados num quadro do tipo 2 X 2 (2 linhas e 2 colunas), no qual as freqüências observadas estão distribuídas pelas *caselas*, ao lado das freqüências esperadas (teóricas), que figuram entre parênteses (Tabela 10.2). Observe-se que essas freqüências esperadas baseiam-se *apenas* na ação do acaso, donde decorre uma pressuposição (provisória) de que a hipótese nula esteja correta. Note-se, também, que os totais marginais na Tabela 10.2 foram obtidos pela soma horizontal ou vertical das freqüências das caselas, ou seja: 15 e 25 representam as somas das *linhas* (somas horizontais); 20 e 20, as somas das *colunas* (somas verticais). O número total de sujeitos ($N = 40$), sito é, o tamanho da amostra, pode ser obtido somando-se os totais marginais horizontais ou os totais marginais verticais.*

TABELA 10.1 Freqüências Observadas num Estudo de Permissividade Relacionada com Orientação Política

	Orientação Política	
Métodos de Educação de Crianças	*Liberais* f_o	*Conservadores* f_o
Permissivo	5	10
Não-permissivo	15	10
Total	20	20

Tendo conseguido as freqüências observadas e as esperadas para o problema em foco, o valor do qui-quadrado pode ser calculado mediante a fórmula.**

$$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e}$$

onde

f_o = freqüência observada por casela
f_e = freqüência esperada por casela
χ^2 = qui-quadrado

De acordo com a fórmula do χ^2, devemos (1) subtrair cada freqüência esperada da freqüência observada correspondente, (2) quadrar a diferença, (3) dividir o quadrado de cada diferença pela freqüência esperada adequada*** e (4) somar os quocientes parciais para obter o valor do qui-quadrado (χ^2_o).

Os dados da Tabela 10.2 podem ser usados para ilustrar o procedimento acima:

*N.T.: Obviamente, a soma dos totais marginais horizontais deve ser *igual* à soma dos totais marginais verticais. Essa, aliás, é uma prova elementar de coerência.

** N.T.: O qui-quadrado assim calculado é comumente chamado *qui-quadrado observado* e designa-se por χ^2_o.

*** N.T.: A freqüência esperada adequada é a que figura, em cada casela, entre parênteses.

TABELA 10.2 Dados da Tabela 10.1 Dispostos numa Tabela 2 X 2*

Freqüência Observada — *Freqüência Esperada* — *Total Marginal (Horizontal)*

	Liberais	*Conservadores*	
Permissivo	5 (7,5)	10 (7,5)	15
Não-permissivo	15 (12,5)	10 (12,5)	25
	20	20	$N = 40$

$$\chi^2 = \frac{(5-7{,}5)^2}{7{,}5} + \frac{(10-7{,}5)^2}{7{,}5} + \frac{(15-12{,}5)^2}{12{,}5} + \frac{(10-12{,}5)^2}{12{,}5}$$

$$= \frac{(-2{,}5)^2}{7{,}5} + \frac{(2{,}5)^2}{7{,}5} + \frac{(2{,}5)^2}{12{,}5} + \frac{(-2{,}5)^2}{12{,}5}$$

$$= \frac{6{,}25}{7{,}5} + \frac{6{,}25}{7{,}5} + \frac{6{,}25}{12{,}5} + \frac{6{,}25}{12{,}5}$$

$$= 0{,}83 + 0{,}83 + 0{,}50 + 0{,}50 = 2{,}66$$

Vemos, assim, que $\chi^2 = 2{,}66$. A fim de podermos interpretar esse valor de qui-quadrado, devemos, ainda, determinar o número de graus de liberdade adequado. Esta determinação pode ser feita mediante o emprego da seguinte fórmula independentemente do número de linhas ou de colunas existentes numa tabela:

$$\text{gl} = (l-1)(c-1), \text{ onde}$$

l = número de linhas na tabela de freqüências observadas
c = número de colunas na tabela de freqüências observadas
gl = número de graus de liberdade

Posto que as freqüências observadas na Tabela 10.2 dão origem a duas linhas e duas colunas (2 X 2), vem que:

* N.T.: Leia-se "tabela dois por dois". Além disso, essa tabela é conhecida pelos nomes *tabela de dupla entrada* e *tabela de contingência*.

$$gl = (2-1)(2-1) = (1)(1) = 1$$

Consultando a Tabela E (no fim do livro), encontramos uma lista de valores de qui-quadrado,* os quais são significantes aos níveis (de confiança/significância) de 0,05 e 0,01. Ao nível de 0,05 e 1 grau de liberdade, vemos que o qui-quadrado crítico é 3,84. Para que a hipótese nula possa ser rejeitada, é preciso que esse valor (isto é, 3,84) seja excedido ou, pelo menos, igualado pelo qui-quadrado observado. Como o nosso χ^2 observado (calculado) é somente 2,66–e, portanto, *menor* que o valor tabelado (ou seja, menor que o χ^2_c)–devemos aceitar a hipótese nula e rejeitar a hipótese experimental. As freqüências obtidas (observadas) não se afastam das esperadas o suficiente para podermos explicar a diferença como resultado de *real* diferença populacional.** Em outros termos, tal diferença (3,84 – 2,66) deve ser interpretada como resultante da ação do acaso.

CÁLCULO DAS FREQÜÊNCIAS ESPERADAS (TEÓRICAS)

As freqüências esperadas para cada casela devem refletir a ação do acaso nos termos da hipótese nula (H_0). Se tais freqüências têm que indicar "mesmice" em todas as amostras, devem ser proporcionais aos seus totais marginais, quer para as linhas, quer para as colunas.

Para obter a freqüência esperada relativa a uma determinada casela, simplesmente multiplicamos seus respectivos totais marginais (de linha e de coluna), dividindo, em seguida, o produto por N. Portanto:

$$f_e = \frac{(\text{total marginal "linha"})(\text{total marginal "coluna"})}{N}$$

Para a casela superior esquerda da Tabela 10.2 (liberais permissivos):

$$f_e = \frac{(20)(15)}{40} = \frac{300}{40} = 7,5$$

Da mesma forma, para a casela superior direita da Tabela 10.2 (conservadores permissivos):

$$f_e = \frac{(20)(15)}{40} = \frac{300}{40} = 7,5$$

* N.T.: Os valores tabelados de qui-quadrado são geralmente chamados *qui-quadrados críticos* e notados por χ^2_c.

**N.T.: Insista-se, mais uma vez, no que ficou dito na nota à pág. 176. Dessa comparação entre o χ^2_o e o χ^2_c resulta uma informação a respeito do *tipo de relação* existente entre as variáveis estudadas. Então, se $\chi^2_o < \chi^2_c$, as variáveis são *independentes*; se $\chi^2_o \geqslant \chi^2_c$, as variáveis são *dependentes*. No exemplo em foco, $\chi^2_o < \chi^2_c$ e, por isso, a hipótese experimental foi rejeitada. Isso equivale a dizer que, com certeza de 95% as variáveis "filiação política" e "permissividade na criação dos filhos" são variáveis *independentes*.

Para a casela inferior esquerda da Tabela 10.2 (liberais não-permissivos):

$$f_e = \frac{(20)(25)}{40} = \frac{500}{40} = 12,5$$

Finalmente, para a casela inferior direita da Tabela 10.2 (conservadores não-permissivos):

$$f_e = \frac{(20)(25)}{40} = \frac{500}{40} = 12,5$$

Como veremos mais adiante, o método que acabamos de apresentar pode ser aplicado a qualquer problema de qui-quadrado que implique obter freqüências esperadas (f_e).*

Ilustração

Para resumir passo a passo o procedimento que permite obter o qui-quadrado observado ou calculado (χ^2_o), vamos supor que quiséssemos estudar a relação entre "fumar maconha" e "desejo de cursar o segundo grau" (antigo "colegial"). Poderíamos especificar nossas hipóteses da seguinte forma:

H_o: *a relação entre "fumar maconha" (X) e "fazer o segundo grau" (Y) é de independência, isto é, X e Y são independentes.*
H_a: *a relação entre X e Y é de dependência.*

A fim de testar estas hipóteses ao nível de significância de 0,05, colhemos duas amostras de alunos de primeiro grau (preferentemente das últimas séries): uma, composta de 21 sujeitos cujo objetivo é "prosseguir os estudos" e, outra, com 15 sujeitos cuja decisão é "parar os estudos após conclusão do primeiro grau". Suponha-se que, após a realização da pesquisa com esses 36 sujeitos, resultassem os dados condensados na Tabela 10.3.
Como podemos observar nessa tabela, 15 dos 21 alunos que tencionam fazer o segundo grau fumam maconha; entre os que não pretendem fazer o segundo grau, apenas 5 fumam essa erva. Para descobrir se há uma relação significante entre "fumar maconha" e "continuar os estudos", os passos que devemos efetuar são os seguintes:

* N.T.: Tal afirmação nem sempre é verdadeira! Se o pesquisador estiver trabalhando com χ^2 *de aderência* (teste de), as freqüências teóricas (esperadas) são obtidas a partir da *lei* que rege o particular fenômeno que esteja sendo estudado. Por exemplo, num jogo de cara-ou-coroa, onde C = cara e K = coroa, suponha-se que C = 20 e K = 80, com N = 100 (obviamente). Então, se a moeda for honesta, a freqüência esperada para a variável C deverá ser igual à freqüência esperada para a variável K, ou seja:

	f_o	f_e
C →	20	50
K →	80	50
N →	100	100

Observe-se, pois, que a regra apresentada pelo autor não se aplica ao cálculo dessas freqüências, uma vez que o fenômeno estudado (isto é, "sair cara-ou-coroa") possui uma *lei própria*.

TABELA 10.3 Consumo de Maconha entre Alunos de Primeiro Grau e sua Relação com Objetivos Acadêmicos Posteriores

	Tenciona Cursar o 2.º Grau	
Fuma Maconha	*Sim (fo)*	*Não (fo)*
Sim	15	5
Não	6	10
Total	21	15

PASSO 1: Dispor os dados numa tabela de dupla entrada (tipo 2 X 2).

	Continuam os Estudos	*Param os Estudos*	
Fumantes	15 ( )	5 ( )	20
Não-fumantes	6 ( )	10 ( )	16
	21	15	$N = 36$

PASSO 2: Obter a freqüência esperada (teórica) para cada casela

	15 (11,67)	5 (8,33)	20
	6 (9,33)	10 (6,67)	16
	21	15	$N=36$

(superior esquerda) $f_e = \dfrac{(21)(20)}{36} = \dfrac{420}{36} = 11{,}67$

(superior direita) $f_e = \dfrac{(15)(20)}{36} = \dfrac{300}{36} = 8{,}33$

(inferior esquerda) $f_e = \dfrac{(21)(16)}{36} = \dfrac{336}{36} = 9{,}33$

(inferior direta) $f_e = \dfrac{(15)(16)}{36} = \dfrac{240}{36} = 6{,}67$

PASSO 3: Subtrair cada freqüência esperada da respectiva freqüência observada

	$fo - fe$
(superior esquerda)	$15 - 11{,}67 = 3{,}33$
(superior direita)	$5 - 8{,}33 = -3{,}33$
(inferior esquerda)	$6 - 9{,}33 = -3{,}33$
(inferior direita)	$10 - 6{,}67 = 3{,}33$

PASSO 4: Quadrar essas diferenças

	$(fo - fe)^2$
(superior esquerda)	$(3{,}33)^2 = 11{,}09$
(superior direita)	$(-3{,}33)^2 = 11{,}09$
(inferior esquerda)	$(-3{,}33)^2 = 11{,}09$
(inferior direita)	$(3{,}33)^2 = 11{,}09$

PASSO 5: Dividir o quadrado de cada diferença pela respectiva freqüência esperada

$$\frac{(fo - fe)^2}{fe}$$

(superior esquerda) $\dfrac{11{,}09}{11{,}67} = 0{,}95$

(superior direita) $\dfrac{11{,}09}{8{,}33} = 1{,}33$

(inferior esquerda) $\dfrac{11{,}09}{9{,}33} = 1{,}19$

(inferior direita) $\dfrac{11{,}09}{6{,}67} = 1{,}66$

PASSO 6: Somar esses quocientes para obter o qui-quadrado observado

$$\Sigma \frac{(fo - fe)^2}{fe}$$

0,95
1,33
1,19
1,66

$\chi^2 = 5{,}13$

PASSO 7: Achar o número de graus de liberdade

$$gl = (l - 1)(c - 1) = (2 - 1)(2 - 1) = (1)(1) = 1$$

PASSO 8: Comparar o valor do qui-quadrado observado (obtido) com o qui-quadrado crítico (tabelado). Veja Tabela E (fim do livro)

$$\begin{aligned} \chi^2 \text{ obtido (observado)} &= 5{,}13 \\ \chi^2 \text{ tabelado (crítico)} &= 3{,}84 \\ gl &= 1 \\ P &= 0{,}05 \end{aligned}$$

O passo 8 sugere que, para rejeitarmos a hipótese nula a um nível de significância de 0,05 e com 1 grau de liberdade, o nosso qui-quadrado observado deverá ser *igual a* ou *maior que* 3,84. Como o qui-quadrado que obtivemos é de 5,13 (isto é, $\chi_o^2 > \chi_c^2$), podemos rejeitar a hipótese nula e aceitar a hipótese experimental. Nossos resultados "sugerem" que a proporção de fumantes de maconha é maior entre os alunos que pretendem cursar o segundo grau do que entre os que pretendem parar os estudos no primeiro grau.*

O procedimento que acabamos de ilustrar, passo a passo, pode ser resumido, em forma tabular, como se segue:

	fo	fe	$fo - fe$	$(fo - fe)^2$	$\frac{(fo - fe)^2}{fe}$
(superior esquerda)	15	11,67	3,33	11,09	0,95
(superior direita)	5	8,33	−3,33	11,09	1,33
(inferior esquerda)	6	9,33	−3,33	11,09	1,19
(inferior direita)	10	6,67	3,33	11,09	1,66
					$\chi^2 = 5{,}13$

FÓRMULA PARA O CÁLCULO DO QUI-QUADRADO DE UMA TABELA 2 X 2

Podemos evitar o processo lento e trabalhoso de calcular as freqüências esperadas para um qui-quadrado associado a uma tabela 2 X 2 (2 linhas por 2 colunas), usando a seguinte fórmula:

$$\chi^2 = \frac{N(AD - BC)^2}{(A + B)(C + D)(A + C)(B + D)}\text{, onde}$$

A = freqüência observada para a casela superior esquerda
B = freqüência observada para a casela superior direita
C = freqüência observada para a casela inferior esquerda
D = freqüência observada para a casela inferior direita

* N.T.: Não se vá inferir daí que existe uma relação *obrigatória* de causa e efeito. Além disso, nada indica que essa relação seja *verdadeira*. Primeiro, porque se trata de uma *situação hipotética*; segundo, porque, ainda que fosse verdadeira a relação de dependência, ela seria, quando muito, verdadeira para um particular grupo de sujeitos, numa particular comunidade (norte-americana, no caso).

Numa tabela 2 X 2, a disposição das caselas (A, B, C e D), com os respectivos totais marginais, segue o modelo abaixo:

A	B	$A + B$
C	D	$C + D$
$A + C$	$B + D$	N

Para ilustrar o uso dessa nova fórmula, vamos voltar aos dados da Tabela 10.3 (uso de maconha vs. continuidade dos estudos), para a qual já obtivemos um χ^2 de 5,13. Podemos entrar na nova fórmula diretamente com as freqüências observadas. Assim:

15 A	5 B
6 C	10 D

aplicando a fórmula, vem que:

$$\chi^2 = (15 + 5(6 + 10)(15 + 6)(5 + 10)$$

$$\chi^2 = \frac{36[(15)(10) - (5)(6)]^2}{(15 + 5)(6 + 10)(15 + 6)(5 + 10)} = \frac{36(150 - 30)^2}{(20)(16)(21)(15)}$$

$$= \frac{36(120)^2}{100.800} = \frac{36(14.400)}{100.800} = \frac{518.400}{100.800} = 5{,}14$$

CORREÇÃO DE FREQÜÊNCIAS ESPERADAS PEQUENAS

Se as freqüências esperadas numa tabela 2 X 2 forem muito pequenas (menores que 10 em cada casela), as fórmulas que aprendemos até aqui poderão produzir um qui-quadrado "inflado", isto é, um valor de qui-quadrado *maior* do que o real. Note-se que isso só se aplica às freqüências *esperadas* e não às de fato obtidas no decorrer de uma pesquisa; estas, aliás, podem ser de qualquer tamanho.

A fim de reduzir o "tamanho" do qui-quadrado observado e, assim, chegar-se a um resultado mais "sensato", aplicamos a conhecida *correção de Yates*—sempre que a tabela for do tipo 2 X 2. A correção de Yates consiste em reduzir de meia unidade

(0,5) *todas* as diferenças entre cada freqüência observada e a teórica (esperada) respectiva. Uma vez que o χ^2 depende do tamanho dessas diferenças, conseguimos restringir, também, o tamanho do próprio qui-quadrado observado.* A fórmula "corrigida" do qui-quadrado, para situações em que as freqüências esperadas são muito pequenas, é a seguinte:

$$\chi^2 = \Sigma \frac{(|f_o - f_e| - 0{,}50)^2}{f_e}$$

Na fórmula "corrigida" (acima), as barras verticais que ladeiam $f_o - f_e$ indicam que meia unidade (0,5) deve ser subtraída do *valor absoluto* de cada diferença–ignorando-se os sinais –.

Vamos aplicar a fórmula corrigida aos dados da Tabela 10.3:

$$\chi^2 = \frac{(|15 - 11{,}67| - 0{,}50)^2}{11{,}67} + \frac{(|5 - 8{,}33| - 0{,}50)^2}{8{,}33}$$

$$+ \frac{(|6 - 9{,}33| - 0{,}50)^2}{9{,}33} + \frac{(|10 - 6{,}67| - 0{,}50)^2}{6{,}67}$$

$$= \frac{(3{,}33 - 0{,}50)^2}{11{,}67} + \frac{(-3{,}33 - 0{,}50)^2}{8{,}33} + \frac{(-3{,}33 - 0{,}50)^2}{9{,}33}$$

$$+ \frac{(3{,}33 - 0{,}50)^2}{6{,}67}$$

$$= \frac{(2{,}83)^2}{11{,}67} + \frac{(2{,}83)^2}{8{,}33} + \frac{(2{,}83)^2}{9{,}33} + \frac{(2{,}83)^2}{6{,}67}$$

$$= \frac{8{,}01}{11{,}67} + \frac{8{,}01}{8{,}33} + \frac{8{,}01}{9{,}33} + \frac{8{,}01}{6{,}67}$$

$$= 0{,}69 + 0{,}96 + 0{,}86 + 1{,}20 = 3{,}71$$

O procedimento acima, em que se aplica a fórmula "corrigida" do qui-quadrado, pode se resumido em fórmula tabular:

f_o	f_e	$\lvert f_o - f_e \rvert$	$\lvert f_o - f_e \rvert - 0{,}50$	$(\lvert f_o - f_e \rvert - 0{,}50)^2$	$\frac{(\lvert f_o - f_e \rvert - 0{,}50)^2}{f_e}$
15	11,67	3,33	2,83	8,01	0,69
5	8,33	3,33	2,83	8,01	0,96
6	9,33	3,33	2,83	8,01	0,86
10	6,67	3,33	2,83	8,01	1,20
					$\chi^2 = 3{,}71$

* N.T.: Essa operação é também conhecida pelo nome de *correção de continuidade*.

Como se observa acima, a correção de Yates produz um qui-quadrado ($\chi^2 = 3{,}71$) *menor* do que o obtido por meio das fórmulas não-corrigidas ($\chi^2 = 5{,}13$). No exemplo em foco, nossa decisão quanto à hipótese nula fica inteiramente a mercê do fato de se usar ou não a correção de Yates. Com a fórmula corrigida, aceitamos a hipótese nula (pois $3{,}71 < 3{,}84$), e com as fórmulas não-corrigidas, rejeitamos esta mesma hipótese.

A correção de Yates também se aplica à fórmula alternativa para o cálculo do qui-quadrado numa tabela 2 X 2. Assim:

$$\chi^2 = \frac{N(|AD - BC| - N/2)^2}{(A+B)(C+D)(A+C)(B+D)}$$

Retomando os dados da Tabela 10.3, vem que:

$$\chi^2 = \frac{36[|(15)(10) - (5)(6)| - 36/2]^2}{(15+5)(6+10)(15+6)(5+10)} = \frac{36(|150 - 30| - 18)^2}{(20)(15)(21)(15)}$$

$$= \frac{36(120 - 18)^2}{100.800} = \frac{36(102)^2}{100.800} = \frac{36(10404)}{100.800} = \frac{374.544}{100.800} = 3{,}71$$

COMPARAÇÃO DE VÁRIOS GRUPOS

Até aqui, limitamos nossas ilustrações ao esquema 2 X 2 amplamente empregado. Deve-se enfatizar, entretanto, que é freqüente calcular-se o qui-quadrado para tabelas maiores que as do tipo 2 X 2; nessas tabelas ingressam vários grupos ou categorias. O procedimento passo a passo usado na comparação* de vários grupos é, em essência, igual ao adotado na situação 2 X 2. Vamos ilustrar com um problema 3 X 3 (3 linhas por 3 colunas), embora tal procedimento se aplique para qualquer número de linhas ou colunas.

Imagine-se, mais uma vez, que o problema seja pesquisar a relação entre "orientação política" e "métodos de educação de crianças". Desta vez, porém, vamos admitir

* N.T.: As mesmas observações feitas anteriormente com respeito ao termo "comparação" valem para tabelas maiores que 2 X 2.

que três amostras aleatórias tenham sido extraídas: 32 liberais, 30 moderados e 27 conservadores. Suponha-se, ainda, que os métodos de educação de crianças estejam categorizados em: "permissivos", "moderados" e "autoritários". Então:

Hipótese Nula: A freqüência relativa de métodos permissivos, moderados e autoritários é a mesma tanto para liberais quanto para moderados quanto para conservadores.*

Hipótese Experimental: A freqüência relativa de métodos permissivos, moderados e autoritários não é a mesma para liberais, para moderados e para conservadores.

Vamos admitir que, terminada a coleta de dados, resulte a Tabela 10.4. Nela, verificamos que 7 dos 32 conservadores, 9 dos 30 moderados e 14 dos 27 liberais adotam práticas permissivas relativamente à educação de seus filhos.

TABELA 10.4 Educação de Filhos vs. Orientação Política (Problema 3 X 3)

Método de Educação de Filhos	*Orientação Política*		
	Conservadores *fo*	*Moderados* *fo*	*Liberais* *fo*
Permissivo	7	9	14
Moderado	10	10	8
Autoritário	15	11	5
Total	32	30	27

Tenha-se em mente que tanto a correção de Yates quanto a fórmula alternativa para o χ^2 são aplicáveis *apenas* quando se trata de uma tabela (problema) 2 X 2, não sendo possível, por isso, aplicá-las a vários grupos, como, por exemplo, o caso presente (3 X 3). Para determinar se há uma diferença significativa na Tabela 10.4, temos de utilizar a fórmula original de χ^2, apresentada no início deste capítulo, isto é:

$$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e}$$

* N.T.: Com o objetivo de não se perder de vista que o que se testa é a *independência entre as variáveis*, essas hipóteses podem ser reformuladas, a fim de obter-se maior precisão, para:

H_0: As variáveis P, M e A guardam *relação de independência* com as variáveis X, Y e Z.

H_a: As variáveis P, M e A guardam *relação de dependência* com as variáveis X, Y e Z.

P = método permissivo; M = método moderado; A = método autoritário; X = liberais; Y = moderados e Z = conservadores.

A fórmula acima pode ser empregada no problema 3 X 3 conforme se segue:

PASSO 1: Dispor os dados numa tabela de dupla entrada (tipo 3 X 3)

Métodos Usados na Educação de Crianças	*Orientação Política*			
	Conservadores	*Moderados*	*Liberais*	
Permissivo	7	9	14	30
Moderado	10	10	8 *Freqüência observada*	28
Autoritário	15	11	5	31
Total marginal	32	30	27	N = 89

PASSO 2: Calcular a freqüência esperada para cada casela

7 (10,79)	9 (10,11)	14 (9,10)	30
10 (10,07)	10 (9,44)	8 (8,49)	28
15 (11,41)	11 (10,45)	5 (9,40)	31
32	30	27	N = 89

(superior esquerda) $f_e = \frac{(30)(32)}{89} = \frac{960}{98}$

$= 10{,}79$

(intermediária esquerda) $f_e = \frac{(28)(32)}{89} = \frac{896}{89}$

$= 10{,}07$

(inferior esquerda) $f_e = \frac{(31)(32)}{89} = \frac{992}{89}$

$= 11{,}14$

(intermediária superior) $f_e = \frac{(30)(30)}{89} = \frac{900}{89}$

$= 10{,}11$

(central) $f_e = \frac{(28)(30)}{89} = \frac{840}{89}$

$= 9{,}44$

(intermediária inferior) $f_e = \frac{(31)(30)}{89} = \frac{930}{89}$

$= 10{,}45$

(superior direita) $f_e = \frac{(30)(27)}{89} = \frac{810}{89}$

$= 9{,}10$

(intermediária direita) $f_e = \frac{(28)(27)}{89} = \frac{756}{89}$

$= 8{,}49$

(inferior direita) $f_e = \frac{(31)(27)}{89} = \frac{837}{89}$

$= 9{,}40$

PASSO 3: Subtrair cada freqüência esperada da respectiva freqüência observada

	$f_o - f_e$
(superior esquerda)	$7 - 10{,}79 = -3{,}79$
(intermediária esquerda)	$10 - 10{,}07 = -0{,}07$
(inferior esquerda)	$15 - 11{,}14 = 3{,}86$
(intermediária superior)	$9 - 10{,}11 = -1{,}11$
(central)	$10 - 9{,}44 = 0{,}56$
(intermediária inferior)	$11 - 10{,}45 = 0{,}55$
(superior direita)	$14 - 9{,}10 = 4{,}90$
(intermediária direita)	$8 - 8{,}49 = -0{,}49$
(inferior direita)	$5 - 9{,}40 = -4{,}40$

PASSO 4: Quadrar essas diferenças

	$(f_o - f_e)^2$
(superior esquerda)	$(-3{,}79)^2 = 14{,}36$
(intermediária esquerda)	$(-0{,}07)^2 = 0{,}01$
(inferior esquerda)	$(3{,}86)^2 = 14{,}90$
(intermediária superior)	$(-1{,}11)^2 = 1{,}23$
(central)	$(0{,}56)^2 = 0{,}31$
(intermediária inferior)	$(0{,}55)^2 = 0{,}30$
(superior direita)	$(4{,}90)^2 = 24{,}01$
(intermediária direita)	$(-0{,}49)^2 = 0{,}24$
(inferior direita)	$(-4{,}40)^2 = 19{,}36$

PASSO 5: Dividir cada diferença, ao quadrado, pela respectiva freqüência esperada

	$\frac{(f_o - f_e)^2}{f_e}$
(superior esquerda)	$\frac{14{,}36}{10{,}79} = 1{,}33$
(intermediária esquerda)	$\frac{0{,}01}{10{,}07} = 0{,}00$
(inferior esquerda)	$\frac{14{,}90}{11{,}14} = 1{,}34$
(intermediária superior)	$\frac{1{,}23}{10{,}11} = 0{,}12$
(central)	$\frac{0{,}31}{9{,}44} = 0{,}03$
(intermediária inferior)	$\frac{0{,}30}{10{,}45} = 0{,}03$
(superior direita)	$\frac{24{,}01}{9{,}10} = 2{,}64$
(intermediária direita)	$\frac{0{,}24}{8{,}49} = 0{,}03$
(inferior direita)	$\frac{19{,}36}{9{,}40} = 2{,}06$

PASSO 6: Somar esses quocientes para obter o qui-quadrado observado (calculado)

$$\Sigma \frac{(f_o - f_e)^2}{f_e}$$

$$\begin{array}{rr} & 1{,}33 \\ & 0{,}00 \\ & 1{,}34 \\ & 0{,}12 \\ & 0{,}03 \\ & 0{,}03 \\ & 2{,}64 \\ & 0{,}03 \\ & \underline{2{,}06} \\ \chi^2 = & 7{,}58 \end{array}$$

PASSO 7: Achar o número de graus de liberdade

$$gl = (l - 1)(c - 1) = (3 - 1)(3 - 1) = (2)(2) = 4$$

PASSO 8: Comparar o valor do qui-quadrado obtido (observado) com o do qui-quadrado tabelado (crítico). Veja Tabela E

$$\begin{aligned} \chi^2 \text{ observado (obtido)} &= 7{,}58 \\ \chi^2 \text{ tabelado (crítico)} &= 9{,}49 \\ gl &= 4 \\ P &= 0{,}05 \end{aligned}$$

Necessitamos, portanto, de um valor de qui-quadrado que seja pelo menos igual a 9,49 para podermos rejeitar a hipótese nula. Uma vez que o nosso χ^2 obtido é de apenas 7,58, devemos aceitar a hipótese nula e atribuir as diferenças amostrais à ação do acaso. Não conseguimos descobrir nenhuma prova estatisticamente significante capaz de sustentar a idéia de que as freqüências relativas dos métodos de educação de filhos diferem para liberais, moderados e conservadores.

† QUI-QUADRADO DE ADERÊNCIA

Páginas atrás, demonstrou-se como a estatística qui-quadrado pode ser utilizada para testar hipóteses de independência entre duas variáveis.

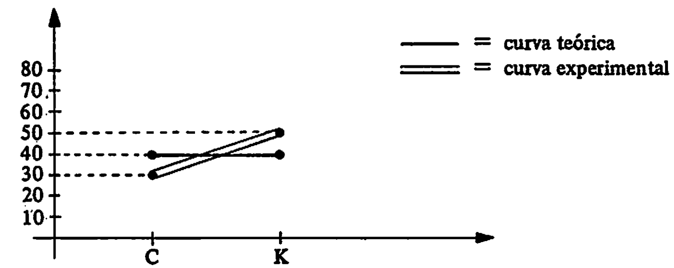
Há situações, entretanto, em que o interesse do pesquisador está voltado para problemas de *ajustamento*, isto é, problemas ligados ao *grau de adaptação* de duas

curvas — uma *experimental*, decorrente de *observação*, e outra *teórica*, resultante de uma *lei*.

Para exemplificar, vamos supor que uma moeda tenha sido lançada 80 vezes, produzindo os seguintes resultados: 30 caras (C) e 50 coroas (K). Com $\alpha = 5\%$, será possível afirmar que essa moeda é "honesta" (eqüiprovável)?

Busquemos, antes de mais nada, um suporte lógico para a resolução do problema. É intuitivo que, sendo eqüiprovável, essa moeda, ao cabo de $n = 80$ lançamentos, deveria ter produzido caras e coroas em quantidades muito próximas. Para *n grande* (e já começa a ser grande a partir de 30), a tendência é igual número de caras e coroas.

Graficamente, essas duas situações — a experimental e a teórica — assumem o seguinte aspecto:



Observação importante: A rigor, o gráfico deveria mostrar *pontos* e não, *linhas*, já que a variável é discreta.

Note-se que do "cruzamento" das linhas no gráfico resulta um ângulo. Pois bem, esse ângulo é *grande* ou *pequeno*? É também intuitivo (ou quase) que, se esse ângulo for pequeno, o ajustamento das duas curvas (linhas) é *bom*. Fala-se, nesse caso, em bom ajustamento ou boa *aderência*. Por igual raciocínio, se o ângulo for grande, o ajustamento das curvas será *ruim*.

Quando o ajustamento é bom, as pequenas oscilações do ângulo podem ser atribuídas ao *acaso*; quando o ajustamento é ruim, o acaso não é suficiente para explicar as diferenças e, aí, cabe pesquisar que fatores estão condicionando a falta de aderência.

Uma boa medida para avaliar-se o grau de ajustamento de duas curvas é a estatística denominada *qui-quadrado de aderência*, cuja fórmula, já conhecida, é

$$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e}$$

onde

f_o = freqüência observada por casela

f_e = freqüência esperada por casela.

Neste caso, a tabela não é do tipo 2 × 2, de modo que o cálculo dos graus de liberdade segue a seguinte fórmula:

$$gl = (l - 1)$$

onde l = número de linhas.
Assim:

	f_o	f_e
C	30	40
K	50	40
Σ	80	80

linhas

$$gl = (l - 1)$$
$$gl = (2 - 1) = 1$$

As hipóteses estatísticas do nosso exemplo são:

H_o: $P(C) = P(K)$ ⟵ Isto só ocorre se a moeda for "honesta".
H_a: $P(C) \neq P(K)$

O cálculo do χ_o^2 (qui-quadrado observado) segue em tudo o que já se viu páginas atrás. Assim:

	f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
C	30	40	− 10	100	100 ÷ 40 = 2,5
K	50	40	10	100	100 ÷ 40 = 2,5
	80	80	0		χ_o^2 = 5,000

Para $\alpha = 5\%$ e 1 gl, o $\chi_c^2 = 3{,}841$. Como o $\chi_o^2 = 5{,}000$, rejeita-se a H_o em favor da H_a. Tudo faz crer que essa moeda não é eqüiprovável. E o ajustamento (aderência) das duas curvas é ruim, isto é, o ângulo formado é grande demais para ser explicado apenas pela ação do acaso.

O cálculo das freqüências esperadas (f_e) depende da lei que rege o fenômeno. No caso de moedas (honestas), a lei é expressa por $p = \frac{1}{2}$; já no caso de dados, a lei é $p = \frac{1}{6}$. E para o cálculo de f_e basta aplicar a seguinte fórmula:

$$f_e = n \cdot p$$

onde

n = total de observações
p = probabilidade de cada observação (lei).

EXERCÍCIOS RESOLVIDOS

1. Um dado de seis faces foi jogado 120 vezes e produziu os seguintes resultados: (F_1 = 15), (F_2 = 16), (F_3 = 25), (F_4 = 20), (F_5 = 20), (F_6 = 24); onde F_i representa "face". Com α = 5%, será possível admitir que esse dado seja eqüiprovável?

Solução

H_o: $P(F_1) = P(F_2) = P(F_3) = P(F_4) = P(F_5) = P(F_6) = \frac{1}{6}$ ← Lei

H_a: Quase idêntica à H_o, só que, em algum ponto, o = passa a ≠.

Como o dado possui 6 faces, haverá 6 linhas na tabela. Então:

$$gl = (l - 1) = (6 - 1) = 5$$

E o $\chi_c^2 = 11{,}070$. (Consultar Tabela E no fim do livro).

Resultado	f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
F_1	15	20	− 5	25	1,250
F_2	16	20	− 4	16	0,800
F_3	25	20	5	25	1,250
F_4	20	20	0	0	0,000
F_5	20	20	0	0	0,000
F_6	24	20	4	16	0,800
	120	120	0		χ_o^2 = 4,100

Observar que as f_e foram obtidas assim:

$$\underset{(n)}{\text{total}} \times \underset{(p)}{\text{lei}} = 120 \cdot \frac{1}{6} = 20$$

Conclusão: como (χ_o^2 = 4,100) < (χ_c^2 = 11,070), H_o *não rejeitada*, ou seja, o dado pode ser considerado honesto (eqüiprovável).

2. Estudos não muito recentes demonstraram que em Buenos Aires os grupos sangüíneos estavam assim distribuídos:

Grupo Sangüíneo	%
O	45,6
A	39,4
B	10,4
AB	4,6

Colhida uma amostra aleatória de 300 sujeitos (em Buenos Aires), verificou-se que:

140 sujeitos eram do grupo O,
122 sujeitos eram do grupo A,
30 sujeitos eram do grupo B e
8 sujeitos eram do grupo AB.

Com $\alpha = 5\%$, vamos testar a hipótese de que os dados obtidos são estatisticamente aceitáveis.

Solução

Em primeiro lugar, precisamos determinar a *lei*. E a lei está expressa em termos porcentuais no corpo do enunciado do problema. Então:

H_o: $P(\text{O}) = 0{,}456 = p_1$
$P(\text{A}) = 0{,}394 = p_2$
$P(\text{B}) = 0{,}104 = p_3$
$P(\text{AB}) = 0{,}046 = p_4$

Esses valores (probabilidades) foram obtidos a partir da divisão das porcentagens por 100.

H_a: Semelhante à H_o, só que, em algum ponto, o = transforma-se em $\neq$.

Dispondo os dados numa tabela e fazendo os cálculos vem:

Tipo Sangüíneo	f_o	f_e*	$(f_o - f_e)$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
O	140	136,8	3,2	10,24	0,07485
A	122	118,2	3,8	14,44	0,12217
B	30	31,2	− 1,2	1,44	0,04615
AB	8	13,8	− 5,8	33,64	2,43768
	300	300,0	0		χ^2_o = 2,68085 ≅ 2,681

* Lembrar que as f_e são obtidas mediante a fórmula $n \cdot p_i$. Por exemplo: a f_e relativa ao grupo A é (300) (0,394) = 118,20.

Ora, $\chi^2_c = 7{,}815$ (com 3 graus de liberdade, ou seja, 4–1 linhas). E como $\chi^2_o < \chi^2_c$, a H_o vai ser ***não rejeitada***, donde aceitarmos a hipótese de que os valores obtidos são estatisticamente aceitáveis.

3. Admita-se a hipótese de que, em determinada região, 84% das pessoas de raça branca sejam portadoras de Rh^+. Suponha-se, ainda, que uma amostra de 500 sujeitos, todos brancos, tenha produzido os seguintes resultados:

Fator Rh	*Número de Portadores*
+	408
−	92

Testando-se a hipótese inicial com $\alpha_1 = 5\%$ e $\alpha_2 = 1\%$, qual a conclusão?

Solução

H_o: $P(Rh^+) = 0{,}84$ ou $P(Rh^-) = 0{,}16$

H_a: $P(Rh^+) \neq 0{,}84$ ou $P(Rh^-) \neq 0{,}16$

Cálculos:

Fator Rh	f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
+	408	420	− 12	144	0,3439
−	92	80	12	144	1,8000
	500	500	0		χ^2_o = 2,1439 ≅ 2,144

Ora, χ^2_c (1 gl, 5%) = 3,841. Logo H_o não rejeitada, pois $\chi^2_o < \chi^2_c$.

Ainda: χ^2_c (1 gl, 1%) = 6,635. Logo H_o não rejeitada, pois $\chi^2_o < \chi^2_c$.

Então, para os dois níveis (α_1 e α_2), a hipótese de que 84% da população branca daquela região é portadora de Rh^+ é estatisticamente aceitável.

4. Certo pesquisador, trabalhando experimentalmente com as Leis de Mendel, cruzou ervilhas *amarelas* (A) – algumas de pele *lisa* (L), outras de pele *rugosa* (r) – com ervilhas *verdes* (v) – também de pele *lisa* (L) algumas, e de pele *rugosa* (r) outras. No final do experimento, colheu 640 sementes assim classificadas:

AL = 320; Ar = 150; vL = 160; vr = 10

Com $\alpha = 0{,}05$, será possível afirmar que os dados obtidos desmentem a Segunda Lei de Mendel*?

* Para maiores informações, consulte-se Cleffi, Norma Maria; *Curso de Biologia - Biologia Celular, Genética e Evolução*; editora HARBRA ltda., SP, capítulo 6.

Solução

Praticamente todos os manuais de Genética, ao tratar das Leis de Mendel, relatam com minúcias o experimento referido no corpo do problema 4. E, dizem os manuais, num conjunto de 16 sementes, 9 são do tipo AL, 3, do tipo Ar, 3, do tipo vL e 1, do tipo vr.

Com essas informações, é fácil determinar a *lei* que rege o fenômeno. Basta procurar as *probabilidades* associadas a cada uma dessas combinações. Assim:

$$\left.\begin{aligned} P(\text{AL}) &= \frac{9}{16} \\ P(\text{Ar}) &= \frac{3}{16} \\ P(\text{vL}) &= \frac{3}{16} \\ P(\text{vr}) &= \frac{1}{16} \end{aligned}\right\} \quad \text{Notar que} \quad P(\text{AL}) + P(\text{Ar}) + P(\text{vL}) + P(\text{vr}) = 1$$

Então:

H_o: As sementes obtidas estão de acordo com a seguinte lei:

$$P(\text{AL}) = 9/16;\ P(\text{Ar}) = 3/16;\ P(\text{vL}) = 3/16;\ P(\text{vr}) = 1/16$$

H_a. As sementes obtidas não estão de acordo com a lei.

Cálculos:

Tipos de Semente	f_o	f_e*	$(f_o - f_e)$	$(f_o - f_e)^2$	$\dfrac{(f_o - f_e)^2}{f_e}$
AL	320	360	− 40	1.600	4,4444
Ar	150	120	30	900	7,5000
vL	160	120	40	1.600	13,3333
vr	10	40	− 30	900	22,5000
	640	640	0		$\chi_0^2 = 47,7777 \cong 47,778$

* O procedimento para o cálculo de f_e é o seguinte:

$$(640)(9/16) = 360$$
$$(640)(3/16) = 120$$
$$(640)(1/16) = 40$$

Ora, $\chi_c^2 = 7,815$ (com 3 gl, 5%). Como $\chi_o^2 > \chi_c^2$, a H_o vai ser rejeitada. Assim, tudo faz crer que os dados obtidos não estão de acordo com a Segunda Lei de Mendel. As razões? Falha na execução do experimento ou falha no caráter do experimentador...!

5. Testar, com $\alpha = 5\%$, se os dados do seguinte quadro têm distribuição normal com os parâmetros 32,2 (média aritmética) e 213,51 (variância).

X (*Notas*)	n_i
0 ⊢ 10	15
10 ⊢ 20	25
20 ⊢ 30	40
30 ⊢ 40	60
40 ⊢ 50	35
50 ⊢ 60	20
60 ⊢ 70	5
	200

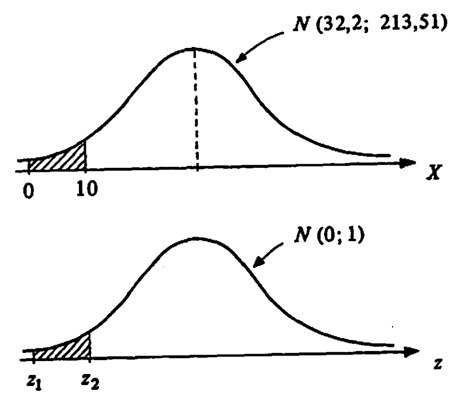
Solução

H_o: X é $N(32,2; 213,51)$.

H_a: X não é $N(32,2; 213,51)$. (Embora possa sê-lo com *outros* parâmetros!)

A primeira coisa que precisamos fazer é transformar os valores de X em valores de z. Assim:

Intervalo 0 —— 10



Então, utilizando, para a transformação de X em z,

$$z = \frac{(X_i \pm 0,5) - \mu}{\sigma}, \text{ vem:}$$

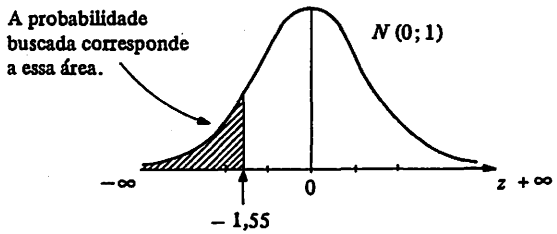
$$z_2 = \frac{10 - 0,5 - 32,2}{\sqrt{213,51}} = \frac{9,5 - 32,2}{14,6} = \frac{-22,7}{14,6} = -1,55$$

E fazendo o mesmo para todos os limites superiores aparentes, isto é, 20, 30, ..., 70, vem:

0 —— 10	corresponde a	$-\infty$ —— $-1,55$
10 ⊢— 20	corresponde a	$-1,55$ ⊢— $-0,87$
20 ⊢— 30	corresponde a	$-0,87$ ⊢— $-0,18$
30 ⊢— 40	corresponde a	$-0,18$ ⊢— $0,50$
40 ⊢— 50	corresponde a	$0,50$ ⊢— $1,18$
50 ⊢— 60	corresponde a	$1,18$ ⊢— $1,87$
60 ⊢— 70	corresponde a	$1,87$ ⊢— $+\infty$

A esses valores de z correspondem probabilidades determinadas. Assim:

$$P(0 < X < 10) = P(X < 10) = P(z < -1,55)$$

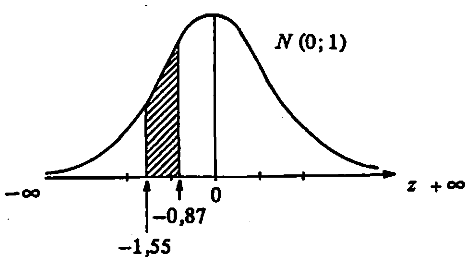


Então:

A probabilidade de 0 até $-\infty$	é 0,5000
A probabilidade de 0 até $-1,55$	é 0,4394
A probabilidade da área hachurada	é 0,0606

De modo análogo:

$$P(10 \leqslant X < 20) = P(-1,55 \leqslant z < -0,87)$$



De 0 a $-1,55$, a probabilidade	é 0,4394
De 0 a $-0,87$, a probabilidade	é 0,3078
A probabilidade da área hachurada	é 0,1316

Por igual raciocínio, chegaremos às demais probabilidades, conforme quadro a seguir:

X	z_i	$P(z_i)$
0 —— 10	$-\infty$ —— $-1,55$	0,0606
10 ⊢— 20	$-1,55$ ⊢— $-0,87$	0,1316
20 ⊢— 30	$-0,87$ ⊢— $-0,18$	0,2364
30 ⊢— 40	$-0,18$ ⊢— $0,50$	0,2629
40 ⊢— 50	$0,50$ ⊢— $1,18$	0,1895
50 ⊢— 60	$1,18$ ⊢— $1,87$	0,0882
60 ⊢— 70	$1,87$ ⊢— $+\infty$	0,0307

Se multiplicarmos agora cada $P(z)$ por $\Sigma n_i = 200$, obteremos os n_i *esperados* – que poderão ser, mediante uma Prova de Qui-quadrado, comparados com os n_i *observados*. Assim:

$n_i = f_o$	$\tilde{n}_i = f_e = P(z_i)(\Sigma n_i)$	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
15	(0,0606) (200) = 12,12	2,88	8,2944	0,6844
25	(0,1316) (200) = 26,32	−1,32	1,7424	0,0662
40	(0,2364) (200) = 47,28	−7,28	52,9984	1,1209
60	(0,2629) (200) = 52,58	7,42	55,0564	1,0471
35	(0,1895) (200) = 37,90	−2,90	8,4100	0,2219
20	(0,0882) (200) = 17,64	2,36	5,5696	0,3157
5	(0,0307) (200) = 6,14	−1,14	1,2996	0,2117
200	≅ 200,00	+12,66 −12,64 ≅ 0,00		$\chi^2_o = 3,6679$ ↑ ≅ 3,668

Atenção, agora! Os graus de liberdade dessa nova tabela devem ser calculados da seguinte maneira: cada estatística calculada (média, variância) "rouba" 1 grau de liberdade; além disso, o fato de Σn_i ser fixo "rouba" mais 1 grau. Daí resulta que:

$$7 - (\underset{\text{Média}}{\underset{\uparrow}{1}} + \underset{\text{Variância}}{\underset{\uparrow}{1}} + \underset{\Sigma n_i}{\underset{\uparrow}{1}}) = 7 - 3 = 4 \text{ gl}$$

Então: χ^2_c (4 gl e 5%) = 9,488. (Ver Tabela E.)

Como: $(\chi^2_o = 3{,}668) < (\chi^2_c = 9{,}488)$, H_o *não rejeitada.*

Conclusão: pode-se admitir que os dados do problema têm distribuição normal com os parâmetros indicados.

REQUISITOS PARA O USO DO QUI-QUADRADO

Independentemente do fato de que os testes não-paramétricos não pressupõem distribuição normal de uma particular variável na população, eles apresentam, também, uma série de requisitos que devem ser levados em conta pelo pesquisador desejoso de proceder a uma escolha inteligente de algum teste de significância. O leitor notará, entretanto, que os requisitos para o uso de testes não-paramétricos são, via de regra, preenchidos com mais facilidade—pelo menos quando comparados com os "correspondentes" paramétricos, tais como a estatística t ou a análise de variância. Com isso em mente, vamos voltar nossa atenção para alguns dos mais importantes requisitos para o uso da prova de significância chamada qui-quadrado:

1. *"Comparação"*entre duas ou mais amostras.*
   Como procurou ilustrar este capítulo, o teste de qui-quadrado é usado para, através da comparação entre um χ^2 observado e um χ^2 crítico (tabelado), levar o pesquisador a decidir se, numa amostra aleatória, duas ou mais variáveis guardam entre si uma *relação de independência*. Isso torna obrigatório que tenhamos, no mínimo, uma tabela 2 X 2 (no mínimo 2 linhas, como também 2 colunas). Na montagem da tabela, é importante que o *mesmo* sujeito *não* figure em mais de uma casela. A amostra deve ser explorada (com relação às variáveis observacionais) exaustivamente e a distribuição dos sujeitos pelas várias caselas deve ser independente**.
2. *Dados pertencentes ao nível nominal de mensuração.*
   O pesquisador só necessita de freqüências, isto é, do número de sujeitos observados para cada situação.

* N.T.: O termo "comparação", como foi anteriormente salientado, deve ser interpretado. "Comparação" supõe um cotejo entre o *resultado obtido* (observado) e o *resultado teórico* (tabelado). No caso de qui-quadrado, a um dado nível de significância e com certo número de graus de liberdade, a "comparação" se faz entre χ^2_o e χ^2_c.

** N.T.: Assim, se numa amostra aleatória de $N = 100$ sujeitos, desejarmos pesquisar a relação entre cor de olhos (claro/escuro) e cor de cabelos (claro/escuro), o mesmo sujeito não pode, relativamente a esse par de variáveis, figurar em mais de *uma* casela. Essa circunstância torna a utilização do qui-quadrado em situações experimentais do tipo *antes-e-depois* algo *restritiva*, embora *não impeditiva*. Em Edwards, Allen L. – *Statistical Analysis*, 3ª ed., Holt, Rinehart and Winston, Inc., págs. 158-159, cita-se o exemplo de um pesquisador que aplicou aos mesmos 90 sujeitos amostrais dois testes: um, *antes* do tratamento e outro, *depois*.

3. *Amostragem aleatória.*
   Os sujeitos amostrais devem ter sido extraídos *aleatoriamente* de uma particular população.
4. *As freqüências esperadas (teóricas) por casela não devem ser muito pequenas.*
   O tamanho mínimo de uma f_e depende da natureza do problema. Numa tabela 2 X 2, nenhuma freqüência esperada deveria ser menor que 5. Além disso, é aconselhável que se use a correção de Yates sempre que, numa tabela 2 X 2, qualquer freqüência esperada seja inferior a 10. Numa situação em que vários grupos estejam sendo "comparados" (por exemplo, 3 X 3 ou 4 X 5), não há uma regra rígida para o estabelecimento de freqüências teóricas mínimas por casela, embora seja recomendável que bem poucas contenham menos que 5. Note-se, porém, que, em qualquer situação, a soma de todas as freqüências esperadas deverá ser igual à soma de todas as freqüências observadas, ou seja:

$$\Sigma f_e = \Sigma f_o$$

†5. Como regra prática, já que este assunto é algo controvertido, acrescente-se a tudo o que ficou dito mais o seguinte: a prova de qui-quadrado deve ser *evitada* sempre que qualquer $f_e < 5$ e, *simultaneamente*, $(\Sigma f_e = \Sigma f_o = N) < 30$. Assim, quando $N < 30$, aconselha-se a usar a *Prova Exata de Fisher.*

† PROVA EXATA DE FISHER

Há situações em que se torna muito difícil conseguir, como ficou dito acima, que $\Sigma f_o \geqslant 30$ e, simultaneamente, que todas as $f_e \geqslant 5$. Quando isso ocorre – e sempre numa tabela 2 × 2 – a Prova Exata de Fisher permite resolver o problema.

Seja a tabela abaixo, onde $n < 30$:

Var. II / *Var. I*	*IIa*	*IIb*	*Totais*
Ia	A	B	A+B
Ib	C	D	C+D
Totais	A+C	B+D	n

Var. I = Variável I, dicotomizada em Ia e Ib.
Var. II = Variável II, dicotomizada em IIa e IIb.

Demonstra-se, por recorrência a uma distribuição chamada *Hipergeométrica*, que, sendo fixos os totais marginais:

$$P(X = i) = \frac{(A+B)!\ (C+D)!\ (A+C)!\ (B+D)!}{n!\ A!\ B!\ C!\ D!}$$

As probabilidades obtidas pela aplicação dessa fórmula referem-se *sempre* a *uma* das extremidades do espaço experimental (R), de sorte que, no caso de uma H_a bicaudal, deverão ser calculadas probabilidades para *ambos* os extremos.

Vamos supor que estejamos interessados em estudar se existe relação de dependência entre as variáveis *sexo* e *teor de fumo*. Vamos supor, também, que só tenhamos conseguido uma amostra de tamanho $n = 16$. Feitas as dicotomizações essenciais, resulta a seguinte tabela de dupla entrada – com a qual vamos trabalhar:

Teor / *Sexo*	*Fraco* (f)	*Forte* (F)	*Totais*
Masculino (H)	4	3	7
Feminino (M)	2	7	9
Totais	6	10	16

As hipóteses que nos interessam são:

$$H_O: P(\text{f/H}) = P(\text{f/M}) \quad \text{ou} \quad P(\text{F/H}) = P(\text{F/M})$$

Isto é: se, variando o sexo, não variar a preferência pelo teor do fumo, é possível admitir que as variáveis *sexo* e *teor* sejam *independentes*.

$$H_a: P(\text{f/H}) \neq P(\text{f/M}) \quad \text{ou} \quad P(\text{F/H}) \neq P(\text{F/M})$$

Seja ainda $\alpha = 5\%$.

O primeiro problema que surge é: *quantos pontos* tem o *espaço experimental* (R) e *que valores* correspondem a esses pontos?

O segundo problema diz respeito à *definição operacional* da *variável de observação* (X). Essa variável, por uma questão de comodidade, terá a definição que resulte do cruzamento das filas de *menores* totais. Assim:

	f	*F*		
H	4	3	7 ←	menor total de *linha*
M	2	7	9	
	6	10	16	
	↑ menor total de *coluna*			

Então, X será definido como "número de homens que preferem baixos teores".

Quanto à determinação do espaço experimental (R), o problema resolve-se por ensaio-e-erro: é substituir o 2 (da casela A) por 0, 1, 3, ..., até que resulte um absurdo, isto é, a tabela "não feche". Mas, *cuidado*: os totais marginais são *fixos*!

$X = 0$

0	7	7
6	3	9
6	10	16

(a)

$X = 1$

1	6	7
5	4	9
6	10	16

(b)

$X = 3$

2	5	7
4	5	9
6	10	16

(c)

$X = 5$

5	2	7
1	8	9
6	10	16

(d)

$X = 6$

6	1	7
0	9	9
6	10	16

(e)

$X = 7$

7	0	7
?	?	9
6	10	16

(f)

Aqui aparece o absurdo mencionado anteriormente: qualquer valor igual a ou maior que 7 "estoura" o total de coluna – que é obrigatoriamente (neste exemplo) igual a 6.

Assim, o espaço experimental compõe-se dos seguintes valores:

R: 0, 1, 2, 3, 4, 5, 6 homens que preferem baixos teores;

donde resulta, também, que R tem 7 pontos, ou seja: 0, 1, 2, 3, 4, 5 e 6.

Agora é calcular as probabilidades associadas a cada valor de R, cuidando para que em nenhuma das extremidades a soma das probabilidades ultrapasse 2,5%. Por quê? Porque estabelecemos, no início, que α seria 5% (embora pudesse ser outro valor); além disso, sendo a H_a bicaudal – e é bicaudal porque fala em *diferenças* ($\neq$) sem, entretanto, fixar o *sentido* da variação – o α deve ser partido ao meio, de modo que cada extremidade de R, abranja, no máximo, $\frac{\alpha}{2}$.

Da aplicação da fórmula resultam as seguintes probabilidades:

$$\text{Tabela (a)} \rightarrow P(X = 0) = \frac{7!\,9!\,6!\,10!}{16!\,0!\,7!\,6!\,3!} = \frac{10!\,9!\,\cancel{7!}\,\cancel{6!}}{16!\,\cancel{7!}\,\cancel{6!}\,3!\,0!} =$$

$$= \frac{\cancel{10!}\cdot\cancel{3!}\cdot\overset{1}{\cancel{4}}\cdot\overset{1}{\cancel{5}}\cdot\overset{3}{\cancel{6}}\cdot\overset{1}{\cancel{7}}\cdot\overset{1}{\cancel{8}}\cdot\overset{\overset{1}{\cancel{3}}}{\cancel{9}}}{\cancel{10!}\cdot 11\cdot\underset{\underset{1}{\cancel{3}}}{\cancel{12}}\cdot 13\cdot\underset{\underset{1}{\cancel{7}}}{\cancel{14}}\cdot\underset{\underset{1}{\cancel{3}}}{\cancel{15}}\cdot\underset{2}{\cancel{16}}\cdot\cancel{3!}\,(1)} = \frac{3}{11\cdot 13\cdot 2} =$$

$$= \frac{3}{286} = 0{,}0104895 \cong 0{,}0105$$

$\therefore\ P(X = 0) \cong 0{,}0105$, isto é, 1,05% (< 2,5%)

$$\boxed{\text{Então: } X = 0 \text{ serve.}}$$

Tabela (b) $\rightarrow P(X = 1) = \dfrac{7!\ 9!\ 6!\ 10!}{16!\ (1)!\ 6!\ 5!\ 4!} = \dfrac{10!\ 9!\ 7!\ \cancel{6!}}{16!\ \cancel{6!}\ 5!\ 4!\ 1} =$

$$= \frac{\cancel{10!}\,\cancel{5!}\,6\cdot 7\cdot 8\cdot 9\cdot\cancel{4!}\,5\cdot 6\cdot 7}{\cancel{10!}\,11\cdot 12\cdot 13\cdot 14\cdot 15\cdot 16\cdot\cancel{5!}\,\cancel{4!}} =$$

$$= \frac{\overset{1}{\cancel{6}}\cdot\overset{1}{\cancel{7}}\cdot\overset{1}{\cancel{8}}\cdot 9\cdot\overset{1}{\cancel{5}}\cdot\overset{\overset{1}{\cancel{3}}}{\cancel{6}}\cdot 7}{11\cdot\underset{2}{\cancel{12}}\cdot 13\cdot\underset{2}{\cancel{14}}\cdot\underset{\underset{1}{\cancel{3}}}{\cancel{15}}\cdot\underset{\underset{1}{\cancel{2}}}{\cancel{16}}} = \frac{9\cdot 7}{11\cdot 2\cdot 13\cdot 2} =$$

$$= \frac{63}{572} = 0{,}1101398 \cong 0{,}1101$$

$\therefore\ P(X = 1) \cong 0{,}1101$, isto é, 11,01% (> 2,5%)

$$\boxed{\text{Então: } X = 1 \text{ não serve.}}$$

Vamos agora calcular as probabilidades associadas aos maiores valores de R, ou seja, 6, 5, 4 etc.

Tabela (e) $\rightarrow P(X = 6) = \dfrac{7!\,9!\,6!\,10!}{16!\ 6!\ (1)!\,0!\ 9!} = 0{,}0008741 \cong 0{,}0009$

$\therefore\ P(X = 6) \cong 0{,}0009$, isto é, 0,09% (< 2,5%)

$$\boxed{\text{Então: } X = 6 \text{ serve.}}$$

Tabela (d) $\rightarrow P(X = 5) = \dfrac{7!\,9!\,6!\,10!}{16!\ 5!\ 2!\ (1)!\,8!} = 0{,}0236013 \cong 0{,}0236$

$\therefore\ P(X = 5) \cong 0{,}0236$, isto é, 2,36% (< 2,5%)

Observar também que

$$P[(X = 6) \text{ ou } (X = 5)] = P(X = 6) + P(X = 5) = 0{,}09 + 2{,}36 = \underset{\underset{<\ 2{,}5\%}{\uparrow}}{2{,}45\%}$$

$$\boxed{\text{Então: } X = 5 \text{ serve.}}$$

É evidente que não precisamos continuar calculando probabilidades, pois

$[P(X = 4) + P(X = 5) + P(X = 6)] > 2{,}5\%$.

Desse modo, R ficou dividido em três partes, a saber:

$$R\colon \boxed{0,}\ \overbrace{1,\ 2,\ 3,\ 4,}^{R_c^*}\ \boxed{5,\ 6}$$

$$R_c$$

R_c = região crítica, isto é, região que leva à rejeição da H_o se $(X = i)$ pertencer a ela.

R_c^* = região não crítica, isto é, região que leva à não rejeição da H_o se $(X = i)$ pertencer a ela.

Ora, nas condições do nosso problema, $X = 4$. E como $(X = 4) \in R_c^*$, a conclusão é: H_o *não rejeitada.* Voltando agora à H_o, vemos que ela admite a relação de *independência* entre as variáveis *sexo do fumante* e *preferência por baixos teores.* E essa hipótese "resistiu" ao teste, não sendo rejeitada.

A esta altura, podemos acrescentar mais o seguinte: o nível de significância *teórico* (α) era de 5%, mas o nível *efetivo*, ou seja, aquele com o qual de fato trabalhamos, foi de 1,05% + 0,09% + 2,36% = 3,5%. Bem menor que 5%!!

OBSERVAÇÕES IMPORTANTES

1. Definir a variável de observação, X, a partir do cruzamento das filas de menores totais apresenta uma grande vantagem: R começa em 0 (zero).

2. Ao colher sua amostra, o pesquisador tem, quase sempre, condições de fazer $(A + B) = (C + D)$. Esse procedimento acarreta-lhe a vantagem de não ter de calcular as probabilidades associadas a ambas as extremidades de R (espaço experimental): basta calcular os valores de um lado e dobrá-los.

3. Existem tabelas que dispensam o cálculo das probabilidades. Essas tabelas, entretanto, são muito extensas e, por isso, pouco confortáveis.

EXERCÍCIOS RESOLVIDOS

1. Com $\alpha = 5\%$, será possível admitir a hipótese de independência entre as variáveis *remédio* e *resultado*?

Resultado / *Remédio*	*Curados*	*Não Curados*	*Totais*
Tomaram	4	1	5
Não tomaram	2	3	5
Totais	6	4	10

Solução

A variável de observação, X, vai ser definida pelo cruzamento das filas de menores totais. Assim:

Resultado / *Remédio*	C	$\tilde{C}$	
T		○	5
$\tilde{T}$			5
	6	4	

Aqui é indiferente, pois 5 = 5. Vamos adotar o 5 que corresponde a T.

$4 < 6$

Então, X = número de pessoas que, embora tenham tomado o remédio, não se curaram.

H_0: Os resultados (curas/não-curas) independem de tratamento com um particular remédio (tomaram/não-tomaram),

ou

$$P(\tilde{C}|T) = P(\tilde{C}|\tilde{T})$$
$$P(C|T) = P(C|\tilde{T})$$

H_a: Os resultados dependem de tratamento com um particular remédio. (Em símbolos, basta que, nas igualdades acima, um dos sinais de = transforme-se em ≠.)

A determinação de R (espaço experimental) faz-se por tentativas sucessivas:

5	0	5
1	4	5
6	4	10

3	2	5
3	2	5
6	4	10

2	3	5
4	1	5
6	4	10

1	4	5
5	0	5
6	4	10

	5	5
	?	5
6	4	10

Este valor "estoura" o quadro.

Então:

R: 0, 1, 2, 3, 4 (Pessoas que não se curaram, embora tenham tomado o medicamento.)

O nosso problema, então, consiste em verificar se $(X = 1)$ – conforme enunciado do problema – permite concluir pela rejeição da H_0.

Aplicando a fórmula conhecida, vem:

$$P(X=0) = \frac{5!\ 5!\ 6!\ 4!}{10!\ 5!\ 0!\ (1)!\ 4!} \cong 0{,}024, \text{ ou seja, } 2{,}4\%$$

$2{,}4\% < 2{,}5\%$

$$P(X=4) = \frac{5!\ \ 5!\ \ 6!\ \ 4!}{10!\ \ (1)!\ \ 4!\ \ 5!\ \ 0!} \cong 0{,}024, \text{ ou seja, } 2{,}4\%$$

Este resultado não precisaria ter sido calculado. Ver *Observações Importantes*, § 2.

2,4% < 2,5%

$$P(X=1) = \frac{5!\ \ 5!\ \ 6!\ \ 4!}{10!\ \ 4!\ \ (1)!\ \ 2!\ \ 3!} \cong 0{,}238, \text{ ou seja, } 23{,}8\%$$

Como 23,8% > 2,5%, $X = 1$ não serve.

Com esses cálculos, R já pode ser dividido em R_c e R_c^*:

$R_c^* \leftarrow$ Se $(X=i) \in R_c^* \rightarrow H_o$ não rejeitada.

R: $\boxed{0,}\ \boxed{1,\ \ 2,\ \ 3,}\ \boxed{4}$

$R_c \leftarrow$ Se $(X=i) \in R_c \rightarrow H_o$ rejeitada.

Ora, $(X = 1) \in R_c^*$. Logo H_o não rejeitada, o que eqüivale a dizer que, nesse particular exemplo, as curas são independentes da ingestão do medicamento.

2. Para testar se determinada droga era capaz de inibir a absorção de álcool pelo organismo humano, realizou-se um experimento com a participação de 15 voluntários (homens, saudáveis, idades entre 25 e 28 anos). Uma hora após a ingestão da droga "milagrosa", todos os voluntários tomaram 2 doses de uísque. Uma hora mais tarde, colheu-se uma amostra do sangue de cada sujeito. Os resultados estão na tabela abaixo:

Teste Revelou / *Droga "Milagrosa"*	*Presença de Álcool* (+)	*Ausência de Álcool* (−)	
Tomaram (T)	2	8	10
Não-tomaram ($\widetilde{T}$)	4	1	5
	6	9	15

Qual a conclusão, com $\alpha = 5\%$?

Solução

X vai ser definida assim: número de voluntários que *não* tomaram a droga, mas cujo teste sangüíneo deu +. (É só cruzar as filas de menores totais.)

Então:

$H_o: P(+ \mid \widetilde{T}) = P(+ \mid T)$
$H_a: P(+ \mid \widetilde{T}) \neq P(+ \mid T)$

Determinação de R:

6	4	10
0	5	5
6	9	15

5	5	10
1	4	5
6	9	15

4	6	10
2	3	5
6	9	15

3	7	10
3	2	5
6	9	15

1	9	10
5	0	5
6	9	15

		10
6	?	5
6	9	15

Este valor "estoura" a tabela.

Então:

R: 0, 1, 2, 3, 4, 5 resultados + sem ingestão da droga.

Como H_a é bicaudal, α deverá ser repartido de forma que em cada extremidade de R haja uma parte de α. (Lembrete: se, na extremidade esquerda de R houver α_1 e, na direita, α_2, $(\alpha_1 + \alpha_2) \leqslant 5\%$.)

A seguir, mediante a aplicação da fórmula já conhecida, vem:

$$P(X=0) = \frac{10!\ \ 5!\ \ 6!\ \ 9!}{15!\ \ 6!\ \ 4!\ \ 0!\ \ 5!} \cong 0{,}042 \ldots\ldots (4{,}2\%)$$

$$P(X=5) = \frac{10!\ \ 5!\ \ 6!\ \ 9!}{15!\ \ (1)!\ 9!\ \ 5!\ \ 0!} \cong 0{,}002 \ldots\ldots (0{,}2\%)$$

$$P(X=1) = \frac{10!\ \ 5!\ \ 6!\ \ 9!}{15!\ \ 5!\ \ 5!\ \ (1)!\ \ 4!} \cong 0{,}252 \ldots\ldots (25{,}2\%)$$

$$P(X=4) = \frac{10!\ \ 5!\ \ 6!\ \ 9!}{15!\ \ 2!\ \ 8!\ \ 4!\ \ (1)!} \cong 0{,}045 \ldots\ldots (4{,}5\%)$$

Um simples exame dessas probabilidades mostra que a única maneira de dividir R, nas condições do problema, é:

$$R:\ \overbrace{\boxed{0,}}^{R_c}\ \overbrace{\boxed{1,\quad 2,\quad 3,\quad 4,}}^{R_c^*}\ \overbrace{\boxed{5}}^{R_c}$$

$$P(X=0)=0{,}042 \qquad\qquad P(X=5)=0{,}002$$

$$P(X=0)+P(X=5)=0{,}042+0{,}002=0{,}044\ \ldots\ldots\ (4{,}4\%<5\%).$$

Então, $(X=4)\in R_c^*$, donde H_o não rejeitada. Tudo leva a crer que não há relação alguma entre a droga "milagrosa" e a absorção de álcool pelo organismo.

O TESTE DA MEDIANA

O qui-quadrado pode ser aplicado a qualquer número de amostras independentes mensuradas no nível *nominal*. Quando os dados forem *ordinais*, o *teste da mediana* constitui um procedimento não-paramétrico simples capaz de determinar o grau de "certeza" com que duas amostras aleatórias foram extraídas de populações que têm a *mesma mediana*.

Para ilustrar o uso do teste da mediana, admitamos que certo pesquisador desejasse estudar reações masculinas/femininas a uma particular situação social tida como embaraçosa. Para criar a situação embaraçosa, pediu a 15 homens e 12 mulheres–cuja "vocação" para cantores era apenas "média"–que cantassem, individualmente, diante de uma audiência de "peritos", algumas canções do repertório popular. O número de minutos que cada sujeito manteve-se disposto a continuar cantando figura na tabela abaixo, onde, quanto menor o número, maior a intensidade da sensação de embaraço.*

Número de Minutos Durante os quais cada Sujeito Cantou			
Homens	*Mulheres*	*Homens*	*Mulheres*
15	12	11	9
18	7	10	11
15	15	8	14
17	16	14	9
17	6	9	
16	8	18	
10	10	16	
13	6		

*N.T.: O exemplo serve *apenas* para ilustrar o procedimento denominado *teste da mediana*. Do ponto de vista experimental, o delineamento está cheio de impropriedades.

PASSO 1: Achar a mediana *depois* de fundir as duas amostras num só grupo. Em Símbolos:

$$\text{Posição da mediana} = \frac{N+1}{2} = \frac{27+1}{2} = 14.^{\text{o}} \text{ dado da tabela}$$

A mediana é o 14.º escore contado a partir de qualquer extremo da tabela–cujos dados deverão ser ordenados *a priori*.

Para encontrar a mediana, dispomos todos os escores (de homens e mulheres) em ordem consecutiva (sem levar em conta a origem da amostra) e localizamos a mediana "combinada":

18
18
17
17
16
16
16
15
15
15
14
14
13
12 ← Mediana (14º escore contado a partir de qualquer extremo)
11
11
10
10
10
9
9
9
8
8
7
6
6

PASSO 2: Contar em cada amostra o número de sujeitos que se situam "acima da mediana" e "não acima da mediana".*

* N.T.: "Não acima da mediana" não é o mesmo que "abaixo da mediana". "Não acima da mediana" vai *até* Mdn = 12, isto é, o 12 pertence a essa categoria.

	Homens f	Mulheres f
Acima da Mediana	10	3
Não Acima da Mediana	5	9
	$N = 27$	

Como vimos, os valores (duração do canto) da tabela anterior, para cada amostra, quando cotejados com a mediana da distribuição, são transformados em freqüências, originando um quadro 2 X 2. No problema em foco, 10 dos 15 homens e apenas 3 das 12 mulheres continuaram a cantar por um período superior ao representado pela mediana do grupo maior (homens + mulheres).

PASSO 3: Realizar um teste de qui-quadrado para verificar se o resultado é significante. Se não houver diferenças ligadas ao sexo com respeito à duração do canto (e ao presumido embaraço que dele decorre), podemos esperar que ambas as amostras se "quebram" no *mesmo* ponto mediano, de sorte que metade dos homens e das mulheres caem *acima* da mediana. Para determinar se as diferenças obtidas entre homens e mulheres são estatisticamente significantes ou apenas resultado de erros de amostragem, fazemos um teste de χ^2.

	Homens	*Mulheres*
Acima da Mediana	10 (*A*)	3 (*B*)
Não Acima da Mediana	5 (*C*)	9 (*D*)
	$N = 27$	

$$\chi^2 = \frac{N(|AD - BC| - N/2)^2}{(A+B)(C+D)(A+C)(B+D)} = \frac{27[\,|(10)(9) - (3)(5)| - \frac{27}{2}\,]^2}{(10+3)(5+9)(10+5)(3+9)}$$

$$= \frac{27(75 - 13{,}5)^2}{32.760} = \frac{102.120{,}75}{32.760} = 3{,}12$$

Consultando a Tabela E no fim do livro, verificamos que o χ^2 deve igualar-se a ou exceder 3,84 (com gl = 1) a fim de que possa ser, ao nível de 0,05, encarado como significante. Visto que o nosso χ^2 obtido é igual a 3,12, não é possível rejeitar a hipótese nula. Há insuficiência de provas para concluirmos, com base nos nossos resultados, que homens e mulheres diferem com respeito a suas reações diante da situação socialmente embaraçosa chamada "cantar em público".

Requisitos para o Uso do Teste da Mediana

As seguintes condições devem ser satisfeitas a fim de tornar possível o uso adequado do teste da mediana numa pesquisa:

1. *Comparação entre duas ou mais medianas independentes.*
   O teste da mediana é usado em comparações entre duas ou mais medianas originárias de amostras independentes.*
2. *Dados ordinais.*
   Para que seja possível aplicar o teste da mediana, é preciso que os dados pertençam, no mínimo, ao nível ordinal. Dados nominais não podem ser usados.
3. *Amostragem aleatória.*
   Dada uma população, as amostras devem ser extraídas aleatoriamente.

† PROVA DE MANN-WHITNEY (Prova *U*)

Dadas duas amostras, A_1 e A_2, de tamanhos n_1 e n_2, respectivamente, é possível, mediante a prova de Mann-Whitney, decidir se $A_1 \in \pi$ e $A_2 \in \pi$, isto é, se *ambas* as amostras podem ser consideradas provenientes da *mesma* população.

Nas condições acima estabelecidas, a estatística t – Prova t de Student –, pode ser usada desde que os dados tenham sido mensurados, *no mínimo*, no nível *intervalar*. Além disso, as amostras devem ser *aleatórias*, *independentes* e a *variável observacional* precisa ter distribuição *normal* (na população).

Em Ciências Humanas, esses requisitos nem sempre são conseguidos e o recurso consiste em lançar mão do teste não-paramétrico denominado Mann-Whitney, cuja única exigência é que os dados pertençam ao nível *ordinal*.

São três os casos em que se desdobra a prova U de Mann-Whitney, todos ligados aos tamanhos das amostras utilizadas. Assim, estabelecida a convenção de que $n_2 > n_1$, temos:

1.º caso: $n_2 \leqslant 8$;
2.º caso: $9 \leqslant n_2 \leqslant 20$;
3.º caso: $n_2 > 20$.

1.º caso: $n_2 \leqslant 8$

Um grupo de 5 adolescentes, escolhidos aleatoriamente, examina, durante 10 minutos, uma relação de nomes de objetos concretos. Em seguida, cada um dos adolescentes procura recompor, de memória e por escrito, a relação original, com a única restrição de que o tempo para essa tarefa seria igual para todos.

Outro grupo, composto de 4 adolescentes, também escolhidos aleatoriamente, examina a mesma relação durante 5 minutos e tenta, a seguir, da mesma forma que o primeiro grupo, reproduzir a lista de memória. A este grupo foi concedido o mesmo tempo que ao primeiro.

* N.T.: A rigor, não é bem isso. A *mediana* foi usada como ponto de partida para a *definição operacional de uma particular variável.* No exemplo estudado acima, o que o pesquisador fez foi testar se a variável "embaraço" estava associada, numa relação de dependência, à variável "sexo".

Na tabela abaixo, figuram os erros cometidos pelos sujeitos dos dois grupos.

Queremos testar, com $\alpha = 5\%$, se existe significativa diferença de desempenho entre os dois grupos relativamente à variável memória associada a tempo de estudo.

T_A $n_1 = 4$	T_B $n_2 = 5$
12	10
19	14
8	15
25	9
	18

T_A = Tratamento A = memória associada a 5 min de estudo.

T_B = Tratamento B = memória associada a 10 min de estudo.

A primeira coisa que precisamos fazer é juntar todos os dados num grupo só, ordenando-os do menor ao maior (ou vice-versa). A seguir, identificar os escores (dados) por A ou B, conforme o tratamento de que resultem: Assim:

8	9	10	12	14	15	18	19	25
A	B	B	A	B	B	B	A	A

Calculemos agora o número de vezes que cada valor do grupo A é precedido de valores do grupo B; calculemos também o número de vezes que cada valor do grupo B é precedido de valores do grupo A.

8	9	10	12	14	15	18	19	25
A	B	B	A	B	B	B	A	A
Precedido de 0 (zero) B	Precedido de 1 A	Precedido de 1A	Precedido de 2 B's	Precedido de 2 A's	Precedido de 2 A's	Precedido de 2 A's	Precedido de 5 B's	Precedido de 5 B's

A variável U_o de Mann-Whitney corresponde à menor das somas de precedência; a maior das somas designa-se por U_o'. Assim:

$$0 + 2 + 5 + 5 = 12 \quad \text{(A's precedidos de B's)}$$
$$1 + 1 + 2 + 2 + 2 = 8 \quad \text{(B's precedidos de A's)}$$

Então: $U_o = 8$ e $U_o' = 12$

Daqui para a frente, trabalharemos somente com o *menor* desses valores ($U_o < U_o'$) e passaremos a denominá-lo "*U* observado".

Desde logo, é conveniente observar a seguinte regra prática: $U_o = n_1 n_2 - U_o'$. Assim:

$$U_o = (4)(5) - 12 = 20 - 12 = 8$$

A regra de decisão para o caso específico de $n_2 \leqslant 8$ é a seguinte: *rejeitar a H_o sempre que $P(U_o) < \alpha$*. Os valores críticos para esse tipo de decisão estão na Tabela L (pág. 372).

Para a comodidade do leitor, reproduzimos abaixo a tabela que contém as probabilidades críticas adequadas ao nosso problema.

$n_2 = 5$ ← Tamanho da amostra *maior*.

Tamanho da amostra *menor*.

U \ n_1	1	2	3	4	5
0	.167	.047	.018	.008	.004
1	.333	.095	.036	.016	.008
2	.500	.190	.071	.032	.016
3	.667	.286	.125	.056	.028
4		.429	.196	.095	.048
5		.571	.286	.143	.075
6			.393	.206	.111
7			.500	.278	.155
8			.607	.365	.210
9				.452	.274
10				.548	.345
11					.421
12					.500
13					.579

Valor de U_o → 8

$P(U_o = 8) = 0{,}365$

Temos, então:

H_o: Tratamento T_B = Tratamento T_A (ou seja, as diferenças não são significativas).

H_a: Tratamento T_B > Tratamento T_A (ou seja, as diferenças são significativas, sendo que o T_B é *melhor* que o T_A)*

Conclusão: como $[P(U_o = 8) = 0{,}365] > (\alpha = 0{,}05)$, H_o *não rejeitada*, isto é, não há evidência estatística de que os tratamentos sejam diferentes.

2º caso: $9 \leqslant n_2 \leqslant 20$

Uma classe de 26 alunos foi dividida aleatoriamente em $n_1 = 10$ alunos (Grupo A) e $n_2 = 16$ alunos (Grupo B). O grupo A estudou regular e diariamente determinado assunto até as vésperas da prova. O grupo B ocupou-se de *outras* atividades e só estudou para a prova à sua véspera.

* As probabilidades constantes da tabela acima referem-se a uma H_a *unicaudal*. Para o caso de uma H_a bicaudal, as probabilidades devem ser *dobradas*.

A tabela abaixo contém as notas que cada aluno tirou nessa prova.

Grupo A $n_1 = 10$	*Grupo B* $n_2 = 16$
8	6
6,5	8
9	6
9,5	6,5
8	7
5	5
7,5	10
7	3,5
10	4
6	4,5
	9
	9
	1,5
	2
	7
	5

Analisar, com $\alpha = 5\%$, se existe diferença entre os dois tratamentos (isto é, métodos de estudo).

A primeira coisa que temos de fazer é ordenar esses valores, do menor ao maior e atribuir-lhes *postos* conforme a *ordem* ocupada na *seqüência geral*. À menor nota vamos atribuir o posto 1, à nota seguinte, o posto 2, à nota seguinte, o posto 3, e assim sucessivamente até a última nota, que é a maior. Quando houver *empates*, isto é, notas *iguais*, o posto será representado pela média aritmética dos postos que teriam sido atribuídos se não existissem empates.

Assim:

Grupo A		*Grupo B*	
Notas	*Postos*	*Notas*	*Postos*
8	19	6	10
6,5	12,5	8	19
9	22	6	10
9,5	24	6,5	12,5
8	19	7	15
5	7	5	7
7,5	17	10	25,5
7	15	3,5	3
10	25,5	4	4
6	10	4,5	5
	171 ↑ P_1	9	22
		9	22
		1,5	1
		2	2
		7	15
		5	7
			180 ← P_2

Somam-se em seguida os postos relativos a cada grupo (A e B). A soma dos postos do grupo A é 171 e a soma dos postos do grupo B, 180.

Calculemos agora U_o. Para tanto, vamos primeiro entrar nas duas seguintes fórmulas:

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - P_1$$

$$U = n_1 n_2 + \frac{n_2(n_2+1)}{2} - P_2,$$

onde

U = estatística U de Mann-Whitney;
n_1 = tamanho da amostra menor;
n_2 = tamanho da amostra maior;
P_1 = soma dos postos do grupo A;
P_2 = soma dos postos do grupo B.

$$U = (10)(16) + \frac{10(10+1)}{2} - 171 = 160 + \frac{(10)(11)}{2} - 171 = 160 + \frac{110}{2} - 171 = 160 + 55 - 171 = 44$$

$$U = (10)(16) + \frac{16(16+1)}{2} - 180 = 160 + \frac{(16)(17)}{2} - 180 = 160 + \frac{272}{2} - 180 = 160 + 136 - 180 = 116$$

$U_o = 44$, pois U_o é, por convenção, o *menor* dos valores*.

A decisão agora se toma com o auxílio da Tabela M (pág. 377). Para comodidade do leitor, essa tabela está reproduzida abaixo:

* É sempre conveniente fazer a prova já mencionada páginas atrás: $U_o = n_1 n_2 - U'_o$. Então:

$$U_o = (10)(16) - 116 = 160 - 116 = \underset{\substack{\uparrow \\ U_o}}{44}$$

Tabela de U_c ($U_c = U$ crítico)
$\alpha = 0{,}025$ (H_a unicaudal) e $\alpha = 0{,}05$ (H_a bicaudal)

n_1 \ n_2	9	10	11	12	13	14	15	16	17	18	19	20
1												
2	0	0	0	1	1	1	1	1	2	2	2	2
3	2	3	3	4	4	5	5	6	6	7	7	8
4	4	5	6	7	8	9	10	11	11	12	13	13
5	7	8	9	11	12	13	14	15	17	18	19	20
6	10	11	13	14	16	17	19	21	22	24	25	27
7	12	14	16	18	20	22	24	26	28	30	32	34
8	15	17	19	22	24	26	29	31	34	36	38	41
9	17	20	23	26	28	31	34	37	39	42	45	48
10	20	23	26	29	33	36	39	42	45	48	52	55
11	23	26	30	33	37	40	44	47	51	55	58	62
12	26	29	33	37	41	45	49	53	57	61	65	69
13	28	33	37	41	45	50	54	59	63	67	72	76
14	31	36	40	45	50	55	59	64	67	74	78	83
15	34	39	44	49	54	59	64	70	75	80	85	90
16	37	42	47	53	59	64	70	75	81	86	92	98
17	39	45	51	57	63	67	75	81	87	93	99	105
18	42	48	55	61	67	74	80	86	93	99	106	112
19	45	52	58	65	72	78	85	92	99	106	113	119
20	48	55	62	69	76	83	90	98	105	112	119	127

Tamanho da amostra maior.

Tamanho da amostra menor.

U crítico $= U_c$

Regra de decisão: *rejeitar a H_o sempre que $U_o < U_c$.*

Então, retomando o problema, temos:

H_o: Tratamento A = Tratamento B
H_a: Tratamento A ≠ Tratamento B (H_a bicaudal)

Conclusão: como $(U_o = 44) > (U_c = 42)$, H_o *não rejeitada*, isto é, não há evidência estatística de que os tratamentos sejam diferentes.

3º caso: $n_2 > 20$.

À medida que n_2 cresce, a distribuição de U tende rapidamente à normal (normal reduzida, onde $\mu = 0$ e $\sigma^2 = 1$). Assim, para $n_2 > 20$, passamos a usar a Tabela B em conjunto com a seguinte fórmula:

$$z_o = \frac{U_o - \dfrac{n_1 n_2}{2}}{\sqrt{\dfrac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

PROBLEMA

Certo professor aplicou o seguinte procedimento a uma classe de 30 elementos: 21 alunos foram por ele chamados pelos próprios nomes, durante um semestre, contingentemente à apresentação das lições de casa; os restantes 9 alunos, por igual período, foram chamados pelo professor de "você", contingentemente à apresentação das lições de casa. Esse professor admitia que, estimulado pelo próprio nome, o aluno era capaz de melhorar seu desempenho acadêmico – desempenho que foi mensurado em termos de notas escolares. Com $\alpha = 5\%$, será possível afirmar que era correta a hipótese desse professor? O quadro a seguir apresenta as notas dos 30 alunos no fim do semestre em que se realizou o experimento.

Tratamento A = T_A — *Alunos chamados por seus próprios nomes.* $n_2 = 21$		*Tratamento B = T_B* — *Alunos chamados pelo pronome "você".* $n_1 = 9$	
Notas	*Postos*	*Notas*	*Postos*
6,5	18	6,5	18
8,0	25,5	3,5	2,5
8,5	27,5	6,0	13
10,0	29,5	7,5	23,5
8,5	27;5	6,0	13
4,0	4,5	3,0	1
7,0	21,5	7,0	21,5
6,0	13	5,5	9,5
5,5	9,5	6,5	18
	176,5	6,0	13
	↑	6,5	18
	P_1	5,0	7
		5,0	7
		6,0	13
		3,5	2,5
		6,5	18
		10,0	29,5
		8,0	25,5
		7,5	23,5
		4,0	4,5
		5,0	7
			288,5
			↑
			P_2

H_o: $T_A = T_B$
H_a: $T_A > T_B$, isto é, tratamento A melhor que tratamento B. (Logo, a H_a é unicaudal.)

Para podermos entrar na fórmula atrás indicada, precisamos, antes, calcular U_o. Assim:

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - P_1$$

$$U = (9)(21) + \frac{9(9+1)}{2} - 176{,}5$$

$$= 189 + 45 - 176{,}5$$

$$U = 57{,}5$$

$$U = n_1 n_2 + \frac{n_2(n_2+1)}{2} - P_2$$

$$U = (9)(21) + \frac{21(21+1)}{2} - 288{,}5 =$$

$$= 189 + 231 - 288{,}5$$

$$U = 131{,}5$$

Como o U_o é o *menor* dos valores, vem:

$$U_o = 57{,}5$$
$$U_o' = 131{,}5$$

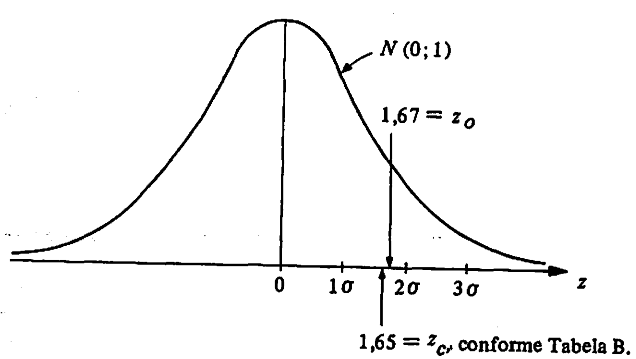
Prova: $U_o = n_1 n_2 - U_o'$

$$U = (9)(21) - 131{,}5 = 189 - 131{,}5 = \underset{\substack{\uparrow \\ U_o}}{57{,}5}$$

Então:

$$z_o = \frac{57{,}5 - \frac{(9)(21)}{2}}{\sqrt{\frac{(9)(21)(9+21+1)}{12}}} = \frac{57{,}5 - 94{,}5}{\sqrt{\frac{5.859}{12}}} = \frac{-37}{\sqrt{488{,}25}} = \frac{-37}{\pm 22{,}1} \cong \mp 1{,}67$$

Vamos agora montar um pequeno gráfico para facilitar a interpretação dos resultados.



Regra de decisão: *rejeitar H_o se $z_o > z_c$ (unicaudal direita)* ou se $z_o < z_c$ *(unicaudal esquerda)*.

Como $(z_o = 1{,}67) > (z_c = 1{,}65)$ ou $(z_o = -1{,}67) < (z_c = -1{,}65)$, decorre que a H_o *vai ser rejeitada.* Tudo parece indicar que chamar os alunos pelos próprios nomes produz melhores resultados que chamá-los pelo pronome "você".

DUPLA ANÁLISE DA VARIÂNCIA POR POSTOS: χ^2 DE FRIEDMAN

No Capítulo 8, introduzimos uma variação da estatística *t* passível de ser usada na comparação de dados resultantes da mesma amostra em dois momentos distintos. No esquema "antes-e-depois", por exemplo, o grau de hostilidade numa amostra de crianças poderia ser mensurado em duas ocasiões distintas: *antes* e *depois* de terem assistido a um violento programa de televisão.

A dupla análise de variância por postos—o χ^2 de Friedman (χ^2_F)—constitui uma aproximação não-paramétrica que permite testar diferenças numa mesma amostra de respondentes que tenham sido mensurados sob, pelo menos, duas condições distintas.

Em símbolos:

$$\chi^2_F = \frac{12}{Nk(k+1)} \Sigma (\Sigma R_i)^2 - 3N(k+1) \qquad \text{onde}$$

k = número de tratamentos, (via de regra representam o número de "condições" em que foram feitas as mensurações dos respondentes)
N = tamanho da amostra (número de respondentes)
ΣR_i = soma dos postos relativos a uma particular mensuração (via de regra representam a soma dos postos relativos a um particular tratamento).

Ilustração

Para ilustrar a aplicação da dupla análise de variância por postos de Friedman, suponha-se que desejemos testar a hipótese de que a hostilidade em crianças varia segundo o grau de violência a que se exponham assistindo a programas de televisão. A fim de estudar a influência da violência televisada, imaginemos que tenhamos condição de expor uma amostra aleatória de dez crianças a três diferentes níveis de violência televisada, num programa que seja essencialmente o mesmo quanto a todos os outros aspectos.* Vamos também admitir que obtivemos os seguintes escores de hostilidade dessas 10 crianças, nas três diferentes condições a que estiveram expostas (a escala vai de 20 a 60, e quanto mais alto o escore, maior a hostilidade:

* N.T.: Em linguagem experimental, fala-se, geralmente, em variáveis de três tipos (muito embora existam outros): variáveis *independentes* (VI), *dependentes* (VD) e *relevantes* (VR). As variáveis independentes são *manipuladas*; as dependentes são *observadas* e *mensuradas*; as relevantes são *mantidas sob controle*. No exemplo em foco, $N = 10$ crianças foram submetidas às VIs "tipos de programas violentos"; as VDs são as "reações individuais" e as VRs constituem as "demais condições ambientais".

PASSO 1: Atribuir postos a todas as condições (tratamentos) a que cada respondente esteve exposto (cada coluna corresponde a um tratamento). A fim de realizarmos a análise de variância de Friedman, trabalhamos diretamente com os postos atribuídos a cada respondente até esgotar-se a lista de mensurações.[2]

Como foi possível verificar acima, o nível de hostilidade da criança A aumentou de 23 para 30 e desse valor para 32 à medida que a "dose" de violência televisada à que ela estava sendo exposta passou de "baixa" para "mediana" e de "mediana" para "alta". Pensando em termos de postos, o escore-hostilidade da criança A foi mais alto (1) no tratamento "violência alta" (coluna A), mediano (2) no tratamento "violência mediana" (coluna M) e baixo (3) no tratamento "violência baixa" (coluna B). Prosseguindo na lista, o mesmo raciocínio aplica-se à criança B: posto 1 para o escore-hostilidade maior (coluna M), posto 2 para o escore-hostilidade mediano (coluna A) e posto 3 para o escore-hostilidade menor (coluna B). Com relação à criança C, o quadro é o seguinte: posto 1 para o escore da coluna A (39), posto 2 para o escore da coluna B (36) e posto 3 para o escore da coluna M (35).* Os postos atribuídos a cada criança sob os diferentes tratamentos figuram na tabela

	Níveis de Violência Televisada		
Criança	*Baixa (B)*	*Mediana (M)*	*Alta (A)*
A	23	30	32
B	41	45	43
C	36	35	39
D	28	29	35
E	39	41	47
F	25	28	27
G	38	46	51
H	40	47	49
I	45	46	42
J	29	34	38

PASSO 2: Somar os pontos relativos a cada tratamento (cada coluna corresponde a um tratamento). Se a hipótese nula estiver correta–e, por isso, não existir diferença significativa entre os tratamentos–,podemos esperar que as somas dos postos sejam iguais entre si, descartados os erros devidos ao acaso. No exemplo em causa, há três condições: violência televisada *baixa*, *mediana* e *alta*. Os postos e respectivas somas totais para cada uma dessas condições (isto é, tratamentos) são:

[2] Na presente ilustração não apareceram postos "espelhados" (empatados). No caso de tais postos ocorrerem (por exemplo, se o escore de hostilidade da criança A tivesse sido precisamente o mesmo para dois ou mais níveis de violência), adote-se o procedimento recomendado para postos "espelhados", Capítulo 11, sob o título "Coeficiente de Correlação de Postos".

* N.T.: Tanto faz que a atribuição de postos se faça, relativamente aos escores, do maior para o menor ou vice-versa.

Criança	*Baixa (B)*	*Posto*	*Mediana (M)*	*Posto*	*Alta (A)*	*Posto*
A	23	3	30	2	32	1
B	41	3	45	1	43	2
C	36	2	35	3	39	1
D	28	3	29	2	35	1
E	39	3	41	2	47	1
F	25	3	28	1	27	2
G	38	3	46	2	51	1
H	40	3	47	2	49	1
I	45	2	46	1	42	3
J	29	3	34	2	38	1

Criança	*(B) Posto*	*(M) Posto*	*(A) Posto*
A	3	2	1
B	3	1	2
C	2	3	1
D	3	2	1
E	3	2	1
F	3	1	2
G	3	2	1
H	3	2	1
I	2	1	3
J	3	2	1
	$\Sigma R = 28$	$\Sigma R = 18$	$\Sigma R = 14$

PASSO 3: Entrar na fórmula para obter o χ^2_p

$$\chi^2_p = \frac{12}{Nk(k+1)} \Sigma (\Sigma R_i)^2 - 3N(k+1)$$

$$= \frac{12}{(10)(3)(3+1)} (28^2 + 18^2 + 14^2) - 3(10)(3+1)$$

$$= \frac{12}{120} (784 + 324 + 196) - 120$$

$$= 0{,}10(1.304) - 120 = 130{,}4 - 120 = 10{,}4$$

PASSO 4: Achar o número de graus de liberdade

$$gl = k - 1 = 3 - 1 = 2$$

PASSO 5: Comparar o χ_p^2 (obtido, observado, calculado) com o χ^2 crítico (tabelado) na Tabela E

$$\begin{aligned} \chi_p^2 \text{ (obtido)} &= 10{,}4 \\ \chi^2 \text{ (tabelado)} &= 5{,}99 \\ gl &= 2 \\ P &= 0{,}05 \end{aligned}$$

O χ_p^2 constitui um valor (de qui-quadrado) que decorre diretamente da soma dos postos para todos os tratamentos. Como conseqüência, podemos comparar o χ_p^2 obtido com o χ^2 tabelado (Tabela E). Com gl = 2 e ao nível de significância de 5%, necessitamos de um qui-quadrado igual a, no mínimo, 5,99 para podermos rejeitar a hipótese nula. Como o χ_p^2 obtido é igual a 10,4, rejeitamos a hipótese nula e aceitamos a hipótese experimental. Acabamos de reunir, com isso, suficientes provas de que a violência televisada realmente exerce influência sobre o comportamento hostil de crianças. Há, pois, diferenças significantes na hostilidade à medida que a violência torna-se mais intensa.

Requisitos para o Uso da Dupla Análise da Variância por Postos (χ^2 de Friedman)

O uso da dupla análise da variância por postos–χ^2 de Friedman–implica que as seguintes condições sejam satisfeitas:

1. *Comparação de uma mesma amostra mensurada sob duas ou mais condições.* O teste de Friedman não pode ser aplicado para testar diferenças entre amostras independentes; ao contrário, ele pressupõe que a mesma amostra de respondentes tenha sido mensurada pelo menos *duas vezes* (ou, então, que os membros de duas ou mais amostras tenham sido aglutinados em função de variáveis específicas).
2. *Dados ordinais.* O teste só pode ser realizado com dados aos quais possam ser atribuídos postos.
3. *O número de respondentes (isto é, o tamanho da amostra) não deve ser muito pequeno.* O tamanho mínimo de N depende do número de tratamentos (k) aos quais os respondentes vão ser expostos. Por exemplo, N deve ser igual a ou maior que 10 quando $k = 3$; já com $k = 4$, $N \geqslant 5$.

ANÁLISE DA VARIÂNCIA POR POSTOS: H DE KRUSKAL-WALLIS

A análise da variância de Kruskal-Wallis é uma alternativa não-paramétrica à análise que se faz por recorrência à estatística F, e pode ser usada para comparar várias amostras independentes desde que os dados sejam de, no mínimo, nível ordinal. Para podermos aplicar a prova de Kruskal-Wallis, devemos calcular a estatística H como se segue:

$$H = \frac{12}{N(N+1)} \sum \left[\frac{(\Sigma R_i)^2}{n} \right] - 3(N+1) \qquad \text{onde}$$

N = número total de dados ou de respondentes
n = número de respondentes por amostra
ΣR_i = soma dos postos por amostra.

Ilustração

Para ilustrar a utilização da prova de Kruskal-Wallis, considere-se a possível influência da variável "idade" sobre a variável "habilidade (capacidade) para arranjar emprego". Vamos supor que, para estudar esse problema, tenhamos extraído amostras aleatórias de populações constituídas, respectivamente, de sujeitos velhos, de meia-idade e de jovens, e que a todos tenha sido dado igual número de dias (prazo) para encontrar emprego. Admitamos que os dados obtidos sejam os seguintes:

Número de Dias Antes de Encontrar Emprego		
Adultos Velhos (X_1)	*Adultos de Meia-idade* (X_2)	*Adultos Jovens* (X_3)
(n = 7) 63	(n = 8) 33	(n = 6) 25
20	42	31
43	27	6
58	28	14
57	51	18
71	64	13
45	12	
	30	

PASSO 1: Atribuir postos ao grupo total de escores (N = 21) e achar a soma desses postos por amostra. Os postos atribuídos aos escores *devem* obedecer ao seguinte critério: ao posto 1 corresponde o *menor* escore (6) e ao posto 21, o *maior* escore (71). Na presente ilustração, tanto 6 quanto 71 representam dias.[3]

X_1	*Posto*	X_2	*Posto*	X_3	*Posto*
63	19	33	12	25	7
20	6	42	13	31	11
43	14	27	8	6	1
58	18	28	9	14	4
57	17	51	16	18	5
71	21	64	20	13	3
45	15	12	2		$\Sigma R_3 = 31$
	$\Sigma R_1 = 110$	30	10		
			$\Sigma R_2 = 90$		

[3] Não apareceram postos "espelhados" (empatados) na ilustração presente. No caso de tais postos ocorrerem (por exemplo, se duas pessoas levarem exatamente 24 dias para encontrar um emprego), seguir o procedimento apresentado no Capítulo 11, sob o título "Coeficiente de Correlação de Postos".

PASSO 2: Entrar na fórmula para achar H.

$$H = \frac{12}{N(N+1)} \sum \left[\frac{(\Sigma R_i)^2}{n}\right] - 3(N+1)$$

$$= \left(\frac{12}{21(21+1)}\right)\left(\frac{110^2}{7} + \frac{90^2}{8} + \frac{31^2}{6}\right) - 3(21+1)$$

$$= \left(\frac{12}{462}\right)\left(\frac{12100}{7} + \frac{8100}{8} + \frac{961}{6}\right) - 66$$

$$= (0{,}03)(1.728{,}57 + 1.012{,}50 + 160{,}17) - 66$$

$$= (0{,}03)(2.901{,}24) - 66 = 87{,}04 - 66 = 21{,}04$$

PASSO 3: Calcular o número de graus de liberdade.

$$gl = k - 1 = 3 - 1 = 2$$

PASSO 4: Comparar H com o qui-quadrado crítico (Tabela E).

$$\begin{aligned} H &= 21{,}04 \\ \chi^2 \text{ crítico (tabelado)} &= 5{,}991 \\ gl &= 2 \\ P &= 0{,}05 \end{aligned}$$

Para rejeitarmos a hipótese nula ao nível de significância de 0,05, com 2 graus de liberdade, nosso H teria que ser igual a ou maior que 5,991. Como obtivemos $H = 21{,}04$, podemos rejeitar a hipótese nula em favor da hipótese alternativa (experimental). Nossos dados revelam que existem diferenças significativas, por idade, na quantidade de dias necessária para arranjar-se um emprego.

Requisitos para o Uso da Análise da Variância de Kruskal-Wallis

A prova de Kruskal-Wallis pressupõe as seguintes condições para seu uso adequado:

1. *Comparação de três ou mais amostras independentes.*
   A prova de Kruskal-Wallis *não* pode ser usada para testar diferenças numa *única* amostra de respondentes mensurados mais de uma vez.
2. *Dados ordinais.*
   Essa prova exige dados que possam ser ordenados e aos quais, por isso mesmo, seja possível atribuir postos.
3. *O tamanho mínimo de cada amostra deve ser 6.*
   Quando $n > 5$ por grupo de respondentes, a significância de H pode ser determinada por recorrência à Tabela E (qui-quadrados críticos). Para testar diferenças entre amostras de tamanho inferior a 6, o leitor deve recorrer a tabelas especiais encontradas em Siegel (1956).

† SUBSEQÜÊNCIAS

Vamos supor que o Departamento de Tráfego deseje saber se a ocorrência de veículos em baixa velocidade (ou em alta velocidade) constitui um fenômeno aleatório. Para a coleta de material para a pesquisa, instala um observador, munido de radar, num ponto estratégico de determinada estrada.

Definindo-se operacionalmente *baixa* velocidade (B) como qualquer velocidade igual ou inferior a 60 km/h, resulta o seguinte critério (convencional):

B $\leqslant$ 60 km/h
A $>$ 60 km/h

A partir desse critério, deseja-se saber se, num determinado período, a ocorrência de veículos com a seguinte distribuição

A A A B A A B B B A B B B B B A A B B B B

permite afirmar que tanto os veículos lentos quanto os rápidos ocorrem em blocos definidos e, portanto, nada aleatórios.

Chamando de *subseqüência* qualquer conjunto de elementos *iguais*, sem interrupção, resulta que os 21 veículos observados fornecem 8 subseqüências, conforme ilustrado abaixo:

$\underbrace{A\ A\ A}_{1}\ \underbrace{B}_{2}\ \underbrace{A\ A}_{3}\ \underbrace{B\ B\ B}_{4}\ \underbrace{A}_{5}\ \underbrace{B\ B\ B\ B\ B}_{6}\ \underbrace{A\ A}_{7}\ \underbrace{B\ B\ B\ B}_{8}$

Ora, esses mesmos 21 veículos, repartidos em 8 rápidos (A) e 13 lentos (B), poderiam ter-se distribuído de várias outras maneiras. Por exemplo:

a) $\underbrace{A}_{1}\ \underbrace{B}_{2}\ \underbrace{A}_{3}\ \underbrace{B}_{4}\ \underbrace{A}_{5}\ \underbrace{B\ B\ B\ B}_{6}\ \underbrace{A\ A}_{7}\ \underbrace{B}_{8}\ \underbrace{A}_{9}\ \underbrace{B}_{10}\ \underbrace{A\ A}_{11}\ \underbrace{B\ B\ B\ B\ B}_{12}$ } 21 veículos e 12 subseqüências

b) $\underbrace{A\ A\ A\ A\ A}_{1}\ \underbrace{B}_{2}\ \underbrace{A\ A\ A}_{3}\ \underbrace{B\ B\ B\ B\ B\ B\ B\ B\ B\ B\ B\ B}_{4}$ } 21 veículos e 4 subseqüências

c) $\underbrace{A}_{1}\ \underbrace{B}_{2}\ \underbrace{A}_{3}\ \underbrace{B}_{4}\ \underbrace{A}_{5}\ \underbrace{B}_{6}\ \underbrace{A}_{7}\ \underbrace{B}_{8}\ \underbrace{A}_{9}\ \underbrace{B}_{10}\ \underbrace{A}_{11}\ \underbrace{B}_{12}\ \underbrace{A}_{13}\ \underbrace{B}_{14}\ \underbrace{A}_{15}\ \underbrace{B\ B\ B\ B\ B\ B}_{16}$ } 21 veículos e 16 subseqüências

d) $\underbrace{B}_{1}\ \underbrace{A}_{2}\ \underbrace{B}_{3}\ \underbrace{A}_{4}\ \underbrace{B}_{5}\ \underbrace{A}_{6}\ \underbrace{B}_{7}\ \underbrace{A}_{8}\ \underbrace{B}_{9}\ \underbrace{A}_{10}\ \underbrace{B}_{11}\ \underbrace{A}_{12}\ \underbrace{B}_{13}\ \underbrace{A}_{14}\ \underbrace{B}_{15}\ \underbrace{A}_{16}\ \underbrace{B\ B\ B\ B\ B}_{17}$ } 21 veículos e 17 subseqüências

$$\text{e) } \underbrace{\text{A A A A A A A A}}_{1}\ \underbrace{\text{B B B B B B B B B B B B B}}_{2} \left.\vphantom{\underbrace{A}_{1}}\right\} \begin{array}{c}\text{21 veículos}\\ \text{e}\\ \text{2 subseqüências}\end{array}$$

O exemplo (e) mostra que o *menor* número de subseqüências resulta da *concentração* de elementos iguais num só bloco; o exemplo (d) mostra que o *maior* número de subseqüências resulta da *alternância* de elementos a partir daquele elemento que figure *mais vezes* no total. Então:

$$\underbrace{\text{A A A A A A A A}}_{1}\ \underbrace{\text{B B B B B B B B B B B B B}}_{2}$$

ou

$$\underbrace{\text{B B B B B B B B B B B B B}}_{1}\ \underbrace{\text{A A A A A A A A}}_{2}$$

produzem 2 subseqüências, enquanto

B A B A B A B A B A B A B A B A B B B B B

produz 17 subseqüências.

Note-se que o número 17 poderia ter sido obtido sem dificuldade aplicando a seguinte fórmula:

$$S_{\text{máx}} = 2n_1 + 1 \qquad \text{(No caso de } n_1 = n_2, S_{\text{máx}} = 2n_1\text{)}$$

onde $S_{\text{máx}}$ = número máximo de subseqüências; n_1 = número de elementos figurantes em *minoria*. Então: número de elementos iguais a A = n_1 = 8, donde resulta que:

$$S_{\text{máx}} = 2\,(8) + 1 = 16 + 1 = 17 \text{ subseqüências}$$

Observação: Por n_2 designar-se-á o total de elementos *diferentes* de A. Neste caso, $n_2 = 13$ (isto é, 13 elementos iguais a B).

Designando por R o espaço experimental, isto é, o conjunto de todos os resultados possíveis, temos:

R : 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17 subseqüências

↑ Não pode haver, em *nenhum* caso, menos que 2 subseqüências. (2)

↑ Não pode haver, *neste* caso, mais que 17 subseqüências. (17)

Voltando ao problema inicial, a pergunta a que se pretende responder é: num conjunto máximo de 17 subseqüências, ocorreriam por mero acaso 8 subseqüências? A resposta depende de:

1. estabelecimento de uma H_o e de uma H_a;
2. fixação do nível de significância, α;
3. comparação do S_o (S obtido) com valores não críticos (S_{nc}) tabelados;
4. interpretação dos resultados.

1. *Hipótese Nula e Hipótese Alternativa*

H_o: Os A's e os B's ocorrem aleatoriamente.
H_a: Os A's e os B's não ocorrem aleatoriamente.

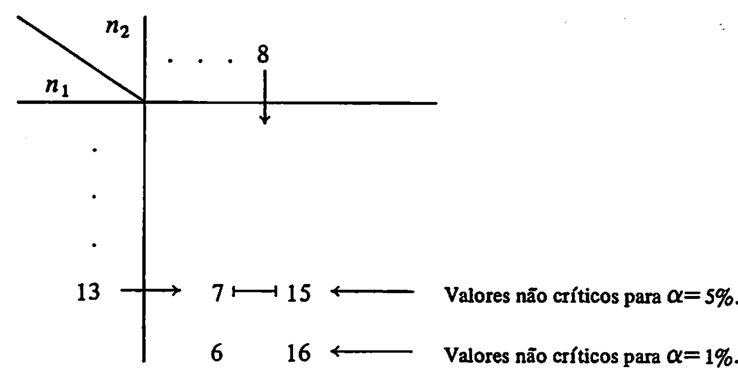
2. *Nível de Significância*

$\alpha = 5\%$.

3. *Comparações*

No caso em pauta, $n_1 = 8$ (8 elementos iguais a A) e $n_2 = 13$ (13 elementos iguais a B) produziram 8 subseqüências, isto é, $S_o = 8$.

Consultando a Tabela J no fim do livro, observamos que para $n_1 = 8$ e $n_2 = 13$ os S_{nc} (valores não críticos) tabelados são 7 e 15. Assim:



Observação: n_1 e n_2 são *intercambiáveis*, donde resulta que, não sendo possível consultar a tabela com $n_1 = 8$ e $n_2 = 13$, basta consultá-la com $n_1 = 13$ e $n_2 = 8$.

Então, feitas as comparações, verificamos que ($S_o = 8$) pertence ao intervalo 7 ⊢⊣ 15.

4. *Interpretação dos Resultados*

O intervalo tabelado abrange um conjunto de valores *não críticos*, isto é, qualquer valor situado *dentro* de seus limites – *inclusive* os extremos – pode ser considerado *aleatório*. Por outras palavras, qualquer S_o pertencente ao intervalo leva à *não rejeição da* H_o.

Obviamente, qualquer valor de S_o situado *fora* dos limites do intervalo (para a esquerda ou para a direita) leva à *rejeição da* H_o. No nosso caso, como $(S_o = 8) \in 7 \vdash\dashv 15$, H_o *não rejeitada*. Que significa isso? Ora, se a H_o estabelecia que os A's e os B's ocorriam aleatoriamente, a não rejeição dessa hipótese leva à aceitação da idéia de que provavelmente os carros observados (em função de suas velocidades) sucederam-se *ao acaso* na estrada. Notar que não estamos afirmando que a ordem de sucessão foi *de fato* casual; estamos, sim, afirmando que *provavelmente* essa ordem foi casual e a probabilidade de estarmos certos é de 95%, já que o nível de significância (α) adotado foi de 5%.

Se, no início do problema, α tivesse sido fixado em 1%, bastaria consultar a tabela do mesmo jeito, porém ler os números que estão abaixo do sinal $\vdash\dashv$. No caso, 6 e 16. Então:

$$(S_o = 8) \in 6 \vdash\dashv 16,$$

o que leva à mesma conclusão, isto é, H_o *não rejeitada* (e agora com 99% de certeza).

O leitor atento já terá observado que dado um conjunto de A's e outro de B's, poderão resultar *poucas* ou *excessivas* subseqüências, dependendo da ordem em que se apresentem os A's e os B's. A lógica subjacente ao teste consiste em rejeitar valores muito pequenos ou muito grandes e não rejeitar valores intermediários, segundo uma dada probabilidade, já que, por mero acaso, *não* deveriam ocorrer valores *extremos*.

EXERCÍCIOS RESOLVIDOS

1. Testar, com $\alpha = 5\%$, se na seqüência abaixo os A's e os B's ocorrem casualmente:

B A A B B B A B A A A A B A B A B B B

Solução

H_o: Os A's e os B's ocorrem aleatoriamente.
H_a: Os A's e os B's não ocorrem aleatoriamente.
$\alpha = 5\%$

$$\left.\begin{array}{l} n_1 = 9 \text{ (9 elementos iguais a A)} \\ n_2 = 10 \text{ (10 elementos iguais a B)} \end{array}\right\} \text{Por convenção, } n_1 \leqslant n_2$$

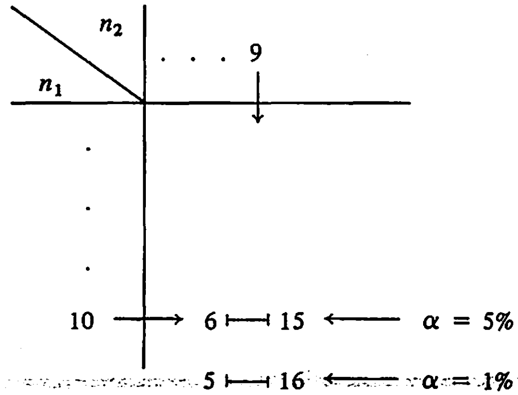
Aplicando $S_{máx} = 2n_1 + 1$, resulta que $S_{máx} = 2(9) + 1 = 19$. Então:

R: 2, 3, 4, . . ., 17, 18, 19 subseqüências *possíveis*.

De acordo com os dados do problema, $S_o = 11$. Assim:

B	A A	B B B	A	B	A A A A	B	A	B	A	B B B
1	2	3	4	5	6	7	8	9	10	11

O que então queremos saber é se 11 subseqüências, no conjunto de 19, formadas a partir de $(n_1 = 9)$ e $(n_2 = 10)$, podem ocorrer por *mero acaso*. Consultando a Tabela J, vem:



Como $(S_o = 11) \in 6 \vdash\dashv 15$, H_o *não rejeitada* e podemos concluir que provavelmente (com 95% de certeza) os A's e os B's aparecem numa ordem casual.

2. Um pescador afirma ter pescado 15 peixes na seguinte ordem:

S T S T S T S T S T T T S T S

onde S = sardinha e T = tainha. Com $\alpha_1 = 5\%$ e $\alpha_2 = 1\%$, testar se essa ordem foi casual.

Solução

H_o: Os S's e os T's ocorrem aleatoriamente.
H_a: Os S's e os T's não ocorrem aleatoriamente.
$\alpha_1 = 5\%$ e $\alpha_2 = 1\%$

S_o = 13, pois

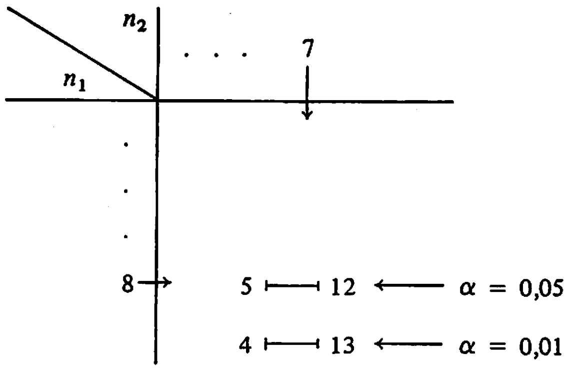
S T S T S T S T S TTT S T S

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

n_1 = 7

n_2 = 8

De acordo com a Tabela J, temos:



Como (S_o = 13) ∉ 5 ⊢ 12, H_o rejeitada. Porém, como (S_o = 13) ∈ 4 ⊢ 13, H_o não rejeitada. Então, com 95% de certeza (para α_1), podemos afirmar que a ordem de ocorrência dos peixes *não* foi casual; entretanto, com 99% de certeza (para α_2), podemos afirmar que a ordem de ocorrência dos peixes foi casual.

3. Uma moeda foi jogada 25 vezes e produziu a seguinte seqüência, onde C = cara e K = coroa:

C C C C K K C K C C C K K C K K C C C C C K K K K

Trabalhar com $\alpha = 5\%$ e testar se os C's e os K's ocorreram aleatoriamente.

Solução

H_o: Os C's e os K's ocorrem aleatoriamente.
H_a: Os C's e os K's não ocorrem aleatoriamente.
$\alpha = 5\%$
$S_o = 10$

Entrando com ($n_1 = 14$) e ($n_2 = 11$) na tabela, resulta o seguinte intervalo não crítico: 9 ⊢ 18. Como (S_o = 10) ∈ 9 ⊢ 18, H_o ***não rejeitada***, isto é, as caras e as coroas provavelmente ocorreram de modo *aleatório*.

Valores Não Tabelados

Observe o leitor que se $(n_1 \leqslant n_2) > 20$, a Tabela J não poderá mais ser usada. Isso, entretanto, não cria problemas, porque já foi demonstrado que se $(n_1 \leqslant n_2) > 10$, a Distribuição de Subseqüências é aproximadamente *normal*. (E o ajustamento será tanto melhor quanto maiores que 10 forem n_1 e n_2.)

Pode-se demonstrar que, arranjados aleatoriamente n_1 elementos de um tipo e n_2 elementos de outro tipo, o número médio de subseqüências é dado por

$$\mu = \frac{2P}{S} + 1\ , \text{ onde } \begin{cases} P = n_1 n_2 \\ S = n_1 + n_2 \end{cases}$$

Pode-se também demonstrar que, nas condições acima, a variância de uma Distribuição de Subseqüências é dada por

$$\sigma^2 = \frac{2P\,(2P - S)}{S^2\,(S-1)}$$

Assim, os extremos do intervalo não crítico podem ser obtidos calculando

$$\mu \pm z_c \cdot \sigma$$

onde z_c é o z crítico tirado de uma tabela de Normal.

A título de ilustração, vamos supor que no nosso problema inicial, o observador tivesse detectado 42 carros, assim distribuídos:

BBABABABBBAAAAABBBAABAAABBBBAAAAAAABBBBBB

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

n_1 = 20 (20 A's)
n_2 = 22 (22 B's)
S_o = 15
$\alpha_1 = 5\%$; $\alpha_2 = 1\%$ (Vamos examinar os dois casos para demonstrar onde ocorrem modificações na fórmula.)

Solução

$$\mu = \frac{2P}{S} + 1 \begin{cases} P = n_1 n_2 = (20)(22) = 440 \\ S = n_1 + n_2 = 20 + 22 = 42 \end{cases}$$

$$\mu = \frac{2\,(440)}{42} + 1 = \frac{880}{42} + 1 = 20{,}95 + 1 = 21{,}95$$

$$\mu = 21{,}95$$

$$\sigma^2 = \frac{2P\,(2P - S)}{S^2\,(S-1)}$$

$$\sigma^2 = \frac{880\,(880-42)}{42^2\,(42-1)} = \frac{880\,(838)}{1.764\,(41)} = \frac{737.440}{72.324} = 10{,}20$$

$$\sigma = \sqrt{10{,}20} \cong 3{,}19$$

$$\sigma = 3{,}19$$

Para $\alpha = 5\%$, $z_c = 1{,}96$ (ver Tabela B, pág. 357).

Então, entrando na fórmula que dá o intervalo não crítico, vem:

$$\mu \pm z_c \cdot \sigma = 21{,}95 \pm 1{,}96\,(3{,}19) = 21{,}95 \pm 6{,}25$$

∴ limite inferior do intervalo $= 21{,}95 - 6{,}25 = 15{,}70$
limite superior do intervalo $= 21{,}95 + 6{,}25 = 28{,}20$

Ora, $(S_o = 15) \notin 15{,}70 \longmapsto 28{,}20$, isto é, ficando o 15 *fora* do intervalo (do lado esquerdo), é possível concluir que os A's e os B's não ocorreram aleatoriamente, donde a *rejeição da* H_o (em favor da H_a).

Para o caso de $\alpha_2 = 1\%$, o $z_c = 2{,}58$ e aí o intervalo não crítico passa a ser:

$$\mu \pm z_c\sigma = 21{,}95 \pm 2{,}58\,(3{,}19) = 21{,}95 \pm 8{,}23$$

∴ limite inferior do intervalo $= 21{,}95 - 8{,}23 = 13{,}72$
limite superior do intervalo $= 21{,}95 + 8{,}23 = 30{,}18$

Neste caso, como $(S_o = 15) \in 13{,}72 \longmapsto 30{,}18$, a hipótese de ocorrência aleatória dos A's e dos B's *não deve ser rejeitada.*

EXERCÍCIOS RESOLVIDOS

1. Vamos resolver o Exercício 3 da pág. 252 por recorrência à Distribuição Normal.

$$\mu = \frac{2P}{S} + 1 \quad \begin{cases} 2P = 2\,(n_1 n_2) = 2\,(11)\,(14) = 308 \\ S = n_1 + n_2 = 11 + 14 = 25 \end{cases}$$

$$\mu = \frac{308}{25} + 1 = 12{,}32 + 1 = 13{,}32$$

$$\sigma^2 = \frac{2P\,(2P-S)}{S^2\,(S-1)} = \frac{308\,(308-25)}{25^2\,(25-1)} = \frac{(308)\,(283)}{(625)\,(24)} = \frac{87.164}{15.000} = 5{,}81$$

$$\sigma = \sqrt{5{,}81} \cong 2{,}41$$

Como $\alpha = 5\%$, $z_c = 1{,}96$, vem:

$$\mu \pm z_c \cdot \sigma = 13{,}32 \pm 1{,}96\,(2{,}41) = 13{,}32 \pm 4{,}72$$

∴ limite inferior do intervalo $= 13{,}32 - 4{,}72 = 8{,}60$
limite superior do intervalo $= 13{,}32 + 4{,}72 = 18{,}04$

Então, $(S_o = 10) \in 8{,}60 \longmapsto 18{,}04$ e a H_o é *não rejeitada.*

Observação: Os valores 8,60 e 18,04 arredondados para o inteiro mais próximo dão 9 e 18. Notar que esses valores (9 e 18) correspondem precisamente aos limites não críticos encontrados na Tabela J.

2. Com $\alpha = 4\%$ (H_a bicaudal), testar a hipótese de que os A's e os B's ocorreram casualmente na seqüência

BABAABABABABABABBABABBBABABABABABABBABABABBBABABAB

43

Solução

$n_1 = 22$
$n_2 = 28$
$S_o = 43$

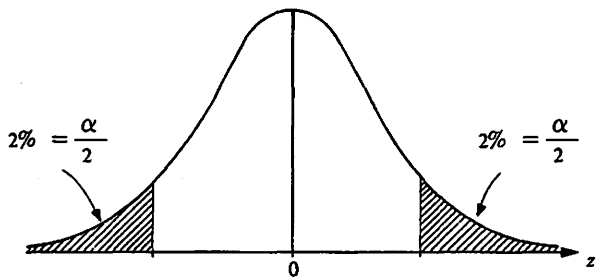
$$\mu = \frac{2P}{S} + 1 = \frac{2\,(22)\,(28)}{22+28} + 1 = \frac{1.232}{50} + 1 = 24{,}64 + 1 = 25{,}64$$

$$\sigma^2 = \frac{2P(2P-S)}{S^2(S-1)} = \frac{1.232\,(1.232-50)}{50^2\,(50-1)} = \frac{1.232\,(1.182)}{2.500\,(49)} =$$

$$= \frac{1.456.224}{122.500} = 11{,}89$$

$$\sigma = \sqrt{11{,}89} \cong 3{,}45$$

Para o cálculo do z_c precisamos atribuir a cada cauda uma área que corresponda a $\frac{\alpha}{2} = 2\%$. Assim:



Ora, se metade da área abrangida por uma curva normal vale, em termos probabilísticos, 0,50, a área que resulta da subtração de $\frac{\alpha}{2} = 2\%$ é

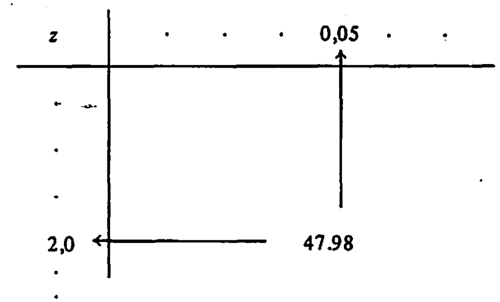
$$0{,}50 - 0{,}02 = 0{,}48$$

Na Tabela B, pág. 357, procura-se o equivalente porcentual a 0,48 que é 48,00%. Lamentavelmente não se encontrará esse valor, mas algo muito próximo que é 47,98% (aprox. por *falta*). Poder-se-ia também usar 48,03% (aprox. por *excesso*). Vamos preferir o primeiro valor, já que a diferença é *menor*. Assim:

$$\left.\begin{array}{l} 48{,}00 - 47{,}98 = 0{,}02 \\ 48{,}03 - 48{,}00 = 0{,}03 \end{array}\right\} 0{,}02 < 0{,}03$$

(Quanto *menor* a diferença, mais *próximo* está o valor encontrado do valor desejado.)

Então, é ler na moldura da Tabela B (pág. 357) os dados que compõem o z_c. Assim:



Portanto, para $\alpha = 4\%$ (H_a bicaudal), $z_c = 2{,}05$.

Agora é entrar na fórmula $\mu \pm z_c\sigma$:

$$25{,}64 \pm 2{,}05\,(3{,}45) = 25{,}64 \pm 7{,}07$$

$\therefore$ limite inferior do intervalo $= 25{,}64 - 7{,}07 = 18{,}57$
limite superior do intervalo $= 25{,}64 + 7{,}07 = 32{,}71$

Como $(S_o = 43) \notin 18{,}57 \longmapsto 32{,}71$, H_o *rejeitada* (com 96% de certeza), isto é, os A's e os B's *não* se distribuem aleatoriamente.

3. Na Tabela H (pág. 363), vamos escolher uma fileira qualquer, por exemplo, a quarta. Nessa fileira, lêem-se os seguintes números:

6 5 6 3 1 6 8 6 7 2 0 7 2 3 2 1 5 0 9

Ora, sendo a Tabela H um conjunto de números *aleatórios*, gostaríamos de testar se a ordem em que eles ocorrem é, de fato, *casual*. Vamos trabalhar com $\alpha = 5\%$.

Solução

H_o: Os números dessa seqüência ocorrem em ordem casual.
H_a: Os números dessa seqüência não ocorrem em ordem casual.
$\alpha = 5\%$

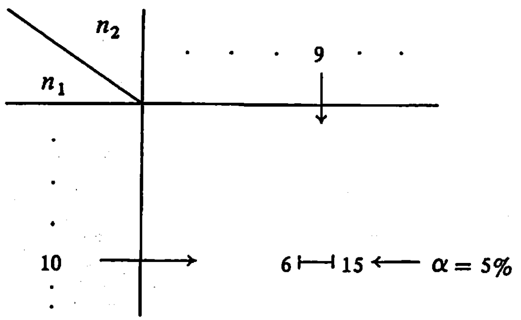
Para que possamos "enxergar" as subseqüências, vamos distribuir P aos números *pares* e I aos *ímpares*. Desse artifício resulta o seguinte:

$$\underline{P}\ \underline{I}\ \underline{P}\ \underline{I\ I}\ \underline{P\ P\ P}\ \underline{I}\ \underline{P\ P}\ \underline{I}\ \underline{P}\ \underline{I}\ \underline{P}\ \underline{I\ I}\ \underline{P}\ \underline{I}$$

14

Então $S_o = 14$
$n_1 = 9$ (I's)
$n_2 = 10$ (P's).

Localizando n_1 e n_2 na Tabela J, pág. 366, vem:



Como ($S_o = 14$) $\in 6 \vdash\!\!\dashv 15$, H_o *não rejeitada*, isto é, com 95% de certeza, os dígitos ocorrem, nessa linha, casualmente.

Alternativamente, poderíamos tentar resolver o problema por recorrência à Normal (embora n_1 sendo *menor* que 10 não ofereça aproximações satisfatórias!).

$$\mu = \frac{2P}{S} + 1 = \frac{2(9)(10)}{9+10} + 1 = \frac{180}{19} + 1 = 9{,}47 + 1 = 10{,}47$$

$$\sigma^2 = \frac{2P(2P-S)}{S^2(S-1)} = \frac{180(180-19)}{19^2(19-1)} = \frac{180(161)}{361(18)} = \frac{28.980}{6.498} = 4{,}46$$

$$\sigma = \sqrt{4{,}46} \cong 2{,}11$$

Então $\mu \pm z_c \cdot \sigma = 10{,}47 \pm 1{,}96\,(2{,}11) = 10{,}47 \pm 4{,}14$.

$\therefore$ limite inferior do intervalo $= 10{,}47 - 4{,}14 = 6{,}33$
limite superior do intervalo $= 10{,}47 + 4{,}14 = 14{,}61$

Como ($S_o = 14$) $\in 6{,}33 \vdash\!\!\dashv 14{,}61$, H_o *não rejeitada*, isto é, chegamos à *mesma* solução.

PROBLEMAS SUPLEMENTARES

1. *Observaram-se 12 meninos de 4 anos e 12 meninas também de 4 anos durante duas sessões lúdicas de 15 minutos cada uma, classificando-se o modo de brincar de cada criança, nos dois períodos, conforme a incidência e o grau de agressividade. Com tais resultados, pretende-se testar a hipótese de que há diferença no grau de agressividade, conforme o sexo.
Os dados obtidos estão na tabela abaixo:

Meninos	*Meninas*
86	55
69	40
72	22
65	58
113	16
65	7
118	9
45	16
141	26
104	36
41	20
50	15

Trabalhar com $\alpha = 5\%$.

Solução

Atribuindo a cada escore masculino a letra H (= homem) e a cada escore feminino a letra M (= mulher), resulta a seguinte nova tabela:

Meninos	*Meninas*
86 H	55 M
69 H	40 M
72 H	22 M
.	.
.	.
.	.
41 H	20 M
50 H	15 M

* Problema extraído de Siegel, Sidney, *Estatística Não-Paramétrica*. São Paulo: McGraw-Hill do Brasil, 1975, págs. 158 e 159. (Tradução de Alfredo Alves de Farias.)

Dispondo esses escores (todos, isto é, dos meninos e das meninas) em ordem crescente (ou decrescente – tanto faz), vem:

7, 9, 15, 16, 16, ..., 104, 113, 118, 141,

donde as seguintes subseqüências:

$$\underbrace{\text{M M M M M M M M M M}}_{1}\ \underbrace{\text{H H H}}_{2}\ \underbrace{\text{M M}}_{3}\ \underbrace{\text{H H H H H H H H H}}_{4}$$

H_o: Os M's e os H's distribuem-se aleatoriamente.
H_a: Os M's e os H's não se distribuem aleatoriamente.

Então: $(n_1 = 12); (n_2 = 12); (S_o = 4); (S_{\text{mín}} = 2); (S_{\text{máx}} = 25)$.

E o que queremos testar é se nesse contexto a ocorrência de 4 subseqüências é algo que pode ser considerado casual.

$$\mu = \frac{2P}{S} + 1 \longrightarrow \mu = \frac{2\,(12)\,(12)}{12+12} + 1 = \frac{288}{24} + 1 = 12 + 1 = 13$$

$$\sigma^2 = \frac{2P\,(2P-S)}{S^2\,(S-1)} \longrightarrow \sigma^2 = \frac{288\,(288-24)}{(24)^2\,(24-1)} = \frac{288\,(264)}{576\,(23)} =$$

$$= \frac{76.032}{13.248} \cong 5{,}74$$

$$\sigma^2 = +\sqrt{5{,}74} \cong 2{,}4$$

Intervalo de confiança com 95% de certeza:

$$\mu \pm z_c\,(\sigma) = 13 \pm 1{,}96\,(2{,}4) = 13 \pm 4{,}7$$

∴ limite inferior do intervalo = 13 – 4,7 = 8,3
limite superior do intervalo = 13 + 4,7 = 17,7

Como $(S_o = 4) \notin 8{,}3 \vdash\!\dashv 17{,}7$, H_o *rejeitada*, isto é, os meninos e as meninas, relativamente ao modo de brincar (nas condições do experimento), *não* se distribuem casualmente. Esses meninos e essas meninas podem, portanto, ser considerados como pertencentes a *populações distintas*. E dizer que pertencem a populações distintas equivale a dizer que esses grupos *diferem* quanto à variável observada.

2. Certo jogador desconfia de que as cartas de determinado baralho não foram bem misturadas. Por isso, ao "cortar", retém um conjunto de 15 delas e, ao examiná-las, verifica que, na ordem em que se sucedem, são:

3C – 2ϕ – 7P – 7E – 8ϕ – 9P – 5C – 3E – 2C – 6P – 4ϕ – 6E – 4C – 4E – 5ϕ,

onde (C = copas), (ϕ = ouros), (P = paus) e (E = espadas).

Com $\alpha = 5\%$, será possível sustentar a hipótese de que houve burla?

Solução

Comecemos recordando que se as cartas estiverem bem misturadas não deverá ser possível prever qualquer delas a partir da anterior.

Ora, sabemos que os naipes *copas* e *ouros* são *vermelhos* (V) e os naipes *paus* e *espadas* são *pretos* (P). Assim, atribuindo P's ou V's às cartas acima, sem mudar-lhes a ordem, resulta o seguinte:

$$\underline{\text{V}}\ \underline{\text{V}}\ \underline{\text{P P}}\ \underline{\text{V}}\ \underline{\text{P}}\ \underline{\text{V}}\ \underline{\text{P}}\ \underline{\text{V}}\ \underline{\text{P}}\ \underline{\text{V}}\ \underline{\text{P}}\ \underline{\text{V}}\ \underline{\text{P}}\ \underline{\text{V}}$$

13

onde n_1 = 7 (elementos do tipo P)
n_2 = 8 (elementos do tipo V)
S_o = 13.

Então:

H_o: Os V's e os P's distribuem-se aleatoriamente (no caso de as cartas estarem bem misturadas).
H_a: Os V's e os P's não se distribuem aleatoriamente.

Consultando a Tabela J no fim do livro, vem:

n_2 \ n_1	. . . 7 . . .
.	
.	
.	
8	$5 \vdash\!\dashv 12$ ⟵ $\alpha = 5\%$
.	

Conclusão: $(S_o = 13) \notin 5 \vdash\!\dashv 12$, donde H_o ser *rejeitada*. Em outras palavras, rejeita-se a hipótese de que as cartas tenham sido bem misturadas. Deve ter havido, por conseguinte, tentativa de burla.

† EXPERIMENTO BINOMIAL
(Prova Binomial)

Um experimento diz-se *binomial* quando apresenta, simultaneamente, as seguintes características:

1. consiste de n tentativas (lances, extrações etc.) idênticas;
2. os resultados são mutuamente excludentes, ou seja, ocorrendo o resultado a não ocorrerá o resultado b;
3. os eventos são independentes: a ocorrência de um não influi na ocorrência do outro;
4. o experimento implica reposição, o que garante que P (sucesso) $= p$ seja constante.

Seja agora uma moeda M cuja eqüiprobabilidade ("honestidade") queiramos testar. As quatro primeiras coisas que devemos fazer são:

1. escrever H_o e H_a em termos probabilísticos;
2. decidir qual a dimensão do erro que estamos dispostos a tolerar (α);
3. arranjar uma situação (experimento) que permita testar as hipóteses;
4. definir operacionalmente X, isto é, a variável de observação.

Então:

1. H_o: $P(C) = P(K) = 0{,}5 \longleftarrow$ Eqüivale a dizer que a moeda é "honesta".

$$H_a:\ P(C) \neq P(K), \text{ donde } \begin{cases} P(C) \neq P(K) \longrightarrow H_a \text{ bicaudal} \\ P(C) > P(K) \longrightarrow H_a \text{ unicaudal} \\ P(C) < P(K) \longrightarrow H_a \text{ unicaudal} \end{cases}$$

A escolha depende do que suspeitemos.

Vamos supor que tenhamos boas razões para suspeitar que $P(C) > 0{,}5$. Então

$$H_a:\ P(C) > 0{,}5 \longrightarrow \text{unicaudal (direita)}$$

2. $\alpha = 0{,}05$, isto é, queremos obter resposta com 95% de certeza, no mínimo.

3. Experimento: jogar a moeda certo número de vezes e anotar o número de ocorrências de determinado tipo (conforme a definição de X).
Por exemplo: jogar M 8 vezes e controlar as ocorrências do tipo "cara" (C).

4. Definição operacional de X: número de vezes que sair "cara" (C).

Logo de saída verificamos que o nosso espaço experimental, R, possui 9 pontos, que são:

R: 0, 1, 2, 3, 4, 5, 6, 7 e 8 ocorrências de C

Esse espaço experimental (R) deverá agora ser repartido em dois pedaços. Assim:

R:	R_c^*	R_c
	Região não-crítica, isto é, região de *não rejeição* da H_o.	Região crítica, isto é, região de *rejeição* da H_o.

A primeira parte do problema consiste em determinar entre que dois valores de R passa o segmento que separa R_c^* de R_c. Para tanto, vamos calcular as probabilidades associadas à *cauda direita* de R. Por quê? Porque se suspeitamos que a $P(C) > 0{,}5$ deveremos obter um número "grande" de "caras". (E essa suspeita deve estar presente quando se redige a H_a.)

Recorrendo ao binômio gerador das probabilidades que convêm ao nosso problema, temos:

$$[P(F) + P(S)]^n = (q + p)^n \longrightarrow (q + p)^8$$

Desenvolvendo $(q + p)^8$, vem:

$$(q+p)^8 = \underset{P(X=0)}{q^8} + \underset{P(X=1)}{8q^7p} + \underset{P(X=2)}{28q^6p^2} + \underset{P(X=3)}{56q^5p^3} + \underset{P(X=4)}{70q^4p^4} +$$

$$+ \underset{P(X=5)}{56q^3p^5} + \underset{P(X=6)}{28q^2p^6} + \underset{P(X=7)}{8qp^7} + \underset{P(X=8)}{p^8}$$

E agora é examinar as probabilidades* associadas a esses valores extremos de R. Assim:

$$P(X = 8) = (0{,}5)^8 \cong 0{,}004 < 0{,}05. \text{ Logo, 8 } \textit{serve.}$$
$$P(X = 7) = 8(0{,}5)(0{,}5)^7 \cong 0{,}031 < 0{,}05$$

* Essas probabilidades já estão calculadas. É só consultar a Tabela K (pág. 369).

Além disso, $[P(X=8) \text{ ou } P(X=7)] = 0{,}004 + 0{,}031 = 0{,}035 < 0{,}05$.
Logo, 7 *serve*.

$P(X=6) = 28\,(0{,}5)^2\,(0{,}5)^6 \cong 0{,}109 > 0{,}05$. Logo, 6 *não serve*.

Pronto! R já está dividido:

$$R:\ \underbrace{0,\ 1,\ 2,\ 3,\ 4,\ 5,\ 6,}_{R_c^*}\Big|\underbrace{7,\ 8}_{R_c}$$

Vamos agora supor *realizado* o experimento e que $X = 6$ caras.

Ora, dispomos de duas regiões críticas para resolver o problema: R_{c_1}: 8 e R_{c_2}: 7 e 8. Como $(X = 6)$, podemos usar R_{c_1} ou R_{c_2} para concluir que

$(X = 6) \notin R_c$, donde H_0 *não rejeitada*.

Assim, embora o nosso α seja 5%, o erro *máximo* que estamos cometendo, adotando a R_{c_2}, será de 3,5%, donde a resposta: M é eqüiprovável (honesta) com 96,5% de certeza.

A segunda parte do problema consiste em responder à seguinte pergunta: "E se $P(C) > 0{,}5$?" Notar que se $P(C) > 0{,}5$, a H_0 será *falsa*, embora, *pelo teste*, ela não tenha sido rejeitada. E isso é perfeitamente possível, já que α não é zero (nem poderia jamais ser! – a menos que abríssemos mão de "fazer ciência" para "fazer milagre"...!!).

Então é preciso controlar a *sensibilidade* do teste (*poder*), a fim de garantir que não se cometerá um *erro tipo II* – que consiste em não rejeitar a H_0 quando, de fato, ela deveria ter sido rejeitada.

$$\text{Se } P(C) > 0{,}5, \text{ então } P(C) = \begin{cases} 0{,}6 \\ 0{,}7 \\ 0{,}8 \\ 0{,}9 \\ \text{etc.} \end{cases}$$

(Só usamos esses valores para facilitar a consulta à Tabela K, pág. 369.)

Vejamos então o que ocorre para cada uma dessas probabilidades:

$P(X = i)$	$B(8; 0{,}6)$ ↓ $P(C) = 0{,}6$	$B(8; 0{,}7)$ ↓ $P(C) = 0{,}7$	$B(8; 0{,}8)$ ↓ $P(C) = 0{,}8$	$B(8; 0{,}9)$ ↓ $P(C) = 0{,}9$
$i = 8$	0,017	0,058	0,168	0,430
$i = 7$	0,090	0,198	0,336	0,383
$i = 7; 8$	0,107	0,256	0,504	0,813
Em %	10,7%	25,6%	50,4%	81,3%

Essas probabilidades representam os *vários poderes* associados a conjecturas específicas. Por exemplo:

1. se considerarmos R_{c_1}: 8 e $P(C) = 0{,}6$, a probabilidade de rejeitarmos a H_0 se ela for, de fato, falsa será igual a 0,017 – 1,7% – valor ridiculamente baixo, pois estaremos "comendo gato por lebre" em 98,3% das situações;

2. se considerarmos R_{c_2}: 7 e 8 e $P(C) = 0{,}7$, a probabilidade de rejeitarmos a H_0 se ela for, de fato, falsa será 0,256 – 25,6% – valor ainda muito baixo, pois estaremos *decidindo mal* em $(100 - 25{,}6) = 74{,}4\%$ dos casos.

Conclusão: o experimento é *ótimo* para perceber a *veracidade* da H_0 se, de fato, ela é *verdadeira*. E com a vantagem de o α efetivo ser 0,035 – menor que 0,05 – se a R_c usada compreender os valores 7 e 8. Mas o experimento é *ruim* para perceber a *falsidade* da H_0 se, de fato, ela é *falsa* – a menos que $p \geqslant 0{,}8$. (Pois, a partir de $p \geqslant 0{,}8$, o *poder* começa a assumir valores aceitáveis: 50,4%, 81,3%.)

Para que o poder fosse maior, teríamos de melhorar o modelo experimental. Há várias maneiras de fazê-lo; uma delas é aumentar o tamanho da amostra (n).

RESUMO

Os estatísticos desenvolveram certo número de testes não-paramétricos de significância–testes cujos requisitos não incluem distribuição normal da variável de observação nem nível intervalar de mensuração. O qui-quadrado (χ^2), o mais conhecido dos testes não-paramétricos, é usado em comparações entre freqüências. Quando as diferenças entre as freqüências esperadas e as observadas são muito grandes, rejeitamos a hipótese nula e aceitamos a hipótese experimental (alternativa), que acena com uma verdadeira diferença na população. Esse é o requisito básico para que um qui-quadrado observado seja significante. Existem ainda outros procedimentos não-paramétricos: o teste da mediana determina se há uma diferença significativa entre as medianas de duas amostras; a dupla análise da variância de Friedman permite fazer comparações entre mensurações distintas da mesma amostra (duas mensurações, no mínimo); a análise da variância por postos de Kruskal-Wallis dá margem a uma comparação entre várias amostras independentes.

† A prova de subseqüências permite analisar se, dado um conjunto de A's e outro de B's, a ordem em que ocorrem esses elementos pode ser considerada casual. A prova binomial, aplicável a algumas variáveis que resultem de contagem, é uma aplicação interessante do binônio de Newton à teoria das probabilidades. Finalmente, a prova de Mann-Whitney é a alternativa não-paramétrica à prova t, da mesma forma que a prova de Kruskal-Wallis, não-paramétrica, é a alternativa à análise de variância, paramétrica, estudada no Capítulo 9.

PROBLEMAS

1. Amostras aleatórias de homens e de mulheres foram entrevistadas com o objetivo de pesquisar-se o comportamento de "fumar cigarros". Descobriu-se que, de 29 homens, 15 eram fumantes; igualmente que, de 30 mulheres, 20 tinham o hábito de fumar. Teste a hipótese nula de que a freqüência relativa de homens que são fumantes é igual à de mulheres que também o são. Que indicam seus resultados? ($\alpha = 5\%$)

2. Dois grupos de estudantes fizeram exames finais de Estatística. Somente um grupo recebeu preparação formal para o exame; o outro grupo leu o texto recomendado, mas nunca compareceu às aulas. Enquanto que 22 dos 30 membros do primeiro grupo (os "freqüentadores") passaram no exame, apenas 10 dos 28 do segundo grupo (os "ausentes") lograram aprovação. Teste a hipótese nula de que a freqüência relativa dos "freqüentadores" que passaram nas provas finais é igual à freqüência relativa dos "ausentes" que também passaram. Que é que seus resultados indicam? ($\alpha = 5\%$)

3. Aplicando a correção de Yates, realize uma prova de qui-quadrado para o seguinte problema 2 X 2:*

M / *S*	*A*	*B*
♂	16	8
♀	7	11

S = sexo do fumante
M = marca do cigarro
($\alpha = 5\%$)

4. Aplicando a correção de Yates, realize uma prova de qui-quadrado para o seguinte problema 2 X 2:

Curados / *Vacinados*	*Sim*	*Não*
Sim	8	12
Não	10	5

($\alpha = 5\%$)

5. Aplicado a correção de Yates, realize uma prova de qui-quadrado para o seguinte problema 2 X 2:

* N.T.: É importante salientar duas coisas: (1)–não se faz ***correção de continuidade*** (correção de Yates) em tabelas com ***dois ou mais*** graus de liberdade; (2)–mesmo em tabelas com ***um*** grau de liberdade, só tem sentido aplicar tal correção se a hipótese nula tiver sido ***rejeitada***.

Cabelos / *Olhos*	*Esc.*	*Clar.*
Escuros	20	14
Claros	5	10

($\alpha = 5\%$)

6. Faça uma prova de qui-quadrado para o seguinte problema 3 X 3:

Candidato / *Região*	*A*	*B*	*C*
Sul	20	17	5
Centro	15	16	16
Norte	4	14	18

($\alpha = 5\%$)

7. Faça um teste de qui-quadrado para o seguinte problema 4 X 2:

Juiz(a) / *Cand.*	♂	♀
Maria	25	6
Lúcia	19	10
Adriana	15	15
Patrícia	8	20

($\alpha = 5\%$)

8. Faça uma prova de qui-quadrado para o seguinte problema 2 X 3:

Margarinas / *Consumid.*	*A*	*B*	*C*
Jovens	8	10	15
Velhos	12	10	9

($\alpha = 5\%$)

9. Pediu-se a duas amostras (aleatórias) de alunos que lessem e em seguida avaliassem um conto escrito por um novo autor. À metade dos estudantes (isto é, à primeira amostra) foi dito que o texto era de autoria de uma mulher, enquanto que à outra metade (segunda amostra), que a autoria era de um homem. Resultaram as seguintes avaliações, onde escores mais altos indicam atitudes mais favoráveis:

X_1 *(Foi-lhes Dito que o Texto era de Autoria de uma Mulher)*	X_2 *(Foi-lhes Dito que o Texto era de Autoria de um Homem)*
6	6
5	8
1	8
1	2
3	5
4	6
3	3
6	8
5	6
5	8
1	2
3	2
5	6
6	8
6	4
3	3

($\alpha = 5\%$)

Aplicando a prova da mediana, decida se existe uma diferença significativa entre as medianas desses grupos. Será que as avaliações feitas pelos estudantes sofreram influência da variável "sexo do autor"?

10. Aplicando o teste da mediana, determine se há uma diferença significativa entre as medianas das seguintes amostras de escores:

X_1		X_2	
7		4	
8	9	7	3
7	5	3	2
6	9	2	2.
7	8	3	6
7	9	4	4
8	7	7	5
9	9	4	4
7	9	5	4
6		6	4
9		2	3

($\alpha = 5\%$)

11. Uma amostra de 14 crianças, constituinte do grupo "harmonia e identificação", foi mensurada em dois momentos distintos: *antes* e *depois* de sua participação numa tarefa de classe cujo objetivo era tornar os alunos mais dependentes uns dos outros na consecução de uma promoção curricular.
Resultaram os seguintes escores-identificação (escores mais altos indicam maior "harmonia grupal"):

	Tarefa Cooperativa	
Aluno	*Antes Tempo 1*	*Depois Tempo 2*
A	62	75
B	51	53
C	60	62
D	43	51
E	49	52
F	45	46
G	73	62
H	66	68
I	57	55
J	63	69
K	43	45
L	46	45
M	67	68
N	61	67

Aplicando a dupla análise de variância de Friedman (por postos), determine se há uma diferença significativa entre os resultados obtidos no Tempo 1 e Tempo 2 relativamente à variável "harmonia grupal". ($\alpha = 5\%$)

12. Aplicando a dupla análise de variância por postos—χ^2 de Friedman—verifique se há uma diferença significativa entre os escores produzidos por uma amostra de 11 respondentes em três momentos distintos: T_1, T_2 e T_3. ($\alpha = 5\%$)

Respondente	T_1	T_2	T_3
A	60	62	64
B	53	54	50
C	59	65	71
D	65	66	68
E	55	63	61
F	71	74	76
G	57	58	63
H	77	76	79
I	63	65	70
J	54	59	62
K	63	62	65

13. Pesquisadores analisaram a variável "alienação política" entre três amostras de universitários: bacharelandos em "profissões liberais", em "engenharia" e em "artes". Cada amostra forneceu um grupo de escores, conforme abaixo indicado (quanto mais alto o escore maior a alienação):

X_1 *(Prof. Liberais)*	X_2 *(Engenharia)*	X_3 *(Artes)*
100	101	97
110	90	98
95	92	99
93	100	100
106	90	104
102	96	103
	92	

Aplicando a análise da variância de Kruskal-Wallis, verifique se há uma diferença significativa entre as citadas amostras de universitários com respeito ao nível de alienação política. ($\alpha = 5\%$)

14. Aplicando a prova de Kruskal-Wallis, verifique se há uma diferença significativa entre as seguintes amostras de dados:

X_1	X_2	X_3
125	100	95
100	99	90
122	105	86
127	103	96
115	116	88
129	98	89
130		

($\alpha = 5\%$)

† EXERCÍCIOS PROPOSTOS

1. Testar com $\alpha = 5\%$, se a distribuição abaixo pode ser considerada normal com média aritmética = 39,0 e variância = 76,56.

X	f
20 ⊢— 25	5
25 ⊢— 30	10
30 ⊢— 35	12
35 ⊢— 40	15
40 ⊢— 45	28
45 ⊢— 50	10
50 ⊢— 55	5
55 ⊢— 60	5

2. Testar, com $\alpha = 5\%$, se a distribuição abaixo pode ser considerada normal com os parâmetros:

74,0 = média aritmética e
59,85 = variância.

X	f
50 ⊢— 55	2
55 ⊢— 60	3
60 ⊢— 65	5
65 ⊢— 70	15
70 ⊢— 75	30
75 ⊢— 80	25
80 ⊢— 85	15
85 ⊢— 90	10

3. Para testar a eqüiprobabilidade de uma moeda, um pesquisador lançou-a 160 vezes, obtendo os seguintes resultados: 132 caras e 28 coroas. Qual a conclusão relativa à H_0 com $\alpha = 5\%$?

4. Para testar a honestidade de um dado, um pesquisador lançou-o 200 vezes e obteve os seguintes resultados:

Face	*Ocorrências*
1	20
2	18
3	40
4	35
5	60
6	27
	200

Com $\alpha = 1\%$, qual a conclusão relativa à H_0?
E se $\alpha = 5\%$, altera-se a conclusão?

5. Com $\alpha = 5\%$, será possível afirmar que existe relação de dependência entre as marcas A e B, de carros, e o sexo do(a) proprietário(a)?

Sexo \ *Marca*	*A*	*B*	
M	3	8	11
F	5	2	7
	8	10	18

6. O departamento do pessoal de determinada empresa, para averiguar se existe relação entre sexo e tendência a chegar tarde ao trabalho, fez a seguinte coleta de dados em dia aleatoriamente escolhido:

Sexo \ *Tendência*	*Chegam Cedo*	*Chegam Tarde*	
♂	9	3	12
♀	7	5	12
	16	8	24

Qual sua conclusão relativamente à H_o?
Trabalhe com $\alpha = 5\%$.

7. Em certo hospital, 20 crianças epilépticas foram submetidas a exames eletroencefalográficos. O objetivo da pesquisa era determinar se havia relação de dependência entre região cerebral (anterior/posterior) do foco epiléptico e hemisfério cerebral (direito/esquerdo). Eis os dados:

Região \ *Hemisfério*	*Direito*	*Esquerdo*	
Anterior	3	8	
Posterior	7	2	
			20

Considerar a casela de freqüência 2 para o início dos cálculos.

Com $\alpha = 5\%$, qual sua conclusão?

a) H_o rejeitada.
b) H_o não rejeitada.
c) H_o não pode ser testada por falta de dados.

8. Mesmos dados do problema 6. Apesar de α ter sido fixado em 5%, qual o nível *efetivo* desse experimento?

9. Mesmos dados do problema 3. De quantos pontos consta o espaço experimental (R)?

10 Suponha que $(n_1 = 9)$, $(n_2 = 21)$, $(P_1 = 176{,}5)$, $(P_2 = 288{,}5)$.
a) Qual o valor de U_o?
b) Qual o valor de U'_o?
c) Se a H_a for unicaudal, qual sua conclusão relativamente à H_o, com $\alpha = 5\%$?
d) Se a H_a for bicaudal, qual sua conclusão relativamente à H_o, com $\alpha = 10\%$?

11. Em determinada escola, 7 crianças foram alfabetizadas pelo método A e 11 pelo método B. Ao final do ano, as provas de leitura produziram as seguintes notas:

Método A	*Método B*
80	60
35	65
70	50
70	85
85	93
90	70
92	61
	40
	87
	75
	45

Pode-se afirmar, com $\alpha = 5\%$, que o método A é superior ao método B?

12. Uma moeda, jogada 18 vezes, produziu os seguintes resultados (nesta ordem): C C K K K C C K K K C C K K K C C K. Com $\alpha = 1\%$, será possível afirmar que a ordem em que os C's e os K's apareceram foi casual?

13. Testar, com $\alpha = 5\%$, se os A's e os B's que compõem a seqüência abaixo ocorrem casualmente:

A B B B A B A A B A A A A B A A B B A B A B A A A A

14. Numa fila de cinema, as pessoas distribuem-se, quanto ao sexo, na seguinte ordem: ♂ ♂ ♀ ♂ ♀ ♀ ♀ ♀ ♀ ♂ ♀ ♀ ♀ ♀ ♀ ♀ ♂ ♀ ♀ ♀. Será possível afirmar que a ordem de chegada dessas pessoas foi aleatória? (Trabalhr com $\alpha = 0{,}05$.)

15. Tome-se a coluna nº 13 da Tabela H, pág. 363. Será possível afirmar que os números, na ordem em que ali aparecem, são casuais? Trabalhar com $\alpha = 5\%$.

16. Sena $n_1 = 12$ e $n_2 = 18$. Quais os limites do intervalo não crítico (R_c^*) para α=6% numa prova de subseqüências? (Considere a H_a bicaudal.) Dê a resposta em números inteiros.

17. Participaram de um concurso de elegância infantil 38 crianças de 10 anos. O número de meninos era igual ao de meninas. Essas crianças dispunham de um farto e variado guarda-roupa, devendo elas próprias fazer as escolhas. As notas que elas receberam estão indicadas a seguir:

Notas dos meninos:

10, 26, 35, 99, 12, 16, 94, 61, 58, 65, 43, 38, 51, 70, 92, 77, 84, 89, 90

Notas das meninas:

18, 27, 36, 100, 79, 54, 52, 57, 41, 44, 50, 67, 63, 72, 88, 96, 85, 91, 86

Será possível afirmar, com $\alpha = 5\%$, que os critérios de escolha diferem conforme o sexo?

18. Para testar se uma pessoa era capaz de captar transmissões telepáticas de outra pessoa, idealizou-se o seguinte experimento: (a) uma moeda honesta (eqüiprovável) seria lançada 12 vezes por alguém escolhido criteriosamente; (b) o telepata transmissor (T) leria o resultado na própria moeda, após cada lançamento, e passaria o resultado mentalmente (sem verbalizar em voz alta) ao sensitivo receptor (R); (c) o sensitivo receptor (R) iria anotando, após cada transmissão, o resultado obtido. Acrescente-se que foram tomadas providências ambientais para garantir a incomunicabilidade não telepática entre T e R.
Após os 12 lançamentos, foram os seguintes os resultados:

Jogada	*Mensagem Transmitida por T*	*Mensagem Captada por R*	*Conclusão*
1	C	C	A
2	C	K	E
3	K	K	A
4	K	C	E
5	C	C	A
6	K	K	A
7	K	C	E
8	K	K	A
9	K	C	E
10	C	C	A
11	K	C	E
12	K	K	A

C = cara A = acerto
K = coroa E = erro

Se a variável de observação, X, for o *número de acertos* (A), qual será sua conclusão relativamente à H_0? Trabalhe com $\alpha_1 = 5\%$ e, depois, com $\alpha_2 = 1\%$.

19. Sabe-se que para $S_0 = V$ a H_0 foi não rejeitada a 5%. Então, para $\alpha = 1\%$, a H_0 deveria ter sido rejeitada. Verdadeiro ou falso? Por quê?

20. 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 absolutamente comunicação alguma entre os degustadores. Ao cabo do experimento, verificou-se que 15 respostas estavam corretas.

Pode-se afirmar, com $\alpha = 0{,}05$, que os degustadores conseguiram, *de fato*, reconhecer A?

21. Sabe-se que para $S_0 = W$ a H_0 foi rejeitada a 5%. Então, para $\alpha = 1\%$, a H_0 deveria ter sido (rejeitada/não rejeitada). Justificar.

11

Correlação

Características tais como orientação política, inteligência e classe social *variam* de um respondente para outro e são, por isso, chamadas *variáveis*. Em capítulos anteriores, ocupamo-nos com o estabelecimento da presença ou ausência de relação entre duas variáveis quaisquer, que, daqui por diante, chamaremos X e Y; por exemplo, relação entre orientação política (X) e método de educar filhos (Y), entre classe social (X) e inteligência (Y) ou entre orientação acadêmica (X) e uso de entorpecentes (Y). Auxiliados pela razão t, pela análise de variância ou pelo qui-quadrado, procuramos, páginas atrás, descobrir se uma diferença entre duas ou mais amostras podia ser considerada estatisticamente significativa—reflexo de uma verdadeira diferença populacional—e não apenas o resultado de erro de amostragem (ação do acaso).

FORÇA DA CORRELAÇÃO

Descobrir a existência de uma relação não esclarece muito a respeito do grau de associação ou *correlação* entre duas variáveis. Muitas são as relações estatisticamente significantes; poucas expressam correlação *perfeita* ou exata. Ilustremos: sabemos que peso e estatura são variáveis associadas, uma vez que, quanto mais alta a pessoa, maior tende a ser seu peso. Há numerosas exceções à regra, entretanto. Algumas pessoas altas pesam muito pouco; algumas pessoas baixas pesam muito. Da mesma forma, uma relação entre orientação para estudos universitários e uso de entorpecentes não prenuncia a possibilidade de encontrarmos centenas de não-viciados entre estudantes que pretendam continuar a vida acadêmica ou muitos viciados entre os que não planejam freqüentar a universidade.

As correlações, na verdade, variam com respeito à sua *força*. Podemos visualizar diferenças, quanto à força de correlação, por meio de um *diagrama de dispersão*, que é um gráfico capaz de mostrar a maneira pela qual os valores de duas variáveis, X e Y, distribuem-se ao longo da faixa dos possíveis resultados. Convencionalmente, num dia-

grama de dispersão, a variável X localiza-se no eixo horizontal, enquanto que a variável Y, no vertical.*

Na Figura 11.1, temos dois diagramas de dispersão: em ambos, há uma representação gráfica da relação entre "anos de escola" (X) e "renda" (Y). Observe-se que essa relação, na Figura 11.1(a), refere-se a homens, enquanto que na Figura 11.1(b), a mulheres. Note-se, ainda, que, nesses diagramas de dispersão, *todos* os pontos resultam do "cruzamento" de *dois* escores—educação e renda—relativos a *um* particular respondente. (Em outras palavras: a *cada* respondente correspondem *dois* escores e estes, em conjunto, dão origem a *um* ponto do diagrama de dispersão.) Na Figura 11.1(a), por exemplo, vemos que a um homem com 4 anos de escolarização corresponde uma renda de Cr$ 68.000,00 (anuais), enquanto que a um homem com 13 anos de estudos, a renda é de Cr$ 170.000,00.**

Podemos dizer que a força de correlação entre X e Y aumenta à medida que os pontos, no diagrama de dispersão, mais compactamente se agrupam em torno de uma *linha reta* imaginária. Portanto, a Figura 11.1(a) (homens) representa uma correlação *mais forte* do que a 11.1(b) (mulheres), muito embora ambos os diagramas indiquem que a renda *tende* a aumentar com o aumento dos anos de escolarização. Tais dados, na verdade, sugerem fortemente que a renda das mulheres (com relação à dos homens) está menos relacionada com o nível de escolaridade por elas atingido.

SENTIDO DA CORRELAÇÃO***

A correlação pode ser classificada, quanto ao *sentido*, em *positiva* ou *negativa*. Uma *correlação positiva* indica que os respondentes que obtiveram escores *altos* na variável X tendem a obter escores também *altos* na variável Y. De forma recíproca, respondentes que obtêm escores *baixos* em X tendem a obter escores também *baixos* em Y (e, nesse caso, a correlação *é também* positiva). A correlação positiva pode ser ilustrada a partir da relação entre escolaridade e renda. Como já vimos antes, respondentes com muitos anos de escolaridade tendem a apresentar rendas anuais maiores do que aqueles que freqüentaram a escola por poucos anos.

Diz-se que há *correlação negativa* quando, com relação aos mesmos respondentes, à medida que se obtêm escores *altos* na variável X, há a propensão de se obterem escores *baixos* na Y. Reciprocamente, ocorrerá *também correlação negativa* se, em correspondência a valores *baixos* na variável X, existir uma tendência a valores *altos* na variável Y. A relação entre escolaridade (anos de) e renda (anual) *não* caracteriza uma correlação

---

* N.T.: Trata-se de um sistema cartesiano de eixos ortogonais.

** N.T.: Em ambos os casos, mas principalmente no primeiro—a julgar pelo valor do salário mínimo vigente—trata-se de empregados com algum grau de especialização.

*** N.T.: No texto original, o Autor fala em *direção* de correlação. Parece que o termo *sentido* é mais rigoroso, já que a *mesma direção r* (reta *r*) pode ser "percorrida" em *dois sentidos* (que se opõem): de A para B e de B para A. Assim:

$$\frac{A \quad \rightarrow \quad B}{\leftarrow}$$

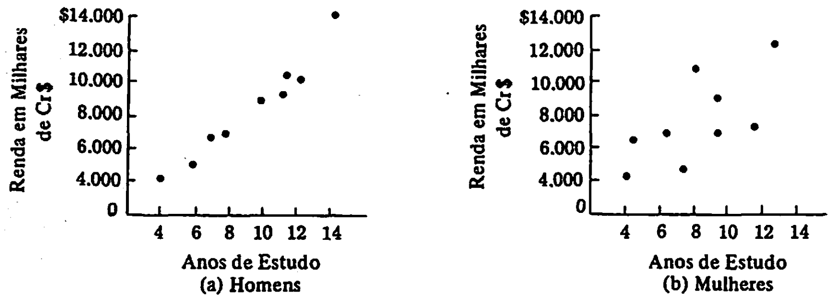


FIGURA 11.1 Diagramas de Dispersão Ilustrativos de Diferenças na "Força da Relação" entre "Escolaridade" e "Renda" (Homens e Mulheres).

negativa, uma vez que respondentes com muitos anos de estudo *não* tendem a apresentar rendas anuais baixas. Um exemplo mais característico de correlação negativa resulta da relação entre escolaridade (anos de) e reprovação nos grupos minoritários. A reprovação tende a diminuir à medida que o nível educacional aumenta. Portanto, sujeitos com pouca educação formal tendem a apresentar maiores índices de reprovação, enquanto que os que freqüentaram por mais tempo geralmente repetiram poucas vezes o ano.

CORRELAÇÃO CURVILÍNEA

Uma correlação positiva ou negativa representa um tipo de relação *linear*. Graficamente, os pontos* do diagrama de dispersão aglomerar-se-ão em torno da linha (reta) imaginária indicada na Figura 11.2(a). Por outro lado, se uma correlação negativa estiver presente, os pontos do diagrama de dispersão circundarão a reta imaginária conforme indica a Figura 11.2(b).

Na maior parte dos casos, os pesquisadores buscam estabelecer correlações lineares, sejam elas positivas ou negativas. É importante notar, entretanto, que nem todas as relações entre X e Y dão origem, necessariamente, a uma linha reta.** Existem muitas *correlações curvilíneas*, indicativas de que uma variável aumenta à medida que a outra *também* aumenta, até ocorrer uma "reversão"; a partir daí, uma das variáveis começa a decrescer, enquanto que a outra *continua* a crescer. Em outras palavras, uma relação entre X e Y que começa positiva torna-se negativa; uma relação que começa negativa torna-se positiva. Para ilustrar uma correlação curvilínea, consideremos a relação entre as variáveis "número de crianças" (tamanho da família) e "situação sócio-econômica". Como evidencia a Figura 11.3, os pontos do diagrama de dispersão tendem a formar uma curva em U—e não uma reta, à semelhança dos exemplos anteriores.

* N.T.: Fala-se costumeiramente em "nuvem de pontos".

** N.T.: Na verdade, uma *linha reta* (perfeita) só resulta de uma função matemática do tipo $Y = a + bX$. Nos demais casos, é a *linha interpolatriz* que melhor se ajusta a todos os pontos da nuvem, que "sugere" uma *reta*.

Assim, famílias da classe média têm pequeno número de filhos: o tamanho da família (Y) aumenta à medida que a situação sócio-econômica torna-se melhor *ou* pior.

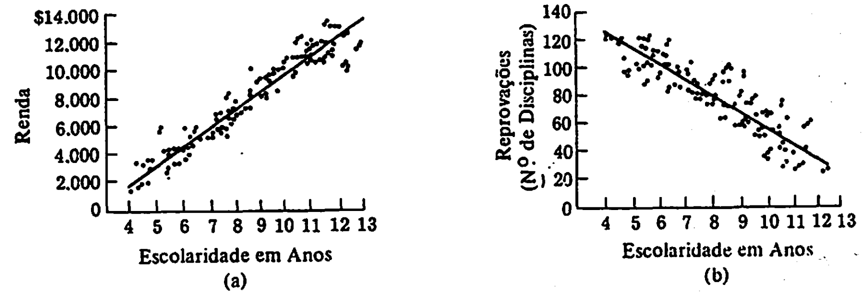


FIGURA 11.2 Diagramas de Dispersão: (a) Correlação Positiva entre Escolaridade e Renda; (b) Correlação Negativa entre Escolaridade e Reprovações.

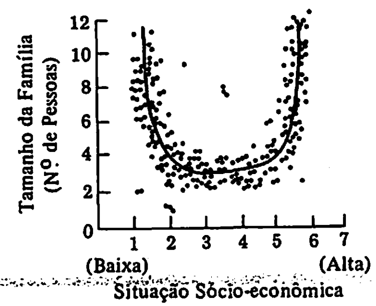


FIGURA 11.3 Relação entre "Situação Sócio-Econômica" (X) e "Tamanho da Família" (Y): Correlação Curvilínea.

COEFICIENTE DE CORRELAÇÃO

O procedimento que permite trabalhar com correlações não-lineares (curvilíneas) situa-se além do escopo deste texto. Por essa razão, vamos voltar nossa atenção para os chamados *coeficientes de correlação linear*, que expressam, numericamente, tanto a *força* quanto o *sentido* da correlação. Tais coeficientes de correlação oscilam entre $-1{,}00$ e $+1{,}00$ conforme se segue:

$-1{,}00$ ⟵ correlação negativa perfeita

⋮

$-0{,}95$ ⟵ correlação negativa forte

⋮

−0,50 ⟵ correlação negativa moderada
:
−0,10 ⟵ correlação negativa fraca
:
−0,00 ⟵ ausência de correlação
:
+0,10 ⟵ correlação positiva fraca
:
+0,50 ⟵ correlação positiva moderada
:
+0,95 ⟵ correlação positiva forte
:
+1,00 ⟵ correlação positiva perfeita

Vemos, pois, que valores numéricos *negativos*, tais como −1,00, −0,95, −0,50 e −0,10, indicam *correlação negativa*, enquanto que valores numéricos *positivos*, como, por exemplo, +1,00, +0,95, +0,50 e +0,10, são indicativos de *correlação positiva*. Em termos de grau de associação, quanto mais próximo de 1,00 em ambos os sentidos, maior a força da correlação. Como tal força é *independente* do seu sentido, podemos dizer que −0,10 e +0,10 são iguais quanto à força (ambos fracos); −0,95 e +0,95 também são iguais quanto a ela (ambos fortes).

COEFICIENTE DE CORRELAÇÃO PARA DADOS INTERVALARES

Com o auxílio do coeficiente de correlação (linear) de Pearson (r), podemos determinar a força e o sentido da relação entre as variáveis X e Y–desde que elas tenham sido mensuradas no nível intervalar. O r de Pearson reflete a extensão em que cada sujeito amostral consegue obter o mesmo escore z nas duas variáveis (X e Y). No caso de correlação positiva, ambos os escores z para um particular respondente têm o mesmo sinal, positivo ou negativo, e localizam-se aproximadamente à mesma distância da média (aritmética) de cada distribuição de escores. Assim, se o escore do sujeito A na variável X cair acima de $\bar{X}$, o seu escore na variável Y também cairá acima de $\bar{Y}$; se o escore de B, em X, cair abaixo de $\bar{X}$, o seu escore em Y também cairá abaixo de $\bar{Y}$. No caso de correlação negativa, os escores z de um particular respondente têm sinais opostos, o que indica que os zs são não só eqüidistantes das respectivas médias mas caem, com relação a elas, em lados opostos. Se o escore do sujeito A, em X, cair acima de $\bar{X}$ (isto é, se $X_A > \bar{X}$), o seu escore em Y cairá abaixo de $\bar{Y}$ (ou seja $Y_A < \bar{Y}$); se o sujeito B cair abaixo de $\bar{X}$ em X (se $X_B < \bar{X}$), ele cairá acima de $\bar{Y}$ (isto é, $Y_B > \bar{Y}$) em Y. A interpretação da correlação positiva e da negativa entre escores z foi ilustrada na Figura 11.4.

Podemos agora definir o r de Pearson como sendo a *média dos produtos dos escores z relativos às variáveis X e Y*. Em símbolos:

$$r = \frac{\Sigma(z_X z_Y)}{N}$$

onde

r = coeficiente de correlação (linear) de Pearson

z_X = escore z de um particular sujeito na variável X, e dado por $\dfrac{X - \bar{X}}{s_X}$

z_Y = escore z de um particular sujeito na variável Y, e dado por $\dfrac{Y - \bar{Y}}{s_Y}$

N = total de pares de dados (escores), isto é, total de "duplas" X e Y.

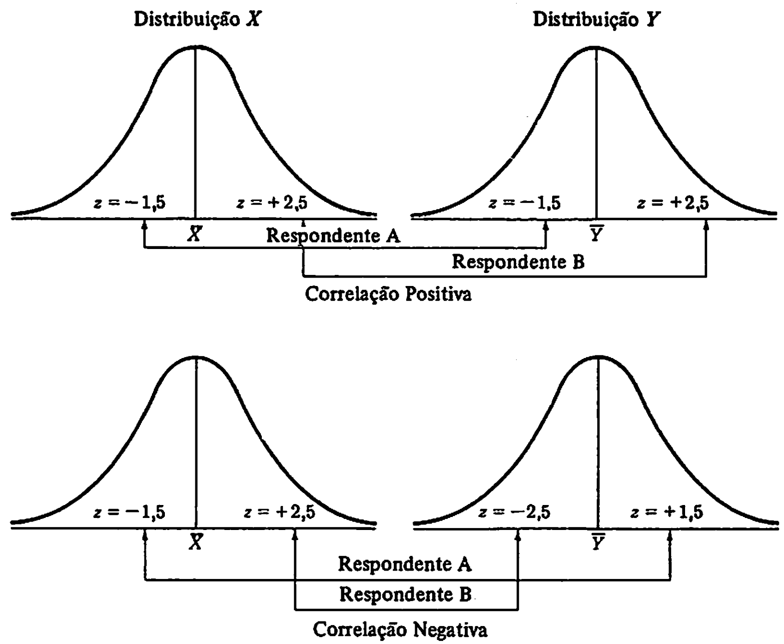


FIGURA 11.4 Intepretação Gráfica de uma Correlação Positiva e de uma Negativa com Base no Escore z.

Para ilustrar a aplicação do r de Pearson, vamos usar a fórmula acima na obtenção do coeficiente de correlação (linear) entre a variável X, número de anos completos de freqüência à escola atingido pelo pai, e a variável Y, número de anos de freqüência à escola completados pelo respectivo filho. Os dados da Tabela 11.1 respresentam essa relação numa amostra aleatória de sete respondentes.

TABELA 11.1 Relação entre Nível Educacional do Respondente e do Respectivo Pai. (Nível Educacional Mensurado em Anos Completos de Freqüência à Escola.)

	Anos de Escola	
Criança	*Pais (X)*	*Crianças (Y)*
A	12	12
B	10	8
C	6	6
D	16	11
E	8	10
F	9	8
G	12	11

Para aplicar a fórmula do r de Pearson, precisamos, antes, descobrir o valor de cada uma das seguintes estatísticas: $\bar{X}$, $\bar{Y}$, s_X e s_Y. Assim:

X	X^2	Y	Y^2
12	144	12	144
10	100	8	64
6	36	6	36
16	256	11	121
8	64	10	100
9	81	8	64
12	144	11	121
$\Sigma X = 73$	$\Sigma X^2 = 825$	$\Sigma Y = 66$	$\Sigma Y^2 = 650$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{73}{7} = 10{,}43 \qquad \bar{Y} = \frac{\Sigma Y}{N} = \frac{66}{7} = 9{,}43$$

$$s_X = \sqrt{\frac{\Sigma X^2}{N} - \bar{X}^2} = \sqrt{\frac{825}{7} - (10{,}43)^2} = \sqrt{117{,}86 - 108{,}78} = \sqrt{9{,}08} = 3{,}01$$

$$s_Y = \sqrt{\frac{\Sigma Y^2}{N} - \bar{Y}^2} = \sqrt{\frac{650}{7} - (9{,}43)^2} = \sqrt{92{,}86 - 88{,}92} = \sqrt{3{,}94} = 1{,}98$$

Agora para cada sujeito amostral, calculamos dois escores z: um relativo à variável X, outro à Y; depois, achamos o produto desses escores z:

	X	$X - \bar{X}$	$\frac{X - \bar{X}}{s_X}$	Y	$Y - \bar{Y}$	$\frac{Y - \bar{Y}}{s_Y}$	$z_X z_Y$
A	12	1,57	0,52	12	2,57	1,30	0,68
B	10	−0,43	−0,14	8	−1,43	−0,72	0,10
C	6	−4,43	−1,47	6	−3,43	−1,73	2,54
D	16	5,57	1,85	11	1,57	0,79	1,46
E	8	−2,43	−0,81	10	0,57	0,29	−0,24
F	9	−1,43	−0,48	8	−1,43	−0,72	0,34
G	12	1,57	0,52	11	1,57	0,79	0,41
							$\Sigma(z_X z_Y) = 5{,}29$

Para ilustrar o procedimento que permite obter z_X, z_Y e $z_X z_Y$ acima, vamos examinar as respostas do sujeito A, mensuradas das variáveis X e Y. Já sabemos que $\bar{X} = 10{,}43$ e que $s_X = 3{,}01$. Como $X - \bar{X} = 12 - 10{,}43 = 1{,}57$ para o sujeito amostral A, verificamos que sen $z_X = 1{,}57/3{,}01 = +0{,}52$. (Em outras palavras, os 12 anos de escola do sujeito A caem aproximadamente a meio desvio padrão acima da média da distribuição.) De modo análogo, sabemos que $\bar{Y} = 9{,}43$ e que $s_Y = 1{,}98$. Como $Y - \bar{Y} = 12 - 9{,}43 = 2{,}57$ para o sujeito amostral A, verificamos que sen $z_Y = 2{,}57/1{,}98 = +1{,}30$. (Em outras palavras, esses 12 anos de escola caem aproximadamente a um e um terço desvios padrões acima da média da distribuição.) Para obtermos o $z_X z_Y$ relativo ao sujeito A, é só multiplicar o escore $z = +0{,}52$ pelo escore $z = +1{,}30$ (donde: $0{,}52 \times 1{,}30 = 0{,}68$). Como evidencia a última coluna, à direita, na tabela acima, a soma desses produtos de escores z é 5,29.

Entrando com esses dados na fórmula de Pearson, obtemos:

$$r = \frac{\Sigma(z_X z_Y)}{N} = \frac{5{,}29}{7} = +0{,}75$$

No exemplo acima, o r de Pearson é igual a +0,75, o que indica uma correlação positiva razoavelmente forte entre o nível educacional atingido pelos filhos e o nível educacional atingido pelos respectivos pais. Em outras palavras, os respondentes cujos pais conseguiram um nível alto de escolaridade também apresentaram propensão a chegar a um nível alto de escolaridade; os respondentes cujos pais freqüentaram a escola durante poucos anos também demonstraram uma forte tendência a freqüentar a escola durante pouco tempo.

FÓRMULA PARA O CÁLCULO DO r DE PEARSON

O cálculo do r de Pearson a partir de escores z ajuda a relacionar o tópico da correlação com nossos estudos anteriores sobre escores padronizados e curva normal. Entretanto, a fórmula do r de Pearson baseada em escores z, requer muitos cálculos que consomem bastante tempo. Felizmente, existe uma fórmula para o cálculo do r de Pearson que trabalha de forma direta com escores brutos, eliminando-se, assim, a necessidade de obterem-se os produtos dos escores z para X e Y. A fórmula alternativa para o r de Pearson é

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}$$

onde

r = coeficiente de correlação (linear) de Pearson
N = número de pares de escores brutos ("duplas" formadas de X e Y)
X = escore bruto (na variável X)
Y = escore bruto (na variável Y)

A fim de ilustrar o uso da fórmula alternativa para o cálculo do r de Pearson, vamos voltar aos dados da Tabela 11.1, onde se procurou estabelecer a relação entre número de anos de escola cursados pelo pai (X) e o número de anos de escola cursados pelo respectivo filho (Y). A aplicação dessa fórmula implica que calculemos primeiro X^2, Y^2 e XY, conforme se segue:

X	X^2	Y	Y^2	XY
12	144	12	144	144
10	100	8	64	80
6	36	6	36	36
16	256	11	121	176
8	64	10	100	80
9	81	8	64	72
12	144	11	121	132
$\Sigma X = 73$	$\Sigma X^2 = 825$	$\Sigma Y = 66$	$\Sigma Y^2 = 650$	$\Sigma XY = 720$

Então, o coeficiente de correlação (linear) de Pearson é igual a

$$r = \frac{7(720) - (73)(66)}{\sqrt{[7(825) - (73)^2][7(650) - (66)^2]}}$$

$$= \frac{5040 - 4818}{\sqrt{(5775 - 5329)(4550 - 4356)}} = \frac{222}{\sqrt{(446)(194)}} = \frac{222}{\sqrt{86524}}$$

$$= \frac{222}{294{,}15} = +0{,}75$$

Teste da Significância do r de Pearson

O coeficiente de correlação (linear) de Pearson dá-nos uma medida precisa da *força* e do *sentido* da correlação (existente entre as variáveis) na *amostra* estudada. Se tivermos extraído uma amostra aleatória de uma particular população, podemos ainda querer verificar se a associação obtida entre X e Y existe de fato na *população*, e não resulta meramente de erro amostral (= ação do acaso).

Para testar a significância de uma medida de correlação, geralmente estabelecemos

a hipótese de que não existe correlação na população. Com respeito ao coeficiente de correlação (linear) de Pearson, a hipótese nula fixa que

$$r = 0,$$

enquanto que a hipótese experimental (hipótese alternativa, H_a) estabelece que

$$r \neq 0.$$

À semelhança do que fizemos em capítulos anteriores, testamos a hipótese selecionando um nível de significância igual a 0,05 ou 0,01,* aplicando, a seguir, a prova adequada. Para testarmos a significância do r de Pearson, podemos calcular uma estatística (razão) t com $(N - 2)$ graus de liberdade, sendo N o número de pares de escores. A razão t pode ser obtida mediante a fórmula:

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

onde

t = razão t (t observado, calculado) para testar a significância do r de Pearson
N = número de pares de escores (compostos de X e Y)
r = o próprio r de Pearson (que está sendo posto à prova).

Voltando ao exemplo anterior, podemos testar a significância de um coeficiente de correlação igual a +0,754, resultante da associação entre níveis educacionais de pais e respectivos filhos. Assim:

$$t = \frac{0{,}754\sqrt{5}}{\sqrt{1-(0{,}754)^2}}$$

$$= \frac{0{,}754(2{,}236)}{\sqrt{1-0{,}569}} = \frac{1{,}69}{\sqrt{0{,}431}} = \frac{1{,}69}{0{,}656} = 2{,}58$$

Ao consultar a Tabela C (fim do livro), verificamos que, para ser significante, a razão t deve igualar-se a ou exceder 2,57, ao nível de significância de 0,05, com 5 graus de liberdade. Como o nosso t calculado (t = 2,58) é maior do que o t tabelado (crítico), podemos rejeitar a hipótese nula de que $r = 0$ e aceitar a hipótese experimental de que $r \neq 0$. O nível educacional do respondente e o nível educacional de seu pai estão realmente associados na população.

* N.T.: Ou qualquer outro nível de significância, pois isso depende do rigor com que o pesquisador deseja trabalhar.

Método Simplificado para Testar a Significância de r

Felizmente, o processo ilustrado acima para testar a significância do r de Pearson, foi simplificado; assim, torna-se desnecessário–a não ser a título de exercício–calcular a razão t. É suficiente, pois, consultar a Tabela F, no fim do livro, onde figura uma lista de valores significantes do r de Pearson aos níveis de significância de 0,05 e 0,01, associados a graus de liberdade que vão de 1 a 90. Se for feita uma comparação direta entre o nosso valor calculado de r e o valor tabelado (Tabela F), obteremos o mesmo resultado que conseguiríamos se, de fato, calculássemos a razão t. No caso de o coeficiente de correlação de Pearson (calculado) ser menor que o valor *crítico* (tabelado), devemos aceitar a hipótese nula de que $r = 0$; se, por outro lado, o r calculado for igual a ou maior que o r crítico, rejeitamos a hipótese nula e aceitamos a hipótese experimental de que existe correlação na população.

A título de ilustração, vamos voltar ao nosso exemplo anterior, no qual um coeficiente de correlação igual a +0,754 foi testado por meio de uma razão t e considerado estatisticamente significante. Ao consultar a Tabela F, verificamos que, ao nível de 0,05 e com 5 graus de liberdade, um r, para ser significante e, assim, levar à rejeição da hipótese nula, deve ser, no *mínimo*, igual a 0,754. Vemos, por esse exemplo, que esse método simplificado leva-nos à mesma conclusão a que chegamos quando, pelo processo mais longo, calculamos o t.

Correlação: Uma Ilustração

Para ilustrar passo a passo o procedimento que permite obter o coeficiente de correlação linear de Pearson (r), vamos examinar a relação entre X = anos completos de freqüência à escola e Y = número de reprovações, na seguinte amostra aleatória de dez respondentes:

Respondente	*Anos de Escola* (X)	*Reprovações* (Y)[a]
A	10	1
B	3	7
C	12	2
D	11	3
E	6	5
F	8	4
G	14	1
H	9	2
I	10	3
J	2	10

(a) Os números dessa coluna indicam quantas reprovações cada sujeito teve em sua vida escolar.

Para calcular o r de Pearson, procede-se do seguinte modo:

PASSO 1: Calcular (1) ΣX, (2) ΣX^2, (3) ΣY; (4) ΣY^2 e (5) ΣXY

Respondente	X	X^2	Y	Y^2	XY
A	10	100	1	1	10
B	3	9	7	49	21
C	12	144	2	4	24
D	11	121	3	9	33
E	6	36	5	25	30
F	8	64	4	16	32
G	14	196	1	1	14
H	9	81	2	4	18
I	10	100	3	9	30
J	2	4	10	100	20
	$\Sigma X = 85$	$\Sigma X^2 = 855$	$\Sigma Y = 38$	$\Sigma Y^2 = 218$	$\Sigma XY = 232$
	(1)	(2)	(3)	(4)	(5)

PASSO 2: Entrar com os resultados do passo 1 na fórmula do r de Pearson:

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}$$

$$= \frac{10(232) - (85)(38)}{\sqrt{[10(855) - (85)^2][10(218) - (38)^2]}}$$

$$= \frac{2320 - 3230}{\sqrt{(8550 - 7225)(2180 - 1444)}} = \frac{-910}{\sqrt{(1325)(736)}} = \frac{-910}{\sqrt{975200}}$$

$$= \frac{-910}{987{,}52} = -0{,}92$$

O resultado encontrado indica a existência de apreciável correlação negativa entre "educação" (X) e "reprovações" (Y).

PASSO 3: Achar os graus de liberdade.

$$gl = N - 2 = 10 - 2 = 8$$

PASSO 4: Comparar o r de Pearson obtido com o r de Pearson crítico (Tabela F):

$$\begin{aligned} r \text{ calculado} &= -0{,}92 \\ r \text{ crítico} &= 0{,}63 \\ gl &= 8 \\ P &= 0{,}05 \end{aligned}$$

Como indicam os resultados acima, para que fosse possível rejeitar a hipótese nula de que $r = 0$, ao nível de significância de 5% e com 8 graus de liberdade, seria necessário que o nosso r calculado fosse, no mínimo, igual a 0,63. Como o r calculado é igual a –0,92, rejeitamos a hipótese nula e aceitamos a hipótese experimental.

Em outras palavras, nossos resultados sugerem a existência de correlação entre "educação" e "reprovação" na população da qual nossa amostra foi retirada.

Requisitos para o Uso do Coeficiente de Correlação de Pearson

A fim de empregar-se corretamente o coeficiente de correlação de Pearson como medida de associação entre as variáveis X e Y, os seguintes requisitos devem ser levados em conta:

1. *Correlação* linear*–O r de Pearson só se aplica a correlações lineares entre X e Y.
2. *Dados intervalares*–As variáveis X e Y devem ser mensuradas, no mínimo, a nível intervalar, de sorte que seja possível trabalhar com escores.
3. *Amostragem casual*–Os sujeitos amostrais devem ter sido extraídos aleatoriamente de uma dada população. Se assim não for, não terá nenhum sentido a prova de significância do coeficiente obtido.
4. *Variáveis distribuídas normalmente*–Para que seja possível testar a significância do r de Pearson, é necessário que ambas as variáveis, X e Y, tenham distribuição normal na população. Quando as amostras são pequenas, qualquer descuido na observância dessa normalidade de distribuição pode comprometer seriamente a validade do r de Pearson. Entretanto, esse requisito deixa de ter importância tão grande quando o tamanho das amostras é igual a ou maior que 30.

ANÁLISE DE REGRESSÃO

O estabelecimento de uma correlação entre duas variáveis pode ter utilidade na previsão dos valores de uma delas (Y) a partir do conhecimento dos valores da outra (X). A técnica empregada em tais previsões é conhecida por *análise de regressão*.

* N.T.: Não vá o leitor confundir *correlação* com *relação de causa e efeito*. Duas variáveis, X e Y, podem estar correlacionadas, mas isso não significa que uma delas seja a causa da outra. Mais formalmente, se $Y = f(X)$, nada obriga, além da correlação, que exista relação *causal*. Exemplo: suponha-se que um pesquisador resolvesse estudar a associação existente entre as seguintes variáveis: X = número de palavras que uma criança conhece e Y = tamanho do pé desta mesma criança. Admita-se, agora, o seguinte quadro:

Idade da Criança (em anos)	*Número de Palavras Conhecidas (X)*	*Número do Calçado (= tam. do pé) (Y)*
1	10	17
2	300	20
3	1.000	24
4	3.000	28
5	8.000	30

É óbvio que a correlação vai ser altíssima e positiva, o que não significa que o tamanho do pé possa ser explicado pela dimensão do vocabulário ou vice-versa. Mais coerente é pensar numa *relação triangular*, onde a *própria idade* é responsável tanto pelo crescimento do pé quanto do vocabulário.

Vimos, páginas atrás, que a *força* de uma correlação entre X e Y aumenta à medida que os pontos do diagrama de dispersão concentram-se em torno de uma reta imaginária. Podemos agora identificar essa curva* pelo nome de *reta de regressão*, que é uma reta interpolatriz obtida a partir da nuvem de pontos do diagrama de dispersão. O leitor atento perceberá que muitas são as retas que se "acomodam" aos pontos do diagrama, mas *uma* e *somente uma* constitui a *reta de regressão*. Isso porque o ajustamento dessa reta ao conjunto de pontos é o melhor possível, garantindo-se, assim, boas previsões de Y, a partir do conhecimento de X.

Previsão de Y a partir de X

Imaginem um estudo em que se procure estabelecer a correlação entre X = anos completos de freqüência à escola e Y = renda mensal. Seja uma amostra de seis respondentes, onde $r = +1,00$, isto é, onde há correlação positiva perfeita:

Respondente	*Anos de Escola (X)*	*Renda Mensal em Cr$ (Y)*
A	18	30.000,00
B	6	10.000,00
C	9	15.000,00
D	15	25.000,00
E	12	20.000,00
F	3	5.000,00

Como demonstra a Figura 11.5, podemos marcar os escores acima num gráfico e traçar por entre os pontos obtidos uma reta mediatriz. Essa mediatriz será a reta de regressão que une os escores de cada respondente da amostra. Tal linha (reta) de regressão permite fazer a seguinte previsão: um sujeito que tenha cursado a escola durante 18 anos poderá ganhar Cr$30.000,00 mensais; um outro, que tenha freqüentado apenas 3 anos, Cr$5.000,00 mensais e assim por diante:

Já foi salientado anteriormente que há poucas correlações perfeitas (+1,00 ou −1,00) que a natureza apresenta. Este fato é muito importante porque, como regra geral, as previsões tornam-se mais acuradas à medida que o coeficiente de correlação aproxima-se de ±1,00. Quando a correlação entre duas variáveis for forte, mas não perfeita, é possível, ainda assim, construir uma linha de regressão (previsão) que se "ajuste** bem" ao conjunto de pontos do diagrama. Isso é verdadeiro mesmo que nem todos os pontos da nuvem caiam exatamente *sobre* a reta (o que, aliás, é o mais comum); não ficamos impedidos de fazer previsões, mas devemos aceitar o fato de que a previsão será tanto mais imprecisa quanto mais distante o ponto estiver da linha de regressão. A Figura 11.6 ilustra o caso de uma correlação forte, mas não perfeita.

* N.T.: *Curva* é um nome genérico. Por isso nada impede, como, aliás, é o presente caso, que, na regressão linear, a curva seja uma *reta*.

** N.T.: Esse ajuste é conseguido graças à aplicação de uma "lei" denominada "Lei dos Mínimos Quadrados". As equações de regressão são conseqüência direta dessa lei.

Equação de Regressão
A equação de regressão, em símbolos, resulta na seguinte fórmula:

$$Y' = r\left(\frac{s_Y}{s_X}\right)X - r\left(\frac{s_Y}{s_X}\right)\overline{X} + \overline{Y}$$

onde

Y' = valor teórico (previsto) de Y (*Nota:* como se trata de uma previsão, Y' pode ser diferente de Y)
r = coeficiente de correlação linear de Pearson (r) para a relação entre as variáveis X e Y
s_Y = desvio padrão (amostral) da variável Y
s_X = desvio padrão (amostral) da variável X
X = um particular valor de X
$\overline{X}$ = média aritmética dos valores (amostrais) da variável X
$\overline{Y}$ = média aritmética dos valores (amostrais) da variável Y.

Para ilustrar o uso da fórmula de regressão na previsão dos valores de Y, vamos supor que entre as variáveis X = anos de freqüência à escola e Y = renda mensal tenhamos conseguido $r = +0{,}85$.

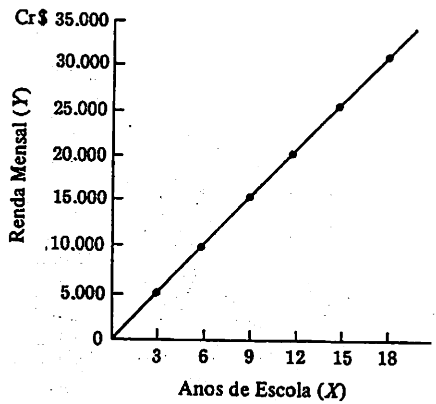


FIGURA 11.5 Linha (Reta) de Regressão Correspondente à Relação entre o Número de Anos Completos de Freqüência à Escola (X) e a Renda Mensal (Y), onde ($r = +1{,}00$)

Seja que:

r = +0,85
s_Y = 0,50
s_X = 0,40
$\overline{X}$ = 10 anos
$\overline{Y}$ = Cr$5.000,00

podemos, a partir daí, estabelecer a equação de regressão conforme se segue:

$$Y' = 0{,}85\left(\frac{0{,}5}{0{,}4}\right) X - 0{,}85\left(\frac{0{,}5}{0{,}4}\right) 10 + 5.000$$

$$= 1{,}06X - 1{,}06(10) + 5.000 = 1{,}06X - 10{,}6 + 5.000$$

$$= 1{,}06X + 4.989{,}4$$

O fim de prever (estimar) o valor de Y para cada X particular, basta entrarmos na fórmula com os valores de X. Por exemplo: qual a renda mensal prevista para um sujeito que tenha ficado 12 anos completos na escola?

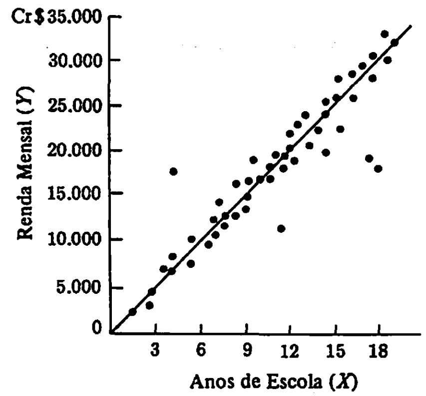


FIGURA 11.6 Linha (Reta) de Regressão Correspondente à Relação entre o Número de Anos Completos de Freqüência à Escola (X) e a Renda Mensal (Y), onde ($r < +1{,}00$)

Entrando com os dados na equação de regressão, vem que:

$$Y' = 1{,}06(12) + 4.989{,}4$$
$$= 12{,}72 + 4.989{,}4$$
$$= 5.002{,}12$$

Portanto, a previsão da renda mensal de alguém que tenha 12 anos de escola é de Cr$5.002,12.

Da mesma forma, podemos prever que um sujeito que tenha completado 6 anos de escola ganhe Cr$4.995,76, ou seja:

$$Y' = 1{,}06(6) + 4.989{,}4$$
$$= 6{,}36 + 4.989{,}4$$
$$= \text{Cr\$}4.995{,}76$$

Análise de Regressão: Ilustração

O estudo da análise de regressão pode ser ilustrado com maior profundidade retomando-se um exemplo anterior em que foram examinadas as variáveis X = nível educacional conseguido por pais e Y = nível educacional conseguido pelos respectivos filhos. Como vimos no início deste capítulo, tal relação, numa amostra de sete respondentes, produziu um r de Pearson igual a 0,75. Recapitulemos:

	Educação	
Respondente	*Pais* (X)	*Respondentes* (Y)
A	12	12
B	10	8
C	6	6
D	16	11
E	8	10
F	9	8
G	12	11

Podemos prever os valores de Y (educação dos filhos) a partir do conhecimento dos valores de X (educação dos respectivos pais), segundo o seguinte procedimento:

PASSO 1: Determinar o coeficiente de correlação linear de Pearson.

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}$$

$$= \frac{7(720) - (73)(66)}{\sqrt{[7(825) - (73)^2][7(650) - (66)^2]}}$$

$$= \frac{5.040 - 4.818}{\sqrt{(5.775 - 5.329)(4.550 - 4.356)}}$$

$$= \frac{222}{294,15} = 0,75$$

PASSO 2: Calcular as médias aritméticas amostrais para X e Y, isto é, $\overline{X}$ e $\overline{Y}$.

$$\overline{X} = \frac{\Sigma X}{N} = \frac{73}{7} = 10,43 \qquad \overline{Y} = \frac{\Sigma Y}{N} = \frac{66}{7} = 9,43$$

PASSO 3: Calcular os devios padrões amostrais para X e Y.

$$s_X = \sqrt{\frac{\Sigma X^2}{N} - \overline{X}^2} = \sqrt{\tfrac{825}{7} - (10,43)^2} = \sqrt{117,86 - 108,79} = \sqrt{9,07} = 3,01$$

$$s_Y = \sqrt{\frac{\Sigma Y^2}{N} - \overline{Y}^2} = \sqrt{\tfrac{650}{7} - (9,43)^2} = \sqrt{92,86 - 88,93} = \sqrt{3,93} = 1,98$$

PASSO 4: Entrar com os valores obtidos nos Passos 1, 2 e 3 na equação de regressão.

$$Y' = r\left(\frac{s_Y}{s_X}\right) X - r\left(\frac{s_Y}{s_X}\right)\overline{X} + \overline{Y}$$

$$= 0,75 \left(\frac{1,98}{3,01}\right) X - 0,75 \left(\frac{1,98}{3,01}\right) 10,43 + 9,43$$

$$= 0,75(0,66)X - 0,75(0,66)10,43 + 9,43 = 0,50X - 5,22 + 9,43$$

$$= 0,50X + 4,21$$

PASSO 5: Determinar os valores de Y' para os vários valores de X.

[Exemplos]

1. Para um respondente cujo pai tenha completado 16 anos de estudos:

$$Y' = 0,50X + 4,21 = 0,50(16) + 4,21 = 8,0 + 4,21 = 12,21$$

2. Para um respondente cujo pai tenha completado 6 anos de estudos:

$$Y' = 0,50X + 4,21 = 0,50(6) + 4,21 = 3,0 + 4,21 = 7,21$$

Conclusão: temos condições de prever (estimar) que respondentes cujos pais tenham concluído 16 anos de estudos cursaram/cursarão 12,21 anos; por igual raciocínio, os filhos de pais que estudaram 6 anos freqüentaram/freqüentarão escola durante 7,21 anos.*

---

* N.T.: Observem-se três aspectos importantes:

1. Análise de regressão não é adivinhação. Também não é um processo de "acertar a mosca". O leitor deve procurar perceber que o conjunto de pontos do diagrama de dispersão define uma *tendência*. A reta interpolatriz (isto é, de regressão) apenas "sintetiza" essa tendência e permite uma generalização, sob a forma de equação, *como se* todos os pontos caíssem sobre ela.

2. Do que ficou dito em (1), acima, decorre que:
   a) se $Y = f(X)$ for *função matemática*, as previsões serão *pontuais*, isto é, *por ponto*, e a precisão será grande;
   b) se $Y = f(X)$ for *função estatística*, as previsões serão *intervalares*, isto é, os valores de Y' pertencerão a um *intervalo*. Daí que $(Y' - Y)$ = erro de estimação.

COEFICIENTE DE CORRELAÇÃO PARA DADOS ORDINAIS

Até aqui, estivemos trabalhando com o r de Pearson, que é um coeficiente de correlação aplicável apenas a dados que possam ser mensurados, no mínimo, ao nível intervalar. Voltamos agora ao problema da determinação do grau de associação entre variáveis ordinais–variáveis às quais é possível atribuir postos ou "graduação" a partir da presença de uma característica particular.

Tomemos um exemplo do campo social. Consideremos a relação entre a "situação sócio-econômica" (X) e o "tempo despendido diante da TV" (Y). Imaginemos que nossa amostra conste de oito respondentes e que, depois da atribuição de postos, tenha resultado o seguinte quadro:

Respondente	Situação Sócio-Econômica (X) Posto		Tempo Gasto Diante da TV (Y) Posto	
A	1	situação sócio--econômica mais alta	2	
B	2		1	maior tempo diante da TV
C	3		3	
D	4		5	
E	5		4	
F	6		8	
G	7		6	
H	8		7	

Como podemos observar, o respondente A recebeu o posto 1 com relação ao nível sócio-econômico, mas posto 2 quanto ao tempo despendido diante da TV; o respondente B recebeu posto 2 com respeito ao nível sócio-econômico e posto 1 quanto ao tempo gasto diante do vídeo.

Para determinar o grau de associação entre essas duas variáveis, podemos aplicar o *coeficiente de correlação de postos de Spearman* (r_s). Em símbolos:

$$r_s = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)}$$

onde

3. A equação $Y' = r\left(\frac{s_Y}{s_X}\right)X - r\left(\frac{s_Y}{s_X}\right)\bar{X} + \bar{Y}$ é conhecida também por *equação de regressão dos Y sobre os X*. Entretanto, pode ocorrer que, dado um Y, o pesquisador resolva calcular o X correspondente. Obviamente, o valor conseguido, à semelhança de Y', será um valor também aproximado, donde: $(X' - X)$ = erro de estimação. Nesse último caso, quando se trata de buscar um X a partir de um Y conhecido, a equação denomina-se *regressão dos X sobre os Y*. Ei-la:

$$X' = r\left(\frac{s_X}{s_Y}\right)Y - r\left(\frac{s_X}{s_Y}\right)\bar{Y} + \bar{X}$$

r_s = coeficiente de correlação de postos

D = diferença entre postos (relativa ao *mesmo* sujeito em *ambas* as variáveis).

N = número de respondentes (isto é, tamanho da amostra).

Vamos considerar os dados tal como figuram na Tabela 11.2 e aplicar a fórmula correlacional de Spearman:

TABELA 11.2 Relação entre Situação Sócio-Econômica e Tempo Gasto Diante da TV

Respondente	Situação Sócio-Econômica X	Tempo Gasto Diante da TV Y	D	D^2
A	1	2	−1	1
B	2	1	1	1
C	3	3	0	0
D	4	5	−1	1
E	5	4	1	1
F	6	8	−2	4
G	7	6	1	1
H	8	7	1	1
				$\Sigma D^2 = 10$

$$r_s = 1 - \frac{6(10)}{8(64-1)} = 1 - \frac{60}{8(63)} = 1 - \frac{60}{504} = 1 - 0{,}12 = +0{,}88$$

Encontramos, portanto, uma correlação positiva forte ($r_s = +0{,}88$), entre "situação sócio-econômica" e "tempo gasto diante do vídeo": respondentes que apresentam uma situação sócio-econômica alta tendem a despender muito tempo diante da TV; respondentes de baixo nível sócio-econômico tendem a ficar pouco tempo assistindo a programas de TV.

Procedimento para Lidar com Postos Empatados (Espelhados)

Na prática, nem sempre é possível atribuir postos ou ordenar nossos respondentes sem que, em algum momento, ocorram postos empatados (espelhados). Poderíamos descobrir, por exemplo, que dois ou mais respondentes despendem exatamente o mesmo tempo diante da televisão, ou que o desempenho acadêmico de dois universitários é semelhante ou, ainda, que diversos respondentes possuem o mesmo QI.

Para ilustrar o procedimento que permite calcular o coeficiente de correlação de postos, no caso de alguns deles estarem empatados, admita-se que estejamos interessados no estabelecimento do grau de associação entre "posição acadêmica" (X) e QI (Y). Suponha-se também que nos seja possível não só ordenar dez alunos graduandos (do melhor para o pior) como também mensurar seus QIs. A tabela resultante apresentaria a seguinte disposição:

Respondente	Classificação X		QI Y
A	10	←	110
B	9	último lugar	90
C	8		104
D	7		100
E	6		110
F	5		110
G	4		132
H	3		115
I	2	primeiro lugar	140
J	1	←	140

Antes de prosseguir com os cálculos que levam à obtenção do r_s—coeficiente de Spearman—vamos atribuir postos aos QIs dos nossos dez graduandos:

Respondente	QI	Posto ao QI	
A	110	7	
B	90	10	
C	104	8	os postos 5, 6 e 7
D	100	9	estão empatados
E	110	6	
F	110	5	
G	132	3	
H	115	4	
I	140	2	os postos 1 e 2
J	140	1	estão empatados

Como podemos observar, os sujeitos I e J obtiveram os QIs mais altos e estão, por isso, empatados (os postos 1 e 2 "pertencem" simultaneamente a ambos). Da mesma forma, os sujeitos A, E e F estão empatados com relação à mesma variável—QI—e, com isso, os postos 5, 6 e 7 encontram-se "bloqueados".

Para determinar a posição exata de cada sujeito, no caso de postos empatados, devemos *somar esses postos* (o "valor" numérico deles) *e dividir o total pelo número de empates.* Portanto, no caso de um QI = 140, em que ficaram bloqueados os postos 1 e 2; o procedimento consiste em achar uma espécie de "média":

$$\frac{1+2}{2} = 1{,}5$$

Por igual raciocínio, relativamente ao QI = 110, o procedimento é:

$$\frac{5+6+7}{3} = 6{,}0$$

Tendo encontrado o "posto-posição" de cada QI, podemos dar continuidade ao cálculo do r_s conforme mostra a Tabela 11.3.

TABELA 11.3 Relação entre Classificação Acadêmica e QI

Respondente	Classificação Acadêmica (X)	QI (Y)	$X - Y = D$	D^2
1	10	6	4,0	16,00
2	9	10	−1,0	1,00
3	8	8	0	0
4	7	9	−2,0	4,00
5	6	6	0	0
6	5	6	−1,0	1,00
7	4	3	1,0	1,00
8	3	4	−1,0	1,00
9	2	1,5	0,5	0,25
10	1	1,5	−0,5	0,25
				$\Sigma D^2 = 24{,}50$

Entrando agora, com os dados acima, na fórmula, vem que:

$$r_s = 1 - \frac{6(24{,}50)}{10(100-1)}$$

$$= 1 - \frac{147}{990} = 1 - 0{,}15 = +0{,}85$$

O coeficiente de correlação de postos resultante indica que existe uma *correlação positiva bem forte* entre *X* e *Y*. Em outras palavras, universitários com QI *alto* tendem a classificar-se *bem* em sua classe e universitários com QI *baixo* tendem a classificar-se *mal.*

Prova de Significância para o Coeficiente de Spearman (r_s)

Como proceder para testar a significância do coeficiente de correlação de Spearman? Exemplo: como podemos saber se o coeficiente obtido ($r_s = 0{,}85$) entre *X* (classificação acadêmica) e *Y* (QI) é significante, isto é, se a "informação" pode ser generalizada para a população toda? Para testar a significância de um r_s calculado, basta consultar a Tabela G, no fim do livro, onde estão arrolados, aos níveis de significância de 5% e 1%, os r_s *críticos*. Observe-se que, neste caso, a consulta leva em conta *apenas o número de pares de escores* (*N*) e não, como noutros testes, os graus de liberdade. Neste problema, $N = 10$ e, conforme a Tabela G, para que um r_s seja *significante* (ao nível de 0,05), é preciso que ele obedeça à seguinte condição: $r_s \geqslant 0{,}648$. Como o nosso r_s observado é igual a 0,85 (portanto *maior* que o r_s crítico), rejeitamos a hipótese nula de que $r_s = 0$ e aceitamos a hipótese experimental de que "classificação acadêmica" e "QI" são variáveis *correlacionadas* não só na amostra como *também* na *população*.

Correlação de Postos: Ilustração

Vamos resumir o procedimento que permite obter o coeficiente de correlação de postos utilizando, para isso, mais um exemplo. Seja o cálculo da relação entre o "grau de participação voluntária em associações" (X) e o "número de amigos íntimos" (Y). Suponha-se, ainda, uma amostra de cinco respondentes, conforme o quadro

Respondente	Participação Voluntária em Associações (X). Posto		Número de Amigos (Y)
A	1	⟵ participa mais	6
B	2		4
C	3		6
D	4		2
E	5	⟵ participa menos	2

Para determinar o grau de associação entre X e Y, procedemos da seguinte maneira:

PASSO 1: Atribuir postos aos respondentes tanto na variável X quanto na Y. Como demonstra o quadro acima, a distribuição de postos com respeito à variável X–participação voluntária em associações–obedece à seguinte regra: posto 1 para o respondente que participa "mais" e 5 para o que participa "menos".

Precisamos, também, atribuir postos aos mesmos respondentes sob a variável Y–número de amigos íntimos. No exemplo em foco, ocorrem postos empatados, conforme a demonstração dada a seguir:

Número de Amigos (Y)	Posto	
6	1	postos 1 e 2 "bloqueados"
4	3	
6	2	
2	4	postos 4 e 5 "bloqueados"
2	5	

Para trabalhar com postos empatados, usamos a "média" dos empates:

Para a primeira e segunda posições: $\dfrac{1+2}{2} = 1{,}5$

Para a quarta e quinta posições: $\dfrac{4+5}{2} = 4{,}5$

Portanto:

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

PASSO 2: Calcular ΣD^2. Para tanto, calcular as diferenças entre os postos (colunas X e Y), quadrar estas diferenças (D^2) e somar os quadrados resultantes (ΣD^2):

X	Y	D	D^2
1	1,5	−0,5	0,25
2	3,0	−1,0	1,00
3	1,5	1,5	2,25
4	4,5	−0,5	0,25
5	4,5	0,5	0,25
		ΣD^2 =	4,00

PASSO 3: Entrar com o resultado do Passo 2 na fórmula do coeficiente de correlação de postos de Spearman:

$$r_s = 1 - \frac{6\Sigma D^2}{N(N^2-1)} = 1 - \frac{6(4)}{5(24)} = 1 - \frac{24}{120}$$

$$= 1 - 0{,}20 = +0{,}80$$

PASSO 4: Comparar o r_s obtido (calculado) com o r_s crítico na Tabela G.

r_s calculado = 0,80
r_s crítico = 1,00
N = 5
P = 0,05

Volvendo à Tabela G, no fim do livro, verificamos que, ao nível de significância de 0,05 e com uma amostra de tamanho 5, necessitaríamos de um r_s observado igual a 1,00 (correlação perfeita) para que fosse possível rejeitar a hipótese nula. Portanto, embora tenhamos conseguido uma forte correlação positiva na nossa amostra, somos obrigados a aceitar a hipótese nula segundo a qual $r_s = 0$. Nossos resultados não podem ser generalizados para a população da qual a amostra foi extraída.

Requisitos Para o Uso do Coeficiente de Correlação de Postos de Spearman

O coeficiente de correlação de postos de Spearman (r_s) deve ser empregado somente quando as seguintes condições podem ser satisfeitas:

1. *Correlação linear*–O r_s (coeficiente de correlação de postos de Spearman) só detecta correlações lineares entre X e Y, quaisquer que sejam essas variáveis.
2. *Dados ordinais*–As variáveis X e Y devem ser passíveis de ordenação; como alternativa, deve ser possível atribuir-lhes postos.
3. *Amostragem casual*–Os sujeitos amostrais devem ter sido extraídos aleatoriamente de uma dada população.

GAMA DE GOODMAN E KRUSKAL

A correlação pode ser interpretada em termos do grau em que os valores de uma variável podem ser previstos (ou adivinhados) a partir do conhecimento dos valores da outra variável. Isso pode ser diretamente observado no coeficiente *gama* de Goodman e Kruskal (G), que é uma alternativa ao coeficiente de correlação de postos (r_s). Aliás, muitos pesquisadores dão preferência ao gama quando se trata de medir o grau de associação entre variáveis mensuradas ao nível ordinal.

A fórmula básica para o cálculo do *gama* é:

$$G = \frac{\Sigma f_a - \Sigma f_i}{\Sigma f_a + \Sigma f_i}$$

onde

f_a = freqüência dos "acordos"
f_i = freqüência das "inversões".

Podemos interpretar os *acordos* e as *inversões* como uma expressão do *sentido* da correlação entre as variáveis X e Y. Um acordo perfeito indica uma correlação positiva perfeita (+1,00): tal situação ocorre quando os sujeitos estudados receberam *postos exatamente na mesma ordem em ambas as variáveis*. O quadro abaixo demonstra esse fato muito bem: um sujeito recebe o posto 1 na variável X e *também* o posto 1 na variável Y; um outro sujeito recebe o posto 2 na variável X e também na Y e assim por diante.

	Postos	
Sujeitos	*X*	*Y*
A	1	1
B	2	2
C	3	3
D	4	4
E	5	5
F	6	6

Por outro lado, uma inversão perfeita indica uma correlação negativa perfeita (−1,00), de forma que os sujeitos estudados recebem postos em ordem rigorosamente inversa nas duas variáveis. Assim, um sujeito que receba o posto 1 em X receberá o último posto em Y; um outro sujeito que receba o posto 2 em X receberá o penúltimo posto em Y e assim por diante.

	Postos	
Sujeitos	*X*	*Y*
A	1	6
B	2	5
C	3	4
D	4	3
E	5	2
F	6	1

Sempre que ocorrer acordo perfeito ou inversão perfeita, torna-se possível prever, com precisão total, o posto de um sujeito numa variável a partir do conhecimento do posto desse sujeito na outra variável. No caso de acordo perfeito, por exemplo, sabemos que a pessoa que recebe posto 3 na variável X também recebe o 3 em Y. Uma vez que a ocorrência de correlação perfeita raramente ocorre na natureza, nossa capacidade para fazer previsões acuradas sobre uma variável a partir do conhecimento que temos da outra vai depender da *quantidade* de acordos ou de inversões na ordenação de postos dos indivíduos em ambas as variáveis.

Coeficiente Gama (G): Ilustração

Para ilustrar o uso de *gama*, vamos admitir que estivéssemos estudando a relação entre (X), o tamanho da população negra de áreas metropolitanas dos Estados Unidos da América do Norte, e (Y), o respectivo nível de discriminação em emprego. Tal estudo poderia ser levado a efeito, por exemplo, analisando os dados relacionados com populações e renda ora disponíveis no *U.S. Bureau of the Census.**

Suponha-se que tenhamos sido capazes de ordenar e atribuir postos às seis maiores áreas metropolitanas dos EUA relativamente a X = tamanho da população negra e Y = nível de discriminação em emprego. Seja então o quadro:

Área Metropolitana	*Tamanho da População Negra (X)*	*Nível de Discriminação em Emprego (Y)*
A	6	4
B	1	2
C	2	3
D	5	5
E	4	6
F	3	1

Vemos, assim, que a área metropolitana A apresentou o menor número de negros e classificou-se em quarto lugar com respeito à discriminação; a área metropolitana B apresentou a maior população negra e classificou-se em segundo lugar quanto à discriminação e assim por diante.**

PASSO 1: Redistribuir os dados de forma que a variável X fique perfeitamente ordenada (do posto menor ao maior). Para determinar o grau de associação entre o tamanho da população negra (X) e a respectiva discriminação em emprego (Y), levamos, primeiro, a variável X a uma tabela em que seus postos figurem de forma ordenada de 1 a 6; os postos da variável Y *apenas acompanham* os postos de X, o que fatalmente determina

* N.T.: Serviço de Levantamento de Dados Demográficos dos EUA.
** N.T.: Observe-se que o *menor posto* (em ambas as variáveis) corresponde à *maior população* ou ao *maior nível discriminador.*

uma *ordem caótica*.* As freqüências dos acordos e das inversões na coluna desordenada (*Y*) indicam o quanto essa coluna de postos difere de uma ordenação perfeita, seja ela positiva (1, 2, 3, 4, 5, 6) ou negativa (6, 5, 4, 3, 2, 1):

Área Metropolitana	*Tamanho da População Negra (X)*	*Nível de Discriminação em Emprego (Y)*
B	1	2
C	2	3
F	3	1
E	4	6
D	5	5
A	6	4

PASSO 2: Obter as freqüências dos acordos. Para obter tais freqüências (f_a), começamos com o primeiro posto no topo da coluna dos *Y*s (isto é, o posto que corresponde à área metropolitana B). Para cada posto, contamos o número de *dados situados acima dele na tabela, porém que tenham menor valor numérico*. O número de postos acima do primeiro (sem levar em consideração seu valor numérico) é *sempre zero* (uma vez que, do ponto de vista estritamente "visual", não há nenhum posto antes do primeiro na tabela). Como resultado, entramos com um zero na coluna "acordos" para a área metropolitana B. Observando o segundo posto (3) listado na coluna *Y* (área metropolitana C), contamos o número de postos que estão acima dele na tabela e que têm menor valor numérico. Vemos que apenas o posto 2 está, na tabela, acima dele (3). E, uma vez que esse posto (2) é menor que 3, entramos com o valor 1 na coluna de acordos. Em continuação ao processo, o próximo posto listado (para a área metropolitana F) é 1. Como os postos que estão acima dele, na tabela, são 3 e 2–e como ambos (3 e 2) são *maiores* que 1–entramos com um zero na coluna de acordos. O posto seguinte na coluna dos *Y*s é 6 e corresponde à área metropolitana E. Os postos que, na tabela, estão acima dele, são 1, 3 e 2 e estes, sem exceção, são *menores* que 6. Em outras palavras, se $(1 < 6)$, $(3 < 6)$ e $(2 < 6)$, há 3 postos menores que 6 e, por isso, entramos com o valor 3 na coluna de acordos. Procedendo de forma idêntica com relação aos demais postos da coluna *Y* (relativos às áreas D e A), vamos encontrar duas vezes o valor 3, que deve ser levado à coluna de acordos. Assim:

Área Metropolitana	*Tamanho da População Negra (X)*	*Nível de Discriminação em Emprego (Y)*	*Acordos*
B	1	2	0
C	2	3	1
F	3	1	0
E	4	6	3
D	5	5	3
A	6	4	3

* N.T.: Se esse procedimento de ordenação tivesse sido feito com relação à variável *Y*, a ordem caótica estaria associada aos postos da variável *X*. Por outro lado, entenda-se por *ordem caótica* qualquer seqüência numérica que não apresente uma *lei de formação*.

PASSO 3: Obter as freqüências das inversões. Para encontrar as freqüências das inversões, novamente começamos com o primeiro posto da coluna dos *Y*s (isto é, o posto que corresponde à área metropolitana B). Desta vez, entretanto, para *cada* posto, o procedimento será o seguinte: *contar o número de dados que, na tabela, situam-se acima de um particular posto e que, em valor numérico, são maiores do que ele.* Iniciando com o posto que figura no topo da coluna *Y*, verificamos que, em termos de *tabela*, não ocorre nenhum dado antes dele (isto é, antes do posto 2); por isso, entramos com um zero na coluna das inversões. Tomando, agora, o segundo posto na coluna dos *Y*s (3), isto é, o que corresponde à área metropolitana C, contamos o número de dados que o antecedem e são maiores do que ele. Uma simples inspeção da tabela revela-nos que *antes* do posto 3 só figura um dado: o relativo ao posto 2 (área metropolitana B) e que, sendo $2 < 3$, só é possível entrar com um zero na coluna das inversões. Prosseguindo lista abaixo, o próximo posto é o que corresponde à área metropolitana F (posto 1). Como antes desse posto, 1, na coluna dos *Y*s, figuram os dados (postos) 3 e 2, e sendo $(3 > 1)$ e $(2 > 1)$, entramos com o valor 2 na coluna das inversões. Ao descer mais um pouco encontramos o posto 6, relativo à área metropolitana E. Observamos que, na tabela, três dados precedem esse posto: 1, 3 e 2; porém, como nenhum deles supera, em valor numérico, o 6, novamente escrevemos zero na coluna das inversões. O processo continua até que se esgotem todos os postos da coluna *Y*, dando origem aos dois seguintes valores na coluna das inversões: 1 e 2 (o leitor poderá fazer seus próprios cálculos para comprovar).

Área Metropolitana	*Tamanho da População Negra (X)*	*Nível de Discriminação em Emprego (Y)*	*Inversões*
B	1	2	0
C	2	3	0
F	3	1	2
E	4	6	0
D	5	5	1
A	6	4	2

PASSO 4: Obter Σf_a e Σf_i. Uma vez que todos os acordos e todas as inversões tenham sido estabelecidos, somamos separadamente essas colunas, conforme abaixo:

	Acordos	*Inversões*
B	0	0
C	1	0
F	0	2
E	3	0
D	3	1
A	3	2
	$\Sigma f_a = 10$	$\Sigma f_i = 5$

PASSO 5: Entrar com Σf_a e Σf_i na fórmula do coeficiente gama:

$$G = \frac{\Sigma f_a - \Sigma f_i}{\Sigma f_a + \Sigma f_i} = \frac{10 - 5}{10 + 5} = \frac{5}{15} = +0{,}33$$

Um coeficiente *gama* igual a +0,33 indica a presença de uma correlação positiva fraca. Essa correlação baseia-se numa dominância de acordos: o número de acordos é, aproximadamente, 33% maior do que o número de inversões entre o tamanho da população negra (X) e a discriminação em emprego (Y) nos EUA.*

Procedimento Para Postos Empatados (Espelhados)

Como já foi visto com relação ao coeficiente de correlação de postos (de Spearman), nem sempre é possível evitar postos empatados ao nível ordinal de mensuração. Na verdade, os pesquisadores se vêem, com freqüência, às voltas com medidas ordinais, brutas, que dão origem a um número bem grande de postos empatados. Nos casos em que um número muito grande deles ocorrer, o coeficiente gama torna-se uma medida de associação particularmente útil. A fórmula básica para o cálculo do coeficiente gama pode ser empregada mesmo quando há postos empatados, desde que as freqüências dos acordos e das inversões sejam calculadas mediante uma "adaptação".

Vamos ilustrar o procedimento de obtenção do coeficiente gama numa situação de postos empatados. Suponha-se que um pesquisador quisesse examinar a relação entre "classe social" e "adesão voluntária a uma associação". Suponha-se, ainda, que ele tivesse obtido os seguintes dados após examinar questionários distribuídos a e respondidos por 80 moradores de uma cidade: entre 29 respondentes da classe alta (A), 15 demonstraram grande interesse, 10 interesse moderado e 4 pouco interesse em aderir voluntariamente a uma determinada associação; entre 25 respondentes da classe média (B), 8 demonstraram grande interesse, 10 interesse moderado e 7 pouco interesse; e, para finalizar, entre 26 respondentes da classe baixa (C), 7 demonstraram grande interesse, 8 interesse moderado e 11 pouco interesse. Observe-se que postos empatados ocorrem em todas as posições. Por exemplo, 29 respondentes empataram suas respostas na coluna classe alta, no ponto mais alto da variável X.

PASSO 1: Redispor os dados numa tabela de freqüências.

Adesão Voluntária a uma Associação (Y)	*Alta (A)*	*Classe Social (X)* *Média (B)*	*Baixa (C)*
Alta	15	8	7
Moderada	10	10	8
Baixa	4	7	11
	29	25	26
		$N = 80$	

* N.T.: É bom insistir no fato de que a preocupação do Autor foi ilustrar um *procedimento estatístico* e não criticar um quadro social eventualmente existente nos EUA. Além disso, os dados usados bem podem ser *fictícios*.

Observe-se que o quadro acima caracteriza uma tabela 3 X 3, onde existem 9 caselas (3 linhas X 3 colunas = 9 caselas). Para garantir que o sinal do coeficiente gama seja fixado com o devido rigor (positivo ou negativo), a variável X (nas colunas) deve ser sempre disposta em ordem decrescente da esquerda para a direita. Por exemplo, na tabela, a classe social (X) decresce da esquerda para a direita, "assumindo" os valores "alta", "média" e "baixa". De modo análogo, a variável Y, nas linhas, deve decrescer de cima para baixo. Assim, nessa tabela, a adesão voluntária a uma associação (Y) varia, no sentido descendente, assumindo os valores "alta", "moderada" e "baixa".

PASSO 2: Calcular Σf_a. Para encontrar Σf_a, comece pela casela onde $f = 15$ (canto esquerdo superior). Multiplique essa freqüência (15) pela soma de todas as outras freqüências do quadro que, com relação ao 15, estão *à direita e abaixo*. Lendo da esquerda para a direita, observamos que as freqüências que estão abaixo e a direita de 15 são: 10, 8, 7 e 11. Repita-se este procedimento relativamente a todas as caselas que admitam freqüências nas posições (simultâneas) "abaixo e à direita". Assim, trabalhando da esquerda para a direita, vem que:

Classe A/Adesão alta	15(10 + 8 + 7 + 11)	=	15(36)	=	540
Classe B/Adesão alta	8(8 + 11)	=	8(19)	=	152
Classe A/Adesão moderada	10(7 + 11)	=	10(18)	=	180
Classe B/Adesão moderada	10(11)	=	110		

(Note-se que nenhuma outra casela da tabela admite cálculos semelhantes aos que acabamos de fazer, uma vez que para as freqüências 7 (1.ª linha), 8 (2.ª linha), 4, 7 e 11 (última linha) as posições simultâneas "abaixo e à direita" não têm sentido.)

Σf_a é a soma dos produtos acima. Portanto:

$$\begin{aligned} \Sigma f_a &= 540 + 152 + 180 + 110 \\ &= 992 \end{aligned}$$

PASSO 3: Calcular Σf_i. Para encontrar Σf_i, inverta-se o procedimento apresentado no Passo 2, começando com a freqüência que figura no *topo direito* da tabela (7). Desta vez, cada valor vai ser multiplicado pela soma de todas as freqüências que, com relação a ele, figuram *abaixo e à esquerda*. Fazendo a leitura da direita para a esquerda, observamos que as freqüências que estão à esquerda *e* abaixo de 7 são: 10, 10, 7 e 4. Como no passo anterior, repita-se este procedimento até esgotar *todas* as alternativas posicionais. Então, trabalhando da direita para a esquerda, temos:

Classe C/Adesão alta	7(10 + 10 + 7 + 4)	=	7(31)	=	217
Classe B/Adesão alta	8(10 + 4)	=	8(14)	=	112
Classe C/Adesão moderada	8(7 + 4)	=	8(11)	=	88
Classe B/Adesão moderada	10(4)	=	40		

(Note-se que nenhuma outra casela da tabela admite cálculo semelhantes aos que fizemos acima, uma vez que para as freqüências 15 (1.ª linha), 10 (2.ª linha), 11, 7 e 4 (última linha) as posições simultâneas "abaixo e à direita" não têm sentido.)

Σf_i é a soma dos produtos que acabamos de obter. Então:

$$\begin{aligned}\Sigma f_i &= 217 + 112 + 88 + 40 \\ &= 457\end{aligned}$$

PASSO 4: Entrar com os resultados dos Passos 2 e 3 na fórmula do coeficiente gama.

$$G = \frac{\Sigma f_a - \Sigma f_i}{\Sigma f_a + \Sigma f_i} = \frac{992 - 457}{992 + 457} = \frac{535}{1.449} = +0{,}37$$

Um $G = +0{,}37$ indica *correlação positiva, moderadamente fraca*, entre "classe social" e "adesão voluntária". O resultado que acabamos de encontrar sugere uma *dominância de acordos*: estes excedem as inversões em 37% (Note-se que se o gama tivesse sido $-0{,}37$, o significado concreto desse coeficiente seria: *correlação negativa, moderadamente fraca e baseada em inversões.*)

Prova de Significância Para o Gama

Para testar a hipótese nula de que X e Y não estão associados na população, basta transformar o G calculado (observado) em escore z, mediante a aplicação da seguinte fórmula:

$$z = G\sqrt{\frac{\Sigma f_a - \Sigma f_i}{N(1 - G^2)}}$$

onde

G = coeficiente gama calculado (observado)
f_a = freqüência de acordos
f_i = freqüência de inversões.

Na ilustração acima, encontramos $G = 0{,}37$ (como indicador da correlação existente entre "classe social" e "adesão voluntária"). Para testarmos a significância desse valor, entramos na fórmula do z conforme se segue:

$$\begin{aligned} z &= (0{,}37)\sqrt{\frac{992 - 457}{80(1 - 0{,}37^2)}} \\ &= (0{,}37)\sqrt{\frac{535}{80(0{,}86)}} = (0{,}37)\sqrt{\frac{535}{68{,}80}} \\ &= (0{,}37)\sqrt{7{,}78} = (0{,}37)(2{,}79) = 1{,}03 \end{aligned}$$

Consultando a Tabela B, no fim do livro, verificamos que o z deve igualar-se a ou exceder 1,96, a fim de que seja possível rejeitar a hipótese nula a 5%. Como o nosso z calculado ($z = 1{,}03$) é menor que o z crítico (1,96), devemos aceitar a hipótese nula de que $G = 0$ e rejeitar a hipótese experimental de que $G \neq 0$. A correlação que obtivemos não pode ser generalizada para a população da qual a amostra foi extraída.

Requisitos Para o Uso de Gama

Os seguintes requisitos devem ser levados em conta a fim de que seja possível o emprego do coeficiente gama como medida de associação:

1. *Correlação linear*–O coeficiente gama (G) só detecta relações lineares entre X e Y.
2. *Dados ordinais*–As variáveis X e Y devem ser passíveis de atribuição de postos ou de ordenação.
3. *Amostragem casual*–Para que seja possível testar a hipótese nula de que $G = 0$, é preciso que os sujeitos amostrais tenham sido extraídos aleatoriamente de uma dada população.

COEFICIENTE DE CORRELAÇÃO DE DADOS NOMINAIS DISPOSTOS NUMA TABELA 2 × 2

No capítulo anterior, começamos a estudar um teste de significância (para pôr freqüências à prova) conhecido por qui-quadrado. Como simples extensão do teste de qui-quadrado, podemos agora estabelecer o grau de associação entre variáveis que tenham sido mensuradas a nível nominal.

Vamos examinar, mais uma vez, a seguinte hipótese nula:

*A proporção de fumantes de maconha entre alunos orientados para a universidade é igual à dos não-orientados para a universidade.**

No Capítulo 10, essa hipótese nula foi testada numa amostra composta de 21 sujeitos orientados para a universidade e de 15 não-orientados. A pesquisa revelou que 15 dos 21 orientados para a universidade e 5 dos 15 não-orientados *eram* fumantes de maconha (veja Capítulo 10). Eis esse problema 2 × 2, de modo condensado, na Tabela 11.4.

A relação entre "ser fumante de maconha" (X) e "orientação para estudos superiores" (Y) foi testada mediante uma prova de qui-quadrado, conforme se segue:

$$\chi^2 = \frac{36[(15)(10) - (5)(6)]^2}{(15 + 5)(6 + 10)(15 + 6)(5 + 10)} = \frac{36(150 - 30)^2}{(20)(16)(21)(15)} = 5{,}14$$

* N.T.: Orientado para a universidade = que tenciona prosseguir os estudos após o 2.º grau.

TABELA 11.4 Uso de Maconha entre Estudantes Orientados e Não-Orientados para a Universidade. Dados Tirados da Tabela 10.3

	Fumantes	*Não-fumantes*	
Orientados para a Universidade	15	6	21
Não-orientados para a Universidade	5	10	15
	20	16	$N = 36$

Uma vez conseguido o valor do qui-quadrado observado ($\chi_o^2 = 5{,}14$), podemos calcular o *coeficiente fi* (ϕ), que nada mais é senão uma medida capaz de calcular o grau de associação em tabelas 2 X 2. Em símbolos:

$$\phi = \sqrt{\frac{\chi^2}{N}}$$

onde

ϕ = coeficiente fi
χ^2 = qui-quadrado observado (calculado)
N = tamanho da amostra (isto é, número de sujeitos experimentais).

Aplicando a fórmula acima ao problema em foco, resulta que:

$$\phi = \sqrt{\frac{5{,}14}{36}} = \sqrt{0{,}14} = 0{,}37$$

O coeficiente fi obtido ($\phi = 0{,}37$) indica a existência de correlação moderada entre orientação para a universidade e uso de maconha.

Teste de Significância para o Coeficiente Fi
Sem nenhuma dificuldade, felizmente, o coeficiente fi pode ser testado por meio de qui-quadrado cujos valores críticos estão registrados, em grande número, na Tabela E (fim do livro), aos níveis de 0,05 e 0,01. Assim:

$\chi_o^2 = 5{,}14$
$\chi_c^2 = 3{,}84$
$gl = 1$
$P = 0$

Como o χ_o^2 é maior que o χ_c^2, rejeitamos a hipótese nula de que $\phi = 0$ e aceitamos a hipótese experimental de que as variáveis X e Y estão, *de fato*, associadas na população da qual se extraiu a amostra de 36 estudantes.

Requisitos Para o Uso do Coeficiente Fi
O emprego do coeficiente fi, como medida de associação entre duas variáveis—X e Y—leva em conta os seguintes requisitos:

1. *Dados nominais*—Só se trabalha com freqüências.
2. *Tabela 2 X 2*—É fundamental que os dados possam ser dispostos numa tabela 2 X 2 (2 linhas por 2 colunas). Constitui erro empregar o coeficiente fi em tabelas de ordem superior a 2 X 2 (por exemplo: 3 X 3; 4 X 3 etc.).
3. *Amostragem casual*—A fim de que seja possível testar a significância dos coeficientes fi, é importante que a escolha dos elementos amostrais tenha sido aleatória.

COEFICIENTES DE CORRELAÇÃO DE DADOS NOMINAIS DISPOSTOS EM TABELAS DE ORDEM SUPERIOR A 2 X 2

Até este ponto, ocupamo-nos apenas do coeficiente de correlação de dados nominais dispostos em tabelas do tipo 2 X 2. Como vimos no Capítulo 10, há ocasiões em que, apesar de a variável observacional ser nominal, comparam-se as freqüências de vários grupos ou categorias. Vamos ilustrar examinando a seguinte hipótese nula:

A freqüência relativa com que aparecem os métodos de educar chamados "permissivos", "moderados" e "autoritários" é a mesma para "liberais", "moderados" e "conservadores".

Essa hipótese foi testada no Capítulo 10 com os dados dispostos numa tabela 3 X 3. A Tabela 11.5 reproduz esses dados.

A relação entre X e Y foi testada (no Capítulo 10) por qui-quadrado, de acordo com o que se segue:

$$\chi^2 = \frac{(7-10{,}79)^2}{10{,}79} + \frac{(10-10{,}07)^2}{10{,}07} + \frac{(15-11{,}14)^2}{11{,}14} + $$
$$+ \frac{(9-10{,}11)^2}{10{,}11} + \frac{(10-9{,}44)^2}{9{,}44} + \frac{(11-10{,}45)^2}{10.45} + $$
$$+ \frac{(14-9{,}10)^2}{9.10} + \frac{(8-8{,}49)^2}{8{,}49} + \frac{(5-9{,}40)^2}{9{,}40} = $$
$$= 7{,}58$$

TABELA 11.5 Relação entre "Métodos de Educar Filhos" e "Orientação Política". Dados Tirados da Tabela 10.4

Y \ X	Conservadores	Moderados	Liberais	
Permissivos	7	9	14	30
Moderados	10	10	8	28
Autoritários	15	11	5	31
	32	30	27	$N = 89$

No caso em foco, o que buscamos é determiunar a correlação ou o grau de associação entre "orientação política" (X) e "métodos de educar filhos" (Y). Numa tabela de ordem superior a 2 X 2, essa correlação pode ser conseguida mediante uma simples "ampliação" do teste de qui-quadrado, donde resulta o chamado *coeficiente de contingência*, notado C. O valor de C pode ser obtido pela aplicação da seguinte fórmula:

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}}$$

onde

χ^2 = qui-quadrado observado
N = tamanho da amostra
C = coeficiente de contingência.

Com os dados deste problema, resulta, aplicando a fórmula de C, o seguinte:

$$C = \sqrt{\frac{7{,}58}{89 + 7{,}58}} = \sqrt{\frac{7{,}58}{96{,}58}} = \sqrt{0{,}08} = 0{,}28$$

O nosso resultado, isto é, $C = 0{,}28$, indica que a correlação entre "orientação política" e "métodos de educar filhos" é bastante fraca. Essas duas variáveis *podem* estar relacionadas, o que não impede que surjam inúmeras exceções.

Teste de Significância Para o Coeficiente de Contingência

Tal como procedemos no caso do coeficiente fi, a determinação da significância do coeficiente de contingência faz-se mediante consulta direta à Tabela E, no fim do livro. Relativamente à situação que acabamos de examinar acima temos que:

χ_o^2 = 7,58
χ_c^2 = 9,49
gl = 4
P = 0,05

Diante desses resultados ($\chi_o^2 < \chi_c^2$), somos levados a não rejeitar a hipótese nula de que $C = 0$, descartando, por outro lado, a hipótese alternativa (experimental) de que $C \neq 0$. Em outras palavras, a relação "mensurada" exprime uma *circunstância* da particular *amostra* utilizada e não da *população* que lhe deu origem.

Requisitos Para o Uso do Coeficiente de Contingência

A fim de aplicar convenientemente o coeficiente de contingência, o leitor deve ficar atento para os seguintes requisitos:

1. *Dados nominais*–No cálculo de C só se exigem freqüências. Tais freqüências podem ser dispostas numa tabela 2 X 2 ou de ordem superior.
2. *Amostragem casual*–Para tornar factível a prova de significância do C (coeficiente de contingência), todos os elementos amostrais devem ser selecionados aleatoriamente de uma dada população.

Uma Alternativa Para o Coeficiente de Contingência: V de Cramér

Apesar da grande popularidade que o coeficiente de contingência tem entre os pesquisadores, ele apresenta uma notória desvantagem: *o número de linhas e de colunas numa tabela de qui-quadrado exerce influência sobre o valor máximo de C.* Em outras palavras, o valor do coeficiente de contingência *nem sempre* vai variar de 0 a 1, muito embora, em *nenhum* caso, exceda 1. Sob certas condições, o valor máximo de C poderá ser 0,94; sob outras, 0,89 e assim por diante.

Para evitar-se essa desvantagem de C, podemos decidir pelo emprego de um coeficiente de correlação alternativo, cuja finalidade é exprimir o grau de associação entre variáveis nominais em tabelas de ordem superior a 2 X 2. Conhecido por V de Cramér, esse coeficiente *não* depende do tamanho (ordem) da tabela de χ^2 e apresenta os mesmos requisitos de C (coeficiente de contingência). Em símbolos:

$$V = \sqrt{\frac{\chi^2}{N(k-1)}}$$

onde

V = V de Cramér
N = tamanho da amostra

k = número de linhas *ou* de colunas. (Usar o número *menor*, seja ele qual for. Nas tabelas em que o número de linhas é igual ao número de colunas, tanto faz usar um como outro–por exemplo, numa tabela 3 X 3, $k = 3$; numa tabela 4 X 4, $k = 4$–mas no caso de tabelas tal como uma 3 X 5, $k = 3$ *obrigatoriamente.*)

Voltando ao exemplo da Tabela 11.5, em que se examinou a relação entre "orientação política" (X) e "métodos de educar filhos" (Y) (tabela 3 X 3), vem que:

$$V = \sqrt{\frac{7{,}58}{89(3-1)}} = \sqrt{\frac{7{,}58}{89(2)}} = \sqrt{\frac{7{,}58}{178}} = \sqrt{0{,}04} = 0{,}20$$

Conclusão: O V de Cramér, igual a 0,20 neste exemplo, indica que existe uma relação fraca entre as variáveis X e Y.

RESUMO

Neste capítulo, entramos em contato com coeficientes de correlação que expressam numericamente o grau de associação entre as variáveis X e Y. Com o auxílio do coeficiente de correlação de Pearson (r), podemos determinar tanto a força quanto o sentido da relação entre variáveis mensuráveis, no mínimo, ao nível intervalar. Podemos também usar o r de Pearson para prever valores de uma variável (Y) a partir do conhecimento dos valores da outra variável (X).

Existem inúmeras alternativas não-paramétricas para o r de Pearson. A fim de determinar a correlação entre variáveis mensuradas ao nível ordinal, podemos aplicar o coeficiente de correlação de postos de Spearman (r_s). O uso dessa medida de correlação implica que as variáveis X e Y tenham sido ordenadas ou, então, que a elas tenham sido atribuídos postos. Quando ocorrer um número muito grande de postos empatados (espelhados), o coeficiente gama de Goodman e Kruskal (G) constitui uma boa alternativa para o coeficiente de correlação de postos.

Mediante uma simples extensão do teste de significância de qui-quadrado, podemos determinar o grau de associação entre variáveis mensuradas ao nível nominal. Para um problema 2 X 2, empregamos o coeficiente fi (ϕ); para um problema de ordem superior a 2 X 2, podemos usar ou o coeficiente de contingência (C) ou o V de Cramér.

PROBLEMAS

1. Os seis estudantes abaixo responderam a perguntas relativas às suas atitudes com relação a judeus (X) e a porto-riquenhos (Y). Calcule o coeficiente de correlação de Pearson para esses dados e verifique se a correlação é significante.

Alunos	X	Y
A	1	2
B	6	5
C	4	3
D	3	3
E	2	1
F	7	4

2. Calcule o coeficiente de correlação de Pearson para os seguintes conjuntos de escores e verifique se a correlação é significante.

X	Y
2	5
1	4
5	3
4	1

3. Calcule o coeficiente de correlação de Pearson para o seguinte conjunto de escores e verifique se a correlação é significante.

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

4. Calcule o coeficiente de correlação de Pearson para o seguinte conjunto de dados e indique se a correlação é significante.

X	Y
2	1
5	5
1	2
6	8
4	4

5. Calcule o coeficiente de correlação de Pearson para o seguinte conjunto de escores e verifique se a correlação é significante.

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

6. Utilizando-se dos dados do Problema 1, estabeleça a equação de regressão que permite prever o valor de Y (atitude com respeito a porto-riquenhos) a partir dos seguintes valores de X (atitude para com judeus): (a) $X = 5$; (b) $X = 2$; (c) $X = 9$.

7. Utilizando-se dos dados do Problema 5, estabeleça a equação de regressão que permite estimar o valor de Y para os seguintes valores de X: (a) $X = 10$; (b) $X = 2$.

8. Os cinco estudantes abaixo foram ordenados segundo o tempo que levaram para completar uma prova (Y). Por exemplo, quem terminou em primeiro lugar recebeu 1, quem terminou em segundo, 2 e assim por diante. A seguir, as provas foram avaliadas pelo professor, que atribuiu a esses alunos as notas (X) constantes do quadro abaixo. Teste a hipótese nula de que não existe relação entre X (notas) e Y (tempos de prova), isto é, calcule o r_s (coeficiente de correlação de postos de Spearman) e verifique se ele é ou não significante.

X	Y
53	1
91	2
70	3
85	4
91	5

9. Os oito sujeitos abaixo receberam postos em X e notas em Y. Calcule, para esses dados, o coeficiente de correlação de postos e verifique se existe uma relação significante entre essas variáveis.

X	Y
1	32
2	28
3	45
4	60
5	45
6	60
7	53
8	55

10. Sete sujeitos receberam postos tanto na variável X quanto na Y. Para esses dados, calcule o coeficiente de correlação de postos de Spearman e verifique se existe relação significante entre essas variáveis.

X	Y
1	7
3	6
2	5
4	3
5	4
7	2
6	1

11. Os cinco sujeitos abaixo foram "graduados" de 1 a 5 tanto na variável X quanto na Y. Tomando esses dados por base, calcule o coeficiente de correlação de postos e indique se entre X e Y existe relação significante.

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

12. Cinco sujeitos foram avaliados relativamente a duas variáveis: X e Y. Em ambas receberam valores inteiros, de 1 a 5, que podem ser entendidos como postos. Calcule o coeficiente gama e verifique se existe relação significante entre X e Y.

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

13. Um pesquisador distribuiu postos em ordem decrescente a 106 estudantes relativamente às duas variáveis seguintes: consumo de bebidas alcoólicas (X) e consumo diário de drogas (Y). Em ambas as variáveis, o pesquisador dispunha de três alternativas: alto, moderado e baixo. Concluído o estudo, resultaram os dados freqüenciais constantes da tabela abaixo. Determine, pois, com base neles, o grau em que as variáveis X e Y estão associadas, e teste se a relação encontrada é significante. Sugestão: determinar o coeficiente gama.

	Consumo de Álcool		
	Alto	*Moderado*	*Baixo*
Consumo de Drogas	*f*	*f*	*f*
Alto (*f*)	5	7	20
Moderado (*f*)	10	8	15
Baixo (*f*)	15	6	10
		$N = 106$	

14. No Problema 2 do Capítulo 10, o $\chi^2 = 8{,}29$ para a relação entre "comparecimento às aulas" e "notas no exame final de Estatística". Suponha que $N = 58$ e calcule o coeficiente fi, com vistas a determinar o grau de associação existente entre essas variáveis.

15. Dado um problema 2 X 2, onde $N = 138$ e $\chi^2 = 4,02$, calcule o coeficiente fi, a fim de determinar o grau de associação entre as variáveis X e Y.

16. Dado um problema 2 X 2, onde $N = 150$ e $\chi^2 = 3,90$, calcule o coeficiente fi, a fim de estabelecer o grau em que as variáveis X e Y estão associadas.

17. Tendo em mente estabelecer o grau de associação existente entre X e Y num problema 4 X 3, com $N = 100$ e $\chi^2 = 8,05$, calcule: (a) o coeficiente de contingência; (b) o V de Cramér.

18. No Problema 5 do Capítulo 10, $N = 118$ e $\chi^2 = 17,75$. Qual o grau de associação existente entre X e Y nesse problema 4 X 2? Calcule: (a) coeficiente de contingência; (b) V de Cramér.

19. A fim de estabelecer o grau de associação entre X e Y num problema 3 X 3, onde $N = 138$ e $\chi^2 = 10,04$, calcule: (a) o coeficiente de contingência; (b) o V de Cramér.

12 Aplicação de Procedimentos Estatísticos a Problemas de Pesquisa

A terceira parte deste livro contém certo número de procedimentos estatísticos que podem ser aplicados a vários problemas de pesquisa na área de ciências humanas. Os Capítulos 8, 9 e 10 apresentaram alguns procedimentos distintos para determinar se diferenças amostrais (calculadas) eram estatisticamente significativas ou se, ao contrário, resultavam de meros erros de amostragem. Os procedimentos apresentados no Capítulo 11 permitem determinar o grau de associação, a correlação entre duas variáveis.

Cada procedimento estatístico, conforme foi salientado ao longo deste livro, possui um conjunto de pressupostos que devem ser satisfeitos para garantir-se sua aplicação adequada. Ao selecionar procedimentos, o pesquisador deve, portanto, considerar certo número de fatores, tais como:

1. O que procura testar? Diferenças significativas, grau de associação ou ambas as coisas?
2. Relativamente à variável estudada, que nível de mensuração foi conseguido: nominal, ordinal, intervalar?
3. As variáveis estudadas têm ou não distribuição normal na população da qual elas foram extraídas?
4. A pesquisa implica o estudo de amostras independentes ou o estudo da mesma amostra mensurada mais de uma vez?

O Capítulo 12 apresenta algumas situações hipotéticas de pesquisa, onde os critérios acima estão especificados. Sugere-se ao leitor que escolha o procedimento estatístico que lhe pareça mais adequado a cada situação, dentre os constantes da seguinte lista (todos esses procedimentos foram estudados na Parte III do livro):

1. razão (estatística) t
2. análise de variância
3. qui-quadrado
4. teste da mediana
5. análise da variância de Kruskal-Wallis

6. dupla análise da variância de Friedman
7. r de Pearson
8. correlação de postos de Spearman
9. gama de Goodman e Kruskal
10. coeficiente fi
11. coeficiente de contingência
12. V de Cramér

A Tabela 12.1 (pág. 252) situa cada procedimento estatístico em função de alguns pressupostos importantes que devem ser levados em conta para garantir-se sua correta aplicação. Ao examinar as colunas da tabela, deparamo-nos com a primeira decisão vital, relacionada com a seleção do procedimento estatístico: *É nosso intento verificar se existe alguma relação entre as variáveis?* As provas de significância discutidas nos Capítulos 8, 9 e 10 permitem decidir se uma diferença amostral calculada reflete uma *real* diferença na população. Ou, será que, em lugar disso, o que buscamos é medir a força da relação entre duas variáveis? Este é um problema de correlação que pode ser solucionado por meio dos procedimentos estatísticos apresentados no Capítulo 11. Os subtítulos das colunas da Tabela 12.1 alertam o pesquisador que decide empregar uma prova de significância em vez de um procedimento correlacional para um aspecto deveras importante: ele deve certificar-se de que está diante de *amostras independentes* ou, então, de que seus dados resultam de *várias mensurações da mesma amostra.*

As linhas da Tabela 12.1 dirigem nossa atenção para o nível de mensuração em que nossas variáveis foram avaliadas. Se tivermos conseguido mensurar nossas variáveis no nível intervalar, nada nos impede de considerar o emprego de um procedimento paramétrico como, por exemplo, t, F ou r. Se, entretanto, tivermos obtido nível nominal ou ordinal de mensuração, nossa escolha fica limitada a algumas alternativas não-paramétricas.

As soluções para as situações de pesquisa apresentadas a seguir podem ser encontradas no fim deste capítulo.

SITUAÇÕES DE PESQUISA

Situação de Pesquisa 1
Um pesquisador realizou um experimento com vistas a determinar se a idade do conferencista exercia alguma influência na disposição de um grupo de estudantes para ouvi-lo. Numa situação normal de sala de aula, foi dito a 20 alunos que a administração da escola desejava conhecer suas preferências com relação a uma série de conferências previstas. Especificamente, pediu-se-lhes que dessem sua opinião sobre o professor que "poderia ser convidado". O professor foi descrito a *todos* os alunos da *mesma* forma *exceto* num aspecto: à metade dos alunos foi dito que ele (o conferencista) tinha 65 anos, enquanto que à outra metade a informação foi de que ele tinha 25 anos. Solicitou-se, em seguida, que manifestassem sua disposição para comparecer às conferências. Obtiveram-se os seguintes resultados (onde escores mais altos indicam maior disposição):

X_1 *Escores dos Estudantes Para os Quais o Professor Tinha 25 Anos*	X_2 *Escores dos Estudantes Para os Quais o Professor Tinha 65 Anos*
65	78
38	42
52	77
71	50
69	65
72	70
55	55
78	51
56	33
80	59

Que procedimento estatístico aplicaria o leitor para determinar se existe diferença significativa entre esses grupos de alunos relativamente à sua disposição para assistir às conferências?

Situação de Pesquisa 2
Um pesquisador, interessado em saber se a idade de um conferencista era capaz de influenciar a disposição de um grupo de universitários para ouvi-lo, decidiu realizar o seguinte experimento. Numa situação normal de sala de aula, foi dito a 30 alunos que a administração da escola desejava conhecer suas preferências com respeito a uma série de conferências previstas. Especificamente, pediu-se-lhes que dessem sua opinião quanto ao professor que "poderia ser convidado". O professor foi descrito a *todos* os alunos da *mesma* forma *exceto* num particular: a um terço dos alunos foi dito que o conferencista tinha 75 anos; a outro terço que ele tinha 50 anos; ao último terço, que se tratava de um jovem de 25 anos. Pediu-se, a seguir, que manifestassem sua disposição para comparecer às conferências. Obtiveram-se os seguintes resultados (onde escores mais altos indicam maior disposição):

Escores dos Alunos aos Quais foi Dito que o Professor Tinha		
25 Anos X_1	*50 Anos* X_2	*75 Anos* X_3
65	63	67
38	42	42
52	60	77
71	55	32
69	43	52
72	36	34
55	69	45
78	57	38
56	67	39
80	79	46

TABELA 12.1 Critério para a Escolha de um Procedimento Estatístico Adequado

	Testes de Significância (Capítulos 8, 9, 10)		*Correlação (Capítulo 11)*
Nível de Mensuração	*Amostras Independentes*	*Mesma Amostra Medida Duas Vezes*	
Nominal	Qui-quadrado (não-paramétrico) para testar a independência de duas ou mais variáveis)		Coeficiente fi (não-paramétrico para tabelas 2 X 2) Contingência e V de Cramér (não-paramétricos para tabelas maiores que 2 X 2)
Ordinal	Teste da mediana (não-paramétrico para "comparar" duas amostras) Análise de variância de Kruskal-Wallis (não-paramétrico para "comparar" três ou mais amostras)	Dupla análise da variância de Friedman (não-paramétrico para "comparar" a mesma amostra mensurada no mínimo duas vezes)	Correlação de postos de Spearman (não-paramétrico) Gama de Goodman e Kruskal (não-paramétrico para lidar com um número grande de postos "espelhados")
Intervalar	Razão *t* (paramétrico para "comparar" duas amostras) Análise de variância (paramétrico para "comparar" três ou mais amostras)	Razão *t* (paramétrico para "comparar" a mesma amostra medida duas vezes)	r de Pearson (paramétrico)

Que procedimento estatístico aplicaria o leitor para determinar se há diferença significativa entre esses grupos de estudantes relativamente à sua disposição para assistir às conferências?

Situação de Pesquisa 3

Para estudar a relação entre soletração e habilidade para ler, um pesquisador aplicou a um grupo de 20 alunos, selecionados aleatoriamente numa população de graduandos, testes específicos para mensurar essas variáveis. Obtiveram-se os seguintes resultados (onde escores mais altos indicam maior habilidade):

Aluno	*X* *Soletração*	*Y* *Leitura*
A	52	56
B	90	81
C	63	75
D	81	72
E	93	50

Aluno	*X* *Soletração*	*Y* *Leitura*
F	51	45
G	48	39
H	99	87
I	85	59
J	57	56
K	60	69
L	77	78
M	96	69
N	62	57
O	28	35
P	43	47
Q	88	73
R	72	76
S	75	63
T	69	79

Qual o procedimento estatístico que o leitor aplicaria para determinar o grau de associação entre as variáveis "soletração" (*X*) e "habilidade para leitura" (*Y*)?

Situação de Pesquisa 4

Para pesquisar a validade de um particular teste de leitura, um grupo de pesquisadores aplicou o referido teste a uma mostra de 20 alunos cuja habilidade para ler havia sido previamente avaliada pelo professor da classe. Os escores, bem como as avaliações individuais feitas pelo professor, figuram, relativamente a cada aluno, na lista abaixo:

Aluno	*X* *Escore-leitura*	*Y* *Avaliação pelo Professor*
A	28	18
B	50	17
C	92	1
D	85	6
E	76	5
F	69	10
G	42	11
H	53	12
I	80	3
J	91	2
K	73	4
L	74	9
M	14	20
N	29	19
O	86	7
P	73	8
Q	39	16
R	80	13
S	91	15
T	72	14

Qual procedimento estatístico adotaria o leitor para determinar o grau de associação entre os escores-leitura e as avaliações feitas pelo professor?

Situação de Pesquisa 5

Para pesquisar diferenças regionais relacionadas com "auxílio a estranhos", um pesquisador espalhou 400 chaves nas proximidades de caixas de correios localizadas no nordeste, sul, centro-oeste e oeste dos Estados Unidos da América do Norte. Cada uma das chaves estava etiquetada com o nome e endereço do respectivo "perdedor" (dono). O número de chaves devolvidas de cada região (como medida indicativa de "disposição para ajudar") figura na tabela abaixo:

	Região			
	Nordeste	*Sul*	*Centro-oeste*	*Oeste*
	f	*f*	*f*	*f*
Devolvidas	55	69	82	61
Não-devolvidas	45	31	18	39
	100	100	100	100

Que procedimento estatístico aplicaria o leitor para determinar se essas diferenças regionais são estatisticamente significativas?

Situação de Pesquisa 6

Para examinar a relação entre "autoritarismo" e "preconceito", um pesquisador colheu medidas de autoritarismo (com um instrumento especial) e de preconceito (uma lista de adjetivos negativos atribuíveis a negros norte-americanos). A amostra pesquisada consistiu em 950 adultos norte-americanos sorteados aleatoriamente. Obtiveram-se os seguintes resultados: entre 500 respondentes autoritários, 350 eram "preconceituosos" e 150 "tolerantes"; entre 450 não-autoritários, 125 eram "preconceituosos" e 325 "tolerantes".

Que procedimento estatístico usaria o leitor para estudar o grau de associação entre as variáveis "autoritarismo" e "preconceito"?

Situação de Pesquisa 7

Para investigar a relação entre "série escolar" e desempenho acadêmico", pesquisadores estudaram os registros de 186 universitários selecionados aleatoriamente dentre a totalidade de alunos de certa universidade. Obtiveram-se os seguintes resultados:

	Série Escolar			
	1.ª	*2.ª*	*3.ª*	*4.ª*
	f	*f*	*f*	*f*
A – ou melhor	6	5	7	10
B – a B⁺	10	16	19	18
C – a C⁺	23	20	15	7
C ou pior	15	7	6	2
	54	48	47	37

(N = 186)

Que procedimento estatístico escolheria o leitor para determinar o grau de associação entre desempenho acadêmico e série escolar?

Situação de Pesquisa 8

Para investigar a influência da frustração sobre o preconceito, pediu-se a 10 sujeitos que atribuíssem adjetivos negativos, tais como preguiçoso, sujo e imoral, a membros de um grupo minoritário (uma medida de preconceito). Essa atribuição de adjetivos fez-se em dois momentos distintos: *antes* e *depois* de terem esses sujeitos experimentais sido submetidos a uma série de exames difíceis e demorados. Entendeu o pesquisador que os exames assim concebidos teriam condições de criar a situação frustrante desejada. Obtiveram-se os seguintes dados, onde escores mais altos representam maior preconceito:

	Escores Relacionados com Preconceito	
Sujeitos	*Antes do Exame* X_1	*Depois do Exame* X_2
A	22	26
B	39	45
C	25	24
D	40	43
E	36	36
F	27	29
G	44	47
H	31	30
I	52	52
J	48	59

Qual procedimento estatístico usaria o leitor para determinar se existe uma diferença estatisticamente significante entre os escores de preconceito em função da frustração?

Situação de Pesquisa 9

Para pesquisar-se a relação entre o *status* ocupacional real de um respondente e sua classe social subjetiva (isto é, a classe social que o próprio respondente se atribui), um pesquisador pediu a 677 sujeitos que indicassem suas ocupações, bem como a classe social a que eles pertenciam. Entre 190 respondentes de ocupações de *status* superior (liberais–técnicos–executivos), 56 identificaram-se como de "classe alta", 122 como de "classe média" e 12 como de "classe baixa"; entre 221 respondentes de ocupações de *status* médio (vendedores–burocratas–operários qualificados), 42 percebiam-se na "classe alta", 163 na "média" e 16 na "baixa"; entre 266 com ocupações de *status* baixo (trabalhadores semi e não-qualificados), 15 perceberam-se na classe "alta", 202 na "média" e 49 na "baixa".

Que procedimento estatístico aplicaria o leitor para determinar o grau de associação entre *status* ocupacional e classe social subjetiva?

Situação de Pesquisa 10

Para investigar a relação entre as variáveis "opção profissional" e "salário inicial" em graduandos", pesquisadores entrevistaram um grupo de universitários recém-formados. Só se preocuparam, os pesquisadores, com três opções ocupacionais–engenharia, profissões liberais e administração–com o fato de que o respondente estivesse na situação de "primeiro emprego". Os resultados, obtidos de entrevistas com 21 respondentes, figuram na tabela abaixo:

Salários Iniciais (em Cr$)

Engenharia	*Prof. Liberais*	*Administração*
189.000	126.000	189.000
221.400	171.000	162.000
252.000	180.000	144.000
171.000	198.000	167.400
162.000	153.000	189.000
153.000	135.000	180.000
135.000	126.000	126.000

Que procedimento estatístico usaria o leitor para determinar se existe uma diferença significativa entre esses grupos de respondentes com respeito aos seus salários iniciais?

Situação de Pesquisa 11

Para investigar a influência da variável "opção profissional" sobre a variável "salário inicial" em graduandos, pesquisadores entrevistaram um grupo de universitários recém-formados em profissões liberais ou administração. São os seguintes os resultados colhidos desses 16 respondentes:

Salários Iniciais (em Cr$)

Prof. Liberais	*Administração*
126.000	135.000
171.000	162.000
180.000	144.000
198.000	167.400
153.000	189.000
135.000	180.000
126.000	126.000
	144.000
	167.400

Que procedimento estatístico empregaria o leitor para determinar se existe uma diferença significativa entre os salários iniciais correspondentes a essas duas amplas categorias de trabalho e as respectivas opções profissionais?

Situação de Pesquisa 12

Um pesquisador interessado em saber se a idade de um conferencista era capaz de influenciar a disposição de um grupo de universitários para ouvi-lo, decidiu realizar o seguinte experimento. Numa situação normal de sala de aula, foi dito a 130 alunos que a administração da escola desejava conhecer suas preferências com respeito a uma série de conferências previstas. Especificamente, pediu-se-lhes que dessem sua opinião quanto ao professor que "poderia ser convidado". O tal professor foi então descrito a *todos* os alunos da *mesma* forma *exceto* num particular: à metade dos alunos foi dito que ele tinha 65 anos, enquanto que à outra metade, que ele tinha 25. Pediu-se, a seguir, que manifestassem sua disposição para comparecer às conferências. Obtiveram-se os seguintes resultados: relativamente aos que achavam que o professor tinha 65 anos, 22 manifestaram desejo de comparecer, enquanto que 43 expressaram má vontade; com relação aos que supunham o professor com 25 anos, 38 dispuseram-se a comparecer, enquanto que 27 mostraram-se pouco inclinados a assistir às suas conferências.

Que procedimento estatístico usaria o leitor para determinar se existe diferença significativa entre o comportamento de "adesão" desses dois grupos de estudantes, quando a suposta variável controladora da "disposição" é a idade do conferencista?

SOLUÇÕES DAS PESQUISAS PROPOSTAS

Solução da Situação de Pesquisa 1

(*Razão t ou Teste da Mediana*)

A situação de pesquisa 1 consiste em uma comparação entre os escores de duas amostras independentes de alunos. A razão t (Capítulo 8) é empregada quando se trata de fazer comparações entre duas médias (aritméticas) resultantes de dados mensurados ao nível intervalar. O teste da mediana (Capítulo 10) é uma alternativa não-paramétrica que pode ser aplicada sempre que suspeitarmos que os escores não têm distribuição normal na população ou, então, quando o nível intervalar de mensuração não foi atingido.

Solução da Situação de Pesquisa 2

(*Análise de Variância ou Prova de Kruskal-Wallis*)

A situação de pesquisa 2 representa uma comparação de escores de três amostras independentes de alunos. A razão F (análise de variância–Capítulo 9) é empregada em comparações entre três ou mais médias independentes–desde que o nível intervalar tenha sido conseguido. A prova de Kruskal-Wallis (Capítulo 10) é uma alternativa não-paramétrica que pode ser aplicada sempre que tenhamos boas razões para suspeitar (a) da distribuição não-normal dos escores na população e (b) do fato de o nível de mensuração em que foram colhidos os dados ser, no máximo, ordinal.

Solução da Situação de Pesquisa 3

(*r de Pearson*)

A situação de pesquisa 3 é um problema de correlação, uma vez que a solução consiste na determinação do grau de associação entre X (habilidade para soletrar) e Y (habilidade para ler). O r de Pearson (Capítulo 11) pode ser empregado para o estabelecimento da correlação linear entre as variáveis X e Y, desde que elas tenham sido–ambas–mensuradas, no mínimo, ao nível intervalar de mensuração. Se X e Y não tiverem distribuição normal na população, devemos considerar uma alternativa não-paramétrica, tal como o coeficiente de correlação de postos de Spearman (Capítulo 11).

Solução da Situação de Pesquisa 4
(Coeficiente de Correlação de Postos de Spearman)
A situação de pesquisa 4 é um problema de correlação, uma vez que requer o cálculo do grau de associação entre X (escores-leitura) e Y (avaliação da habilidade para leitura, feita pelo professor). O coeficiente de correlação de postos de Spearman (Capítulo 11) pode ser empregado na detecção de uma relação linear entre as variáveis X e Y–se o nível em que essas variáveis foram mensuradas tiver sido, no máximo, o ordinal. O r de Pearson não pode ser empregado, pois ele implica a mensuração de X e Y a nível intervalar, no mínimo. No caso em foco, aos escores-leitura (X) devem ser atribuídos postos de 1 a 20 *antes* da aplicação da fórmula de Spearman.

Solução da Situação de Pesquisa 5
(Qui-quadrado)
A situação de pesquisa 5 consiste em uma comparação entre as freqüências ("devolução" *versus* "não-devolução") observadas em quatro grupos regionalmente divididos (nordeste, sul, centro-oeste e oeste). A prova de significância chamada qui-quadrado (Capítulo 10) é usada em comparações entre freqüências oriundas de duas ou mais amostras. Para o teste de qui-quadrado não se requer mais do que nível nominal de mensuração. Os dados deste problema podem ser dispostos numa tabela 2 X 4, isto é, numa tabela de 2 linhas e 4 colunas. Observe-se que o grau de associação entre "taxa de devolução" (X) e "região" (Y) pode ser expresso pelo chamado coeficiente C de contingência ou pelo V de Cramér (Capítulo 11).

Solução da Situação de Pesquisa 6
(Coeficiente "fi")
A situação de pesquisa 6 é um problema de correlação, onde o grau de associação entre X (autoritarismo) e Y (preconceito) deve ser buscado. O coeficiente "fi" (Capítulo 11) é uma medida de associação passível de ser empregada quando as freqüências (ou os dados de nível nominal) podem ser dispostos numa tabela do tipo 2 X 2 (2 linhas e 2 colunas).

Neste problema, tal tabela assumiria a seguinte forma:

Nível de Preconceito	*Nível de Autoritarismo*		
	Autoritário	*Não-autoritário*	
Preconceituoso	350	120	
Tolerante	150	325	$N = 950$

Solução da Situação de Pesquisa 7
(Gama de Goodman e Kruskal)
A situação de pesquisa 7 é um problema de correlação que consiste na determinação do grau de associação entre X (escala conceitual) e Y (série escolar). O coeficiente gama de Goodman e Kruskal (Capítulo 11) emprega-se para o estabelecimento de uma correlação linear entre X e Y quando a ambas as variáveis foram atribuídos postos, com vários deles empatados. No problema em foco, a escala conceitual foi graduada de A a D ou pior e a série escolar, por seu turno, da 1.ª à 4.ª. Ambas as variáveis (X e Y) deram origem a "medidas" que, transformadas em postos, resultaram em numerosos empates (postos empatados ou "espelhados"). Por exemplo, 54 alunos cursavam a primeira série; 48 estavam na segunda, e assim por diante. O coeficiente de contingência (C) ou o V de Cramér (Capítulo 11) representam alternativas para o coeficiente gama, cujo cálculo pressupõe só dados de nível nominal.

Solução da Situação de Pesquisa 8
(Razão t ou Teste de Friedman)
A situação de pesquisa 8 corresponde a uma comparação do tipo "antes e depois", onde a mesma amostra foi mensurada duas vezes, porém em momentos diferentes. A razão t (Capítulo 8) pode ser empregada na comparação de duas médias resultantes da mesma amostra, desde que o modelo experimental admita mensurações em momentos distintos e que os dados resultantes dessas mensurações sejam, no mínimo, de nível intervalar. O teste de Friedman (Capítulo 10) é uma alternativa não-paramétrica (dupla análise de variância) que se ajusta bem a modelos experimentais do tipo "antes e depois", e pode ser usado sempre que haja razões para suspeitar de distribuição não-normal dos escores na população, ou quando o nível de mensuração dos dados não tenha alcançado o intervalar.

Solução da Situação de Pesquisa 9
(Gama de Goodman e Kruskal)
A situação de pesquisa 9 é um problema de correlação que implica o cálculo do grau de associação entre X (*status* ocupacional) e Y (classe sócio-econômica subjetiva). O coeficiente gama (Capítulo 11) ajusta-se particularmente bem ao problema de detectar a correlação linear entre duas variáveis sujeitas à atribuição de postos com a ocorrência de grande número de empates. Na situação presente, tanto o "*status* ocupacional" quanto a "classe sócio-econômica subjetiva" foram ordenados segundo o critério

Classe Sócio-Econômica Subjetiva (Y)	*Status Ocupacional* (X)		
	Alto *f*	*Médio* *f*	*Baixo* *f*
Alta	56	42	15
Média	122	163	202
Baixa	12	16	49
	190	221	266

"alto", "médio" e "baixo", gerando-se, assim, muitos postos empatados (por exemplo, 221 respondentes apresentaram ocupações de "*status* médio"). Para que se possa calcular o coeficiente gama, os dados devem ser dispostos numa tabela freqüencial, de dupla entrada, como exemplificado no final da página anterior.

O coeficiente de contingência (C) e o V de Cramér constituem alternativas para o coeficiente gama, que pressupõe só dados de nível nominal.

Solução da Situação de Pesquisa 10
(*Análise de Variância ou Prova de Kruskal-Wallis*)
A situação de pesquisa 10 consiste em uma comparação entre os escores de três amostras independentes de respondentes. A razão F (Capítulo 9) pode ser usada quando se trata de comparar três ou mais médias aritméticas independentes, desde que os dados tenham sido colhidos no nível intervalar (no mínimo). A análise de variância de Kruskal-Wallis (Capítulo 10) é uma alternativa não-paramétrica que pode ser empregada sempre que suspeitarmos de que (a) os escores não estejam normalmente distribuídos na população da qual provieram; (b) o nível intervalar de mensuração não tenha sido alcançado.

Solução da Situação de Pesquisa 11
(*Razão t ou Prova de Mediana*)
A situação de pesquisa 11 requer uma comparação entre os escores de duas amostras independentes de sujeitos. A razão t (Capítulo 8) é usada na comparação de duas médias (aritméticas) quando os dados foram colhidos no nível intervalar. A prova da mediana (Capítulo 10) é uma alternativa não-paramétrica que pode ser aplicada quando não tivermos razões para admitir normalidade de distribuição dos escores (na população) ou quando o nível intervalar de mensuração não tiver sido atingido.

Solução da Situação de Pesquisa 12
(*Qui-quadrado*)
A situação de pesquisa 12 consiste numa comparação entre as freqüências ("disposição" *versus* "má vontade") de dois grupos de estudantes (aqueles aos quais se disse que a idade do professor era 65 anos *versus* os que achavam que ele tinha 25). O teste (de significância) de qui-quadrado (Capítulo 10) permite comparar as freqüências resultantes de duas ou mais amostras, sempre que os dados tenham sido colhidos no nível nominal ou, então, resultem de contagem. Os dados dessa situação podem ser levados a uma tabela 2 X 2, isto é, de 2 linhas e 2 colunas. Assim:

Disposição para Comparecer	*Condição Experimental: Foi Dito aos Alunos Que o Professor Tinha*		
	65 Anos (f)	*25 Anos (f)*	
Sim	22	38	
Não	43	27	$N = 130$

APÊNDICES

Apêndice A

Revisão de Fundamentos de Matemática

Os estudantes de Estatística que necessitem de uma revisão de alguns tópicos fundamentais de álgebra e aritmética poderão encontrar, neste apêndice, problemas com números decimais, números negativos e raízes quadradas. Discutiram-se outros problemas de Matemática, no corpo do texto, à medida que foram surgindo. Por exemplo, o Capítulo 1 identifica, define e compara os três níveis de mensuração; o Capítulo 2 discute porcentagens, proporções, razões e taxas; o Capítulo 4 introduz a noção de somatório (Σ).

NÚMEROS DECIMAIS

Ao somar ou subtrair números decimais, tenha o cuidado de escrevê-los uns debaixo dos outros, fazendo coincidir a posição das vírgulas. Por exemplo, para somar 3.210,76, 2,541 e 98,3 procede-se assim:

$$\begin{array}{r} 3.210,76 \\ 2,541 \\ \underline{98,3} \\ 3.311,601 \end{array}$$

Para subtrair 34,1 de 876,62, o procedimento é análogo:

$$\begin{array}{r} 876,62 \\ \underline{-34,1} \\ 842,52 \end{array}$$

Ao multiplicar números decimais, cuide para que sua resposta tenha tantas casas decimais quantas correspondam à soma das existentes no multiplicando e no multiplicador. Exemplo:

Multiplicando ⟶	63,41	2,6	0,0003	0,5
Multiplicador ⟶	X 0,05	X 1,4	X 0,03	X 0,5
Produto ⟶	3,1705	3,64	0,000009	0,25

Elimine sempre, antes de dividir, as casas decimais do divisor. Isso se consegue caminhando com a vírgula para a direita tantas casas quantas sejam necessárias para tornar o divisor um número inteiro. Proceda de forma análoga com relação ao dividendo, movendo a vírgula para a direita tantas casas quantas tenham sido "saltadas" no divisor (isto é, se a vírgula caminhou duas casas no divisor, deve caminhar também duas casas no dividendo). Com esse recurso fica automaticamente determinado o número de casas decimais da resposta.* Exemplos:

Dividendo → 2,44; Divisor → 0,02

$$\frac{2,44}{0,02} = \underset{\substack{\uparrow \\ \text{Quociente}}}{122} \qquad \frac{2,44}{0,02} = \frac{244}{2} = 122$$

$$\frac{0,88}{0,4} = 2,2 \qquad \frac{0,88}{0,4} = \frac{8,8}{4} = 2,2^{**}$$

$$\frac{10,10}{10} = 1,01$$

$$\frac{1.010}{0.10} = 10.100$$

Operações aritméticas freqüentemente produzem respostas com notação decimal; por exemplo, 2,034; 24,7; 86,001 e assim por diante. A questão que se põe agora tem relação com o número de casas decimais que devemos adotar nas nossas respostas. Uma regra simples de seguir consiste em realizar todas as operações com três casas decimais adicionais e arredondar o resultado de forma que ele contenha duas casas mais do que as constantes dos dados originais.

Ilustrando: se os dados originarem-se de um conjunto de números inteiros (por exemplo, 12, 9, 49 ou 15), realizamos as operações com três casas decimais (até milésimos) e expressamos nossa resposta com duas decimais (isto é, em centésimos).*** Por exemplo:

* N.T.: O número de casas decimais da resposta vai ser determinado pelo número de casas decimais do *dividendo após* aplicação do procedimento.

** N.T.: Uma casa decimal na resposta porque 8,8 tem uma casa depois da vírgula.

*** N.T.: Essa regra "prática" contraria a orientação dada pelo próprio Autor, linhas atrás, sobre

3,889 = 3,89
1,224 = 1,22
7,761 = 7,76

O arredondamento para determinada casa decimal faz-se geralmente do seguinte modo: abandona-se o último dígito se ele for menor que 5:

Menor que 5

26,234 = 26,23
14,891 = 14,89
1,012 = 1,01

Entretanto, se o dígito abandonado (o último) for igual a ou maior que 5, adiciona-se uma unidade ao primeiro dígito que sobra (no sentido da direita para a esquerda:

5 ou mais*

26,236 = 26,24
14,899 = 14,90
1,015 = 1,02

Os seguintes números foram arredondados para o inteiro mais próximo:

3,1 = 3
3,5 = 4
4,5 = 5
4,8 = 5

Os seguintes números foram arredondados para o décimo mais próximo:

3,11 = 3,1
3,55 = 3,6
4,45 = 4,5
4,17 = 4,2

---

divisão de números aproximados. Trabalhar com números aproximados obedece a regras *rígidas* e não apenas *práticas*. Entretanto, sugere-se ao leitor uma regra *deveras* prática: esquecer o procedimento acima!

* N.T.: Melhor seria adotar a seguinte regra: se for *mais de 5*, "forçar" o dígito anterior de uma unidade; se for *exatamente 5*, *manter* o dígito anterior se ele for *par* e *"forçá-lo"* de uma unidade se for *ímpar*. Exemplos:

1,015 → 1,02
1,025 → 1,02
1,035 → 1,04

Os seguintes números foram arredondados para o centésimo mais próximo:

3,328 = 3,33
4,823 = 4,82
3,065 = 3,07
3,055 = 3,06

NÚMEROS NEGATIVOS

Ao somar um conjunto de números negativos, certifique-se de que o sinal menos (−) não foi esquecido no resultado. Exemplo:

$$\begin{array}{r} -20 \\ -12 \\ \underline{-6} \\ -38 \end{array} \qquad \begin{array}{r} -3 \\ -9 \\ \underline{-4} \\ -16 \end{array}$$

Para somar um conjunto de números dotados de sinal* (alguns com + , outros com −), agrupam-se primeiro todos os negativos e depois todos os positivos; daí decorrem dois resultados, um negativo (isto é, uma soma negativa) e um positivo (uma soma positiva); subtrai-se, a seguir, um resultado do outro, *mantendo o sinal do número de maior valor absoluto.*** Exemplos:

$$\begin{array}{r} -6 \\ +4 \\ +2 \\ -1 \\ \underline{-3} \\ -4 \end{array} \qquad \begin{array}{r} +4 \\ \underline{+2} \\ +6 \end{array} \qquad \begin{array}{r} -6 \\ -1 \\ \underline{-3} \\ -10 \end{array} \qquad \begin{array}{r} +6 \\ \underline{-10} \\ -4 \end{array}$$

Para subtrair um número negativo, dê-lhe antes sinal + (mais) e siga o procedimento indicado para a adição. O resto recebe o sinal do número de maior valor absoluto. Exemplos:

$$\begin{array}{r} 24 \\ \underline{-(-6)} \\ 30 \end{array}$$

−6 recebe sinal + e é, por isso, adicionado ao 24. Como o número de maior valor absoluto é *positivo*, o resto (resultado) também vai ser positivo (30).

---

* N.T.: Números dotados de sinal chamam-se *números relativos*. Ex.: −5; +8; −0,2 etc. Um número relativo pode ser *fracionário* como também *irracional*: −0,8; $+\frac{1}{3}$; $-\sqrt{3}$; $+\sqrt{0,6}$ etc.

** N.T.: *Valor absoluto* de um número relativo é o próprio número desprovido de sinal. Então, o valor absoluto de −2 é 2 e o valor absoluto de +2 também é 2.

$$\begin{array}{r} -6 \\ -(-24) \\ \hline 18 \end{array}$$

–24 recebe sinal + e é, por isso, subtraído. Como o número de maior valor absoluto é *positivo*, o resto também vai ser positivo (18).

$$\begin{array}{r} -24 \\ -(-6) \\ \hline -(18) \end{array}$$

–6 recebe sinal + e é, por isso, subtraído. Como o número de maior valor absoluto é *negativo*, o resto também vai ser negativo (–18).

Obs.: Quando um número é positivo, ele dispensa o sinal. Então $+2 = 2$.

Ao multiplicar (ou dividir) dois números relativos que têm o mesmo sinal, dê sempre sinal + (mais) ao resultado. Exemplos:

$$(+8) \times (+5) = +40 \qquad \frac{+40}{+8} = +5$$

$$(-8) \times (-5) = +40 \qquad \frac{-40}{-8} = +5$$

No caso de os números terem sinais contrários, o resultado será sempre negativo. Exemplos:

$$(-8) \times (+5) = -40 \qquad \frac{+40}{-8} = -5$$

CÁLCULO DE RAÍZES QUADRADAS MEDIANTE O USO DA TABELA A

Com o auxílio da Tabela A (fim do livro), pode-se facilmente encontrar a raiz quadrada ($\sqrt{n}$) de qualquer número inteiro (n) de 1 a 1.000.

Para calcular a raiz quadrada aproximada de números decimais, bem como de números maiores que 1.000, pode ser útil começar com a coluna de quadrados (n^2), Tabela A. A raiz quadrada de qualquer número multiplicada por si mesma deve reproduzir o número original.* Como resultado, n, na Tabela A, é, de fato, a própria raiz quadrada de n^2.

Tendo em mente tirar o máximo de proveito da coluna n^2 em cálculos aproximados de raízes quadradas, devemos determinar o número de dígitos que precedem a vírgula na resposta. Uma regra prática, neste caso, é a seguinte: contar o número de dígitos que compõem a parte inteira (do número cuja raiz quadrada vai ser extraída) e dividi-lo por 2; o resultado corresponde ao número de dígitos da parte inteira da resposta.

* N.T.: Isso só é verdade se o número for quadrado perfeito. Caso contrário, jamais se consegue uma reprodução do número original. Exemplos:

a) $\sqrt{144} = 12$. Então $(12) \times (12) = 144$

b) $\sqrt{2} = 1{,}4142$ aprox. (Veja Tabela A). Então $(1{,}4142) \times (1{,}4142) = 1{,}9999616$

Exemplos:

$\sqrt{5555} = 74{,}53$ (4) dígitos $\div$ 2 = (2) dígitos

$\sqrt{55{,}55} = 7{,}45$ (2) dígitos $\div$ 2 = (1) dígito

Se o número de dígitos da parte inteira for um número ímpar, daí resultando um "quebrado" na divisão, adicione-se uma unidade a esse número e proceda-se conforme indicação acima. Exemplos:

$\sqrt{555{,}5} = 23{,}57$ (3) dígitos $\div$ 2 = 1,5 dígitos
Então, $3 + 1 = 4$ e $4 \div 2 =$ (2) dígitos

$\sqrt{5{,}555} = 2{,}36$ (1) dígito $\div$ 2 = 0,5 dígito
Então $1 + 1 = 2$ e $2 \div 2 =$ (1) dígito

Pode-se achar a raiz quadrada de qualquer número menor que 1 de acordo com o seguinte procedimento:

1. Arredondar o número para o centésimo mais próximo.

$$\sqrt{0{,}328} = \sqrt{0{,}33}$$
$$\sqrt{0{,}823} = \sqrt{0{,}82}$$
$$\sqrt{0{,}0651} = \sqrt{0{,}07}$$
$$\sqrt{0{,}035} = \sqrt{0{,}04}$$

2. Localizar (Tabela A) a raiz quadrada do número resultante *como se* fosse um inteiro. Para tanto, basta ignorar a vírgula.

$$\sqrt{33} = 5{,}74$$
$$\sqrt{82} = 9{,}06$$
$$\sqrt{7} = 2{,}65$$
$$\sqrt{4} = 0{,}26$$

3. Deslocar a vírgula uma casa para a esquerda e arredondar.

$$\sqrt{0{,}33} = 0{,}57$$
$$\sqrt{0{,}82} = 0{,}91$$
$$\sqrt{0{,}07} = 0{,}27$$
$$\sqrt{0{,}04} = 0{,}2$$

Apêndice B

TABELA A Quadrados, Raízes Quadradas e Inversos dos Números de 1 a 1.000*.

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
1	1	1.0000	1.000000	1.0000
2	4	1.4142	.500000	.7071
3	9	1.7321	.333333	.5774
4	16	2.0000	.250000	.5000
5	25	2.2361	.200000	.4472
6	36	2.4495	.166667	.4082
7	49	2.6458	.142857	.3780
8	64	2.8284	.125000	.3536
9	81	3.0000	.111111	.3333
10	100	3.1623	.100000	.3162
11	121	3.3166	.090909	.3015
12	144	3.4641	.083333	.2887
13	169	3.6056	.076923	.2774
14	196	3.7417	.071429	.2673
15	225	3.8730	.066667	.2582
16	256	4.0000	.062500	.2500
17	289	4.1231	.058824	.2425
18	324	4.2426	.055556	.2357
19	361	4.3589	.052632	.2294
20	400	4.4721	.050000	.2236
21	441	4.5826	.047619	.2182
22	484	4.6904	.045455	.2132
23	529	4.7958	.043478	.2085
24	576	4.8990	.041667	.2041
25	625	5.0000	.040000	.2000
26	676	5.0990	.038462	.1961
27	729	5.1962	.037037	.1925
28	784	5.2915	.035714	.1890
29	841	5.3852	.034483	.1857
30	900	5.4772	.033333	.1826
31	961	5.5678	.032258	.1796
32	1024	5.6569	.031250	.1768
33	1089	5.7446	.030303	.1741
34	1156	5.8310	.029412	.1715
35	1225	5.9161	.028571	.1690

* Tendo em mente preservar as tabelas da edição original, foram elas reproduzidas neste livro *fotograficamente*, o que implica os seguintes cuidados da parte do leitor:
a) em lugar de *ponto* leia-se *vírgula* e
b) aos números iniciados por ponto acrescente-se um *zero*.

TABELA A (continuação)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
36	1296	6.0000	.027778	.1667
37	1369	6.0828	.027027	.1644
38	1444	6.1644	.026316	.1622
39	1521	6.2450	.025641	.1601
40	1600	6.3246	.025000	.1581
41	1681	6.4031	.024390	.1562
42	1764	6.4807	.023810	.1543
43	1849	6.5574	.023256	.1525
44	1936	6.6332	.022727	.1508
45	2025	6.7082	.022222	.1491
46	2116	6.7823	.021739	.1474
47	2209	6.8557	.021277	.1459
48	2304	6.9282	.020833	.1443
49	2401	7.0000	.020408	.1429
50	2500	7.0711	.020000	.1414
51	2601	7.1414	.019608	.1400
52	2704	7.2111	.019231	.1387
53	2809	7.2801	.018868	.1374
54	2916	7.3485	.018519	.1361
55	3025	7.4162	.018182	.1348
56	3136	7.4833	.017857	.1336
57	3249	7.5498	.017544	.1325
58	3364	7.6158	.017241	.1313
59	3481	7.6811	.016949	.1302
60	3600	7.7460	.016667	.1291
61	3721	7.8102	.016393	.1280
62	3844	7.8740	.016129	.1270
63	3969	7.9373	.015873	.1260
64	4096	8.0000	.015625	.1250
65	4225	8.0623	.015385	.1240
66	4356	8.1240	.015152	.1231
67	4489	8.1854	.014925	.1222
68	4624	8.2462	.014706	.1213
69	4761	8.3066	.014493	.1204
70	4900	8.3666	.014286	.1195
71	5041	8.4261	.014085	.1187
72	5184	8.4853	.013889	.1179
73	5329	8.5440	.013699	.1170
74	5476	8.6023	.013514	.1162
75	5625	8.6603	.013333	.1155
76	5776	8.7178	.013158	.1147
77	5929	8.7750	.012987	.1140
78	6084	8.8318	.012821	.1132
79	6241	8.8882	.012658	.1125
80	6400	8.9443	.012500	.1118
81	6561	9.0000	.012346	.1111
82	6724	9.0554	.012195	.1104
83	6889	9.1104	.012048	.1098
84	7056	9.1652	.011905	.1091
85	7225	9.2195	.011765	.1085

TABELA A (continuação)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
86	7396	9.2736	.011628	.1078
87	7569	9.3274	.011494	.1072
88	7744	9.3808	.011364	.1066
89	7921	9.4340	.011236	.1060
90	8100	9.4868	.011111	.1054
91	8281	9.5394	.010989	.1048
92	8464	9.5917	.010870	.1043
93	8649	9.6437	.010753	.1037
94	8836	9.6954	.010638	.1031
95	9025	9.7468	.010526	.1026
96	9216	9.7980	.010417	.1021
97	9409	9.8489	.010309	.1015
98	9604	9.8995	.010204	.1010
99	9801	9.9499	.010101	.1005
100	10000	10.0000	.010000	.1000
101	10201	10.0499	.009901	.0995
102	10404	10.0995	.009804	.0990
103	10609	10.1489	.009709	.0985
104	10816	10.1980	.009615	.0981
105	11025	10.2470	.009524	.0976
106	11236	10.2956	.009434	.0971
107	11449	10.3441	.009346	.0967
108	11664	10.3923	.009259	.0962
109	11881	10.4403	.009174	.0958
110	12100	10.4881	.009091	.0953
111	12321	10.5357	.009009	.0949
112	12544	10.5830	.008929	.0945
113	12769	10.6301	.008850	.0941
114	12996	10.6771	.008772	.0937
115	13225	10.7238	.008696	.0933
116	13456	10.7703	.008621	.0928
117	13689	10.8167	.008547	.0925
118	13924	10.8628	.008475	.0921
119	14161	10.9087	.008403	.0917
120	14400	10.9545	.008333	.0913
121	14641	11.0000	.008264	.0909
122	14884	11.0454	.008197	.0905
123	15129	11.0905	.008130	.0902
124	15376	11.1355	.008065	.0898
125	15625	11.1803	.008000	.0894
126	15876	11.2250	.007937	.0891
127	16129	11.2694	.007874	.0887
128	16384	11.3137	.007813	.0884
129	16641	11.3578	.007752	.0880
130	16900	11.4018	.007692	.0877
131	17161	11.4455	.007634	.0874
132	17424	11.4891	.007576	.0870
133	17689	11.5326	.007519	.0867
134	17956	11.5758	.007463	.0864
135	18225	11.6190	.007407	.0861

TABELA A (continuação)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
136	18496	11.6619	.007353	.0857
137	18769	11.7047	.007299	.0854
138	19044	11.7473	.007246	.0851
139	19321	11.7898	.007194	.0848
140	19600	11.8322	.007143	.0845
141	19881	11.8743	.007092	.0842
142	20164	11.9164	.007042	.0839
143	20449	11.9583	.006993	.0836
144	20736	12.0000	.006944	.0833
145	21025	12.0416	.006897	.0830
146	21316	12.0830	.006849	.0828
147	21609	12.1244	.006803	.0825
148	21904	12.1655	.006757	.0822
149	22201	12.2066	.006711	.0819
150	22500	12.2474	.006667	.0816
151	22801	12.2882	.006623	.0814
152	23104	12.3288	.006579	.0811
153	23409	12.3693	.006536	.0808
154	23716	12.4097	.006494	.0806
155	24025	12.4499	.006452	.0803
156	24336	12.4900	.006410	.0801
157	24649	12.5300	.006369	.0798
158	24964	12.5698	.006329	.0796
159	25281	12.6095	.006289	.0793
160	25600	12.6491	.006250	.0791
161	25921	12.6886	.006211	.0788
162	26244	12.7279	.006173	.0786
163	26569	12.7671	.006135	.0783
164	26896	12.8062	.006098	.0781
165	27225	12.8452	.006061	.0778
166	27556	12.8841	.006024	.0776
167	27889	12.9228	.005988	.0774
168	28224	12.9615	.005952	.0772
169	28561	13.0000	.005917	.0769
170	28900	13.0384	.005882	.0767
171	29241	13.0767	.005848	.0765
172	29584	13.1149	.005814	.0762
173	29929	13.1529	.005780	.0760
174	30276	13.1909	.005747	.0758
175	30625	13.2288	.005714	.0756
176	30976	13.2665	.005682	.0754
177	31329	13.3041	.005650	.0752
178	31684	13.3417	.005618	.0750
179	32041	13.3791	.005587	.0747
180	32400	13.4164	.005556	.0745
181	32761	13.4536	.005525	.0743
182	33124	13.4907	.005495	.0741
183	33489	13.5277	.005464	.0739
184	33856	13.5647	.005435	.0737
185	34225	13.6015	.005405	.0735

TABELA A (continuação)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
186	34596	13.6382	.005376	.0733
187	34969	13.6748	.005348	.0731
188	35344	13.7113	.005319	.0729
189	35721	13.7477	.005291	.0727
190	36100	13.7840	.005263	.0725
191	36481	13.8203	.005236	.0724
192	36864	13.8564	.005208	.0722
193	37249	13.8924	.005181	.0720
194	37636	13.9284	.005155	.0718
195	38025	13.9642	.005128	.0716
196	38416	14.0000	.005102	.0714
197	38809	14.0357	.005076	.0712
198	39204	14.0712	.005051	.0711
199	39601	14.1067	.005025	.0709
200	40000	14.1421	.005000	.0707
201	40401	14.1774	.004975	.0705
202	40804	14.2127	.004950	.0704
203	41209	14.2478	.004926	.0702
204	41616	14.2829	.004902	.0700
205	42025	14.3178	.004878	.0698
206	42436	14.3527	.004854	.0697
207	42849	14.3875	.004831	.0695
208	43264	14.4222	.004808	.0693
209	43681	14.4568	.004785	.0692
210	44100	14.4914	.004762	.0690
211	44521	14.5258	.004739	.0688
212	44944	14.5602	.004717	.0687
213	45369	14.5945	.004695	.0685
214	45796	14.6287	.004673	.0684
215	46225	14.6629	.004651	.0682
216	46656	14.6969	.004630	.0680
217	47089	14.7309	.004608	.0679
218	47524	14.7648	.004587	.0677
219	47961	14.7986	.004566	.0676
220	48400	14.8324	.004545	.0674
221	48841	14.8661	.004525	.0673
222	49284	14.8997	.004505	.0671
223	49729	14.9332	.004484	.0670
224	50176	14.9666	.004464	.0668
225	50625	15.0000	.004444	.0667
226	51076	15.0333	.004425	.0665
227	51529	15.0665	.004405	.0664
228	51984	15.0997	.004386	.0662
229	52441	15.1327	.004367	.0661
230	52900	15.1658	.004348	.0659
231	53361	15.1987	.004329	.0658
232	53824	15.2315	.004310	.0657
233	54289	15.2643	.004292	.0655
234	54756	15.2971	.004274	.0654
235	55225	15.3297	.004255	.0652

TABELA A (continuação)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
236	55696	15.3623	.004237	.0651
237	56169	15.3948	.004219	.0650
238	56644	15.4272	.004202	.0648
239	57121	15.4596	.004184	.0647
240	57600	15.4919	.004167	.0645
241	58081	15.5242	.004149	.0644
242	58564	15.5563	.004132	.0643
243	59049	15.5885	.004115	.0642
244	59536	15.6205	.004098	.0640
245	60025	15.6525	.004082	.0639
246	60516	15.6844	.004065	.0638
247	61009	15.7162	.004049	.0636
248	61504	15.7480	.004032	.0635
249	62001	15.7797	.004016	.0634
250	62500	15.8114	.004000	.0632
251	63001	15.8430	.003984	.0631
252	63504	15.8745	.003968	.0630
253	64009	15.9060	.003953	.0629
254	64516	15.9374	.003937	.0627
255	65025	15.9687	.003922	.0626
256	65536	16.0000	.003906	.0625
257	66049	16.0312	.003891	.0624
258	66564	16.0624	.003876	.0623
259	67081	16.0935	.003861	.0621
260	67600	16.1245	.003846	.0620
261	68121	16.1555	.003831	.0619
262	68644	16.1864	.003817	.0618
263	69169	16.2173	.003802	.0617
264	69696	16.2481	.003788	.0615
265	70225	16.2788	.003774	.0614
266	70756	16.3095	.003759	.0613
267	71289	16.3401	.003745	.0612
268	71824	16.3707	.003731	.0611
269	72361	16.4012	.003717	.0610
270	72900	16.4317	.003704	.0609
271	73441	16.4621	.003690	.0607
272	73984	16.4924	.003676	.0606
273	74529	16.5227	.003663	.0605
274	75076	16.5529	.003650	.0604
275	75625	16.5831	.003636	.0603
276	76176	16.6132	.003623	.0602
277	76729	16.6433	.003610	.0601
278	77284	16.6733	.003597	.0600
279	77841	16.7033	.003584	.0599
280	78400	16.7332	.003571	.0598
281	78961	16.7631	.003559	.0597
282	79524	16.7929	.003546	.0595
283	80089	16.8226	.003534	.0594
284	80656	16.8523	.003521	.0593
285	81225	16.8819	.003509	.0592

TABELA A (continuação)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
286	81796	16.9115	.003497	.0591
287	82369	16.9411	.003484	.0590
288	82944	16.9706	.003472	.0589
289	83521	17.0000	.003460	.0588
290	84100	17.0294	.003448	.0587
291	84681	17.0587	.003436	.0586
292	85264	17.0880	.003425	.0585
293	85849	17.1172	.003413	.0584
294	86436	17.1464	.003401	.0583
295	87025	17.1756	.003390	.0582
296	87616	17.2047	.003378	.0581
297	88209	17.2337	.003367	.0580
298	88804	17.2627	.003356	.0579
299	89401	17.2916	.003344	.0578
300	90000	17.3205	.003333	.0577
301	90601	17.3494	.003322	.0576
302	91204	17.3781	.003311	.0575
303	91809	17.4069	.003300	.0574
304	92416	17.4356	.003289	.0574
305	93025	17.4642	.003279	.0573
306	93636	17.4929	.003268	.0572
307	94249	17.5214	.003257	.0571
308	94864	17.5499	.003247	.0570
309	95481	17.5784	.003236	.0569
310	96100	17.6068	.003226	.0568
311	96721	17.6352	.003215	.0567
312	97344	17.6635	.003205	.0566
313	97969	17.6918	.003195	.0565
314	98596	17.7200	.003185	.0564
315	99225	17.7482	.003175	.0563
316	99856	17.7764	.003165	.0563
317	100489	17.8045	.003155	.0562
318	101124	17.8326	.003145	.0561
319	101761	17.8606	.003135	.0560
320	102400	17.8885	.003125	.0559
321	103041	17.9165	.003115	.0558
322	103684	17.9444	.003106	.0557
323	104329	17.9722	.003096	.0556
324	104976	18.0000	.003086	.0556
325	105625	18.0278	.003077	.0555
326	106276	18.0555	.003067	.0554
327	106929	18.0831	.003058	.0553
328	107584	18.1108	.003049	.0552
329	108241	18.1384	.003040	.0551
330	108900	18.1659	.003030	.0550
331	109561	18.1934	.003021	.0550
332	110224	18.2209	.003012	.0549
333	110889	18.2483	.003003	.0548
334	111556	18.2757	.002994	.0547
335	112225	18.3030	.002985	.0546

TABELA A (continuação)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
336	112896	18.3303	.002976	.0546
337	113569	18.3576	.002967	.0545
338	114244	18.3848	.002959	.0544
339	114921	18.4120	.002950	.0543
340	115600	18.4391	.002941	.0542
341	116281	18.4662	.002933	.0542
342	116964	18.4932	.002924	.0541
343	117649	18.5203	.002915	.0540
344	118336	18.5472	.002907	.0539
345	119025	18.5742	.002899	.0538
346	119716	18.6011	.002890	.0538
347	120409	18.6279	.002882	.0537
348	121104	18.6548	.002874	.0536
349	121801	18.6815	.002865	.0535
350	122500	18.7083	.002857	.0535
351	123201	18.7350	.002849	.0534
352	123904	18.7617	.002841	.0533
353	124609	18.7883	.002833	.0532
354	125316	18.8149	.002825	.0531
355	126025	18.8414	.002817	.0531
356	126736	18.8680	.002809	.0530
357	127449	18.8944	.002801	.0529
358	128164	18.9209	.002793	.0529
359	128881	18.9473	.002786	.0528
360	129600	18.9737	.002778	.0527
361	130321	19.0000	.002770	.0526
362	131044	19.0263	.002762	.0526
363	131769	19.0526	.002755	.0525
364	132496	19.0788	.002747	.0524
365	133225	19.1050	.002740	.0523
366	133956	19.1311	.002732	.0523
367	134689	19.1572	.002725	.0522
368	135424	19.1833	.002717	.0521
369	136161	19.2094	.002710	.0521
370	136900	19.2354	.002703	.0520
371	137641	19.2614	.002695	.0519
372	138384	19.2873	.002688	.0518
373	139129	19.3132	.002681	.0518
374	139876	19.3391	.002674	.0517
375	140625	19.3649	.002667	.0516
376	141376	19.3907	.002660	.0516
377	142129	19.4165	.002653	.0515
378	142884	19.4422	.002646	.0514
379	143641	19.4679	.002639	.0514
380	144400	19.4936	.002632	.0513
381	145161	19.5192	.002625	.0512
382	145924	19.5448	.002618	.0512
383	146689	19.5704	.002611	.0511
384	147456	19.5959	.002604	.0510
385	148225	19.6214	.002597	.0510

TABELA A (continuação)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
386	148996	19.6469	.002591	.0509
387	149769	19.6723	.002584	.0508
388	150544	19.6977	.002577	.0508
389	151321	19.7231	.002571	.0507
390	152100	19.7484	.002564	.0506
391	152881	19.7737	.002558	.0506
392	153664	19.7990	.002551	.0505
393	154449	19.8242	.002545	.0504
394	155236	19.8494	.002538	.0504
395	156025	19.8746	.002532	.0503
396	156816	19.8997	.002525	.0503
397	157609	19.9249	.002519	.0502
398	158404	19.9499	.002513	.0501
399	159201	19.9750	.002506	.0501
400	160000	20.0000	.002500	.0500
401	160801	20.0250	.002494	.0499
402	161604	20.0499	.002488	.0499
403	162409	20.0749	.002481	.0498
404	163216	20.0998	.002475	.0498
405	164025	20.1246	.002469	.0497
406	164836	20.1494	.002463	.0496
407	165649	20.1742	.002457	.0496
408	166464	20.1990	.002451	.0495
409	167281	20.2237	.002445	.0494
410	168100	20.2485	.002439	.0494
411	168921	20.2731	.002433	.0493
412	169744	20.2978	.002427	.0493
413	170569	20.3224	.002421	.0492
414	171396	20.3470	.002415	.0491
415	172225	20.3715	.002410	.0491
416	173056	20.3961	.002404	.0490
417	173889	20.4206	.002398	.0490
418	174724	20.4450	.002392	.0489
419	175561	20.4695	.002387	.0489
420	176400	20.4939	.002381	.0488
421	177241	20.5183	.002375	.0487
422	178084	20.5426	.002370	.0487
423	178929	20.5670	.002364	.0486
424	179776	20.5913	.002358	.0486
425	180625	20.6155	.002353	.0485
426	181476	20.6398	.002347	.0485
427	182329	20.6640	.002342	.0484
428	183184	20.6882	.002336	.0483
429	184041	20.7123	.002331	.0483
430	184900	20.7364	.002326	.0482
431	185761	20.7605	.002320	.0482
432	186624	20.7846	.002315	.0481
433	187489	20.8087	.002309	.0481
434	188356	20.8327	.002304	.0480
435	189225	20.8567	.002299	.0479

TABELA A (continuação)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
436	190096	20.8806	.002294	.0479
437	190969	20.9045	.002288	.0478
438	191844	20.9284	.002283	.0478
439	192721	20.9523	.002278	.0477
440	193600	20.9762	.002273	.0477
441	194481	21.0000	.002268	.0476
442	195364	21.0238	.002262	.0476
443	196249	21.0476	.002257	.0475
444	197136	21.0713	.002252	.0475
445	198025	21.0950	.002247	.0474
446	198916	21.1187	.002242	.0474
447	199809	21.1424	.002237	.0473
448	200704	21.1660	.002232	.0472
449	201601	21.1896	.002227	.0472
450	202500	21.2132	.002222	.0471
451	203401	21.2368	.002217	.0471
452	204304	21.2603	.002212	.0470
453	205209	21.2838	.002208	.0470
454	206116	21.3073	.002203	.0469
455	207025	21.3307	.002198	.0469
456	207936	21.3542	.002193	.0468
457	208849	21.3776	.022188	.0468
458	209764	21.4009	.002183	.0467
459	210681	21.4243	.002179	.0467
460	211600	21.4476	.002174	.0466
461	212521	21.4709	.002169	.0466
462	213444	21.4942	.002165	.0465
463	214369	21.5174	.002160	.0465
464	215296	21.5407	.002155	.0464
465	216225	21.5639	.002151	.0464
466	217156	21.5870	.002146	.0463
467	218089	21.6102	.002141	.0463
468	219024	21.6333	.002137	.0462
469	219961	21.6564	.002132	.0462
470	220900	21.6795	.002128	.0461
471	221841	21.7025	.002123	.0461
472	222784	21.7256	.002119	.0460
473	223729	21.7486	.002114	.0460
474	224676	21.7715	.002110	.0459
475	225625	21.7945	.002105	.0459
476	226576	21.8174	.002101	.0458
477	227529	21.8403	.002096	.0458
478	228484	21.8632	.002092	.0457
479	229441	21.8861	.002088	.0457
480	230400	21.9089	.002083	.0456
481	231361	21.9317	.002079	.0456
482	232324	21.9545	.002075	.0455
483	233289	21.9773	.002070	.0455
484	234256	22.0000	.002066	.0455
485	235225	22.0227	.002062	.0454

TABELA A (continuação)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
486	236196	22.0454	.002058	.0454
487	237169	22.0681	.002053	.0453
488	238144	22.0907	.002049	.0453
489	239121	22.1133	.002045	.0452
490	240100	22.1359	.002041	.0452
491	241081	22.1585	.002037	.0451
492	242064	22.1811	.002033	.0451
493	243049	22.2036	.002028	.0450
494	244036	22.2261	.002024	.0450
495	245025	22.2486	.002020	.0449
496	246016	22.2711	.002016	.0448
497	247009	22.2935	.002012	.0449
498	248004	22.3159	.002008	.0449
499	249001	22.3383	.002004	.0448
500	250000	22.3607	.002000	.0447
501	251001	22.3830	.001996	.0447
502	252004	22.4054	.001992	.0446
503	253009	22.4277	.001988	.0446
504	254016	22.4499	.001984	.0445
505	255025	22.4722	.001980	.0445
506	256036	22.4944	.001976	.0445
507	257049	22.5167	.001972	.0444
508	258064	22.5389	.001969	.0444
509	259081	22.5610	.001965	.0443
510	260100	22.5832	.001961	.0443
511	261121	22.6053	.001957	.0442
512	262144	22.6274	.001953	.0442
513	263169	22.6495	.001949	.0442
514	264196	22.6716	.001946	.0441
515	265225	22.6936	.001942	.0441
516	266256	22.7156	.001938	.0440
517	267289	22.7376	.001934	.0440
518	268324	22.7596	.001931	.0439
519	269361	22.7816	.001927	.0439
520	270400	22.8035	.001923	.0439
521	271441	22.8254	.001919	.0438
522	272484	22.8473	.001916	.0438
523	273529	22.8692	.001912	.0437
524	274576	22.8910	.001908	.0437
525	275625	22.9129	.001905	.0436
526	276676	22.9347	.001901	.0436
527	277729	22.9565	.001898	.0436
528	278784	22.9783	.001894	.0435
529	279841	23.0000	.001890	.0435
530	280900	23.0217	.001887	.0434
531	281961	23.0434	.001883	.0434
532	283024	23.0651	.001880	.0434
533	284089	23.0868	.001876	.0433
534	285156	23.1084	.001873	.0433
535	286225	23.1301	.001869	.0432

TABELA A (continuação)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
536	287296	23.1517	.001866	.0432
537	288369	23.1733	.001862	.0432
538	289444	23.1948	.001859	.0431
539	290521	23.2164	.001855	.0431
540	291600	23.2379	.001852	.0430
541	292681	23.2594	.001848	.0430
542	293764	23.2809	.001845	.0430
543	294849	23.3024	.001842	.0429
544	295936	23.3238	.001838	.0429
545	297025	23.3452	.001835	.0428
546	298116	23.3666	.001832	.0428
547	299209	23.3880	.001828	.0428
548	300304	23.4094	.001825	.0427
549	301401	23.4307	.001821	.0427
550	302500	23.4521	.001818	.0426
551	303601	23.4734	.001815	.0426
552	304704	23.4947	.001812	.0426
553	305809	23.5160	.001808	.0425
554	306916	23.5372	.001805	.0425
555	308025	23.5584	.001802	.0424
556	309136	23.5797	.001799	.0424
557	310249	23.6008	.001795	.0424
558	311364	23.6220	.001792	.0423
559	312481	23.6432	.001789	.0423
560	313600	23.6643	.001786	.0423
561	314721	23.6854	.001783	.0422
562	315844	23.7065	.001779	.0422
563	316969	23.7276	.001776	.0421
564	318096	23.7487	.001773	.0421
565	319225	23.7697	.001770	.0421
566	320356	23.7908	.001767	.0420
567	321489	23.8118	.001764	.0420
568	322624	23.8328	.001761	.0420
569	323761	23.8537	.001757	.0419
570	324900	23.8747	.001754	.0419
571	326041	23.8956	.001751	.0418
572	327184	23.9165	.001748	.0418
573	328329	23.9374	.001745	.0418
574	329476	23.9583	.001742	.0417
575	330625	23.9792	.001739	.0417
576	331776	24.0000	.001736	.0417
577	332929	24.0208	.001733	.0416
578	334084	24.0416	.001730	.0416
579	335241	24.0624	.001727	.0416
580	336400	24.0832	.001724	.0415
581	337561	24.1039	.001721	.0415
582	338724	24.1247	.001718	.0415
583	339889	24.1454	.001715	.0414
584	341056	24.1661	.001712	.0414
585	342225	24.1868	.001709	.0413

TABELA A (continuação)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
586	343396	24.2074	.001706	.0413
587	344569	24.2281	.001704	.0413
588	345744	24.2487	.001701	.0412
589	346921	24.2693	.001698	.0412
590	348100	24.2899	.001695	.0412
591	349281	24.3105	.001692	.0411
592	350464	24.3311	.001689	.0411
593	351649	24.3516	.001686	.0411
594	352836	24.3721	.001684	.0410
595	354025	24.3926	.001681	.0410
596	355216	24.4131	.001678	.0410
597	356409	24.4336	.001675	.0409
598	357604	24.4540	.001672	.0409
599	358801	24.4745	.001669	.0409
600	360000	24.4949	.001667	.0408
601	361201	24.5153	.001664	.0408
602	362404	24.5357	.001661	.0408
603	363609	24.5561	.001658	.0407
604	364816	24.5764	.001656	.0407
605	366025	24.5967	.001653	.0407
606	367236	24.6171	.001650	.0406
607	368449	24.6374	.001647	.0406
608	369664	24.6577	.001645	.0406
609	370881	24.6779	.001642	.0405
610	372100	24.6982	.001639	.0405
611	373321	24.7184	.001637	.0405
612	374544	24.7386	.001634	.0404
613	375769	24.7588	.001631	.0404
614	376996	24.7790	.001629	.0404
615	378225	24.7992	.001626	.0403
616	379456	24.8193	.001623	.0403
617	380689	24.8395	.001621	.0403
618	381924	24.8596	.001618	.0402
619	383161	24.8797	.001616	.0402
620	384400	24.8998	.001613	.0402
621	385641	24.9199	.001610	.0401
622	386884	24.9399	.001608	.0401
623	388129	24.9600	.001605	.0401
624	389376	24.9800	.001603	.0400
625	390625	25.0000	.001600	.0400
626	391876	25.0200	.001597	.0400
627	393129	25.0400	.001595	.0399
628	394384	25.0599	.001592	.0399
629	395641	25.0799	.001590	.0399
630	396900	25.0998	.001587	.0398
631	398161	25.1197	.001585	.0398
632	399424	25.1396	.001582	.0398
633	400689	25.1595	.001580	.0397
634	401956	25.1794	.001577	.0397
635	403225	25.1992	.001575	.0397

TABELA A (continuação)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
636	404496	25.2190	.001572	.0397
637	405769	25.2389	.001570	.0396
638	407044	25.2587	.001567	.0396
639	408321	25.2784	.001565	.0396
640	409600	25.2982	.001563	.0395
641	410881	25.3180	.001560	.0395
642	412164	25.3377	.001558	.0395
643	413449	25.3574	.001555	.0394
644	414736	25.3772	.001553	.0394
645	416025	25.3969	.001550	.0394
646	417316	25.4165	.001548	.0393
647	418609	25.4362	.001546	.0393
648	419904	25.4558	.001543	.0393
649	421201	25.4755	.001541	.0393
650	422500	25.4951	.001538	.0392
651	423801	25.5147	.001536	.0392
652	425104	25.5343	.001534	.0392
653	426409	25.5539	.001531	.0391
654	427716	25.5734	.001529	.0391
655	429025	25.5930	.001527	.0391
656	430336	25.6125	.001524	.0390
657	431649	25.6320	.001522	.0390
658	432964	25.6515	.001520	.0390
659	434281	25.6710	.001517	.0390
660	435600	25.6905	.001515	.0389
661	436921	25.7099	.001513	.0389
662	438244	25.7294	.001511	.0389
663	439569	25.7488	.001508	.0388
664	440896	25.7682	.001506	.0388
665	442225	25.7876	.001504	.0388
666	443556	25.8070	.001502	.0387
667	444889	25.8263	.001499	.0387
668	446224	25.8457	.001497	.0387
669	447561	25.8650	.001495	.0387
670	448900	25.8844	.001493	.0386
671	450241	25.9037	.001490	.0386
672	451584	25.9230	.001488	.0386
673	452929	25.9422	.001486	.0385
674	454276	25.9615	.001484	.0385
675	455625	25.9808	.001481	.0385
676	456976	26.0000	.001479	.0385
677	458329	26.0192	.001477	.0384
678	459684	26.0384	.001475	.0384
679	461041	26.0576	.001473	.0384
680	462400	26.0768	.001471	.0383
681	463761	26.0960	.001468	.0383
682	465124	26.1151	.001466	.0383
683	466489	26.1343	.001464	.0383
684	467856	26.1534	.001462	.0382
685	469225	26.1725	.001460	.0382

TABELA A (continuação)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
686	470596	26.1916	.001458	.0382
687	471969	26.2107	.001456	.0382
688	473344	26.2298	.001453	.0381
689	474721	26.2488	.001451	.0381
690	476100	26.2679	.001449	.0381
691	477481	26.2869	.001447	.0380
692	478864	26.3059	.001445	.0380
693	480249	26.3249	.001443	.0380
694	481636	26.3439	.001441	.0380
695	483025	26.3629	.001439	.0379
696	484416	26.3818	.001437	.0379
697	485809	26.4008	.001435	.0379
698	487204	26.4197	.001433	.0379
699	488601	26.4386	.001431	.0378
700	490000	26.4575	.001429	.0378
701	491401	26.4764	.001427	.0378
702	492804	26.4953	.001425	.0377
703	494209	26.5141	.001422	.0377
704	495616	26.5330	.001420	.0377
705	497025	26.5518	.001418	.0377
706	498436	26.5707	.001416	.0376
707	499849	26.5895	.001414	.0376
708	501264	26.6083	.001412	.0376
709	502681	26.6271	.001410	.0376
710	504100	26.6458	.001408	.0375
711	505521	26.6646	.001406	.0375
712	506944	26.6833	.001404	.0375
713	508369	26.7021	.001403	.0375
714	509796	26.7208	.001401	.0374
715	511225	26.7395	.001399	.0374
716	512656	26.7582	.001397	.0374
717	514089	26.7769	.001395	.0373
718	515524	26.7955	.001393	.0373
719	516961	26.8142	.001391	.0373
720	518400	26.8328	.001389	.0373
721	519841	26.8514	.001387	.0372
722	521284	26.8701	.001385	.0372
723	522729	26.8887	.001383	.0372
724	524176	26.9072	.001381	.0372
725	525625	26.9258	.001379	.0371
726	527076	26.9444	.001377	.0371
727	528529	26.9629	.001376	.0371
728	529984	26.9815	.001374	.0371
729	531441	27.0000	.001372	.0370
730	532900	27.0185	.001370	.0370
731	534361	27.0370	.001368	.0370
732	535824	27.0555	.001366	.0370
733	537289	27.0740	.001364	.0369
734	538756	27.0924	.001362	.0369
735	540225	27.1109	.001361	.0369

TABELA A (continuação)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
736	541696	27.1293	.001359	.0369
737	543169	27.1477	.001357	.0368
738	544644	27.1662	.001355	.0368
739	546121	27.1846	.001353	.0368
740	547600	27.2029	.001351	.0368
741	549081	27.2213	.001350	.0367
742	550564	27.2397	.001348	.0367
743	552049	27.2580	.001346	.0367
744	553536	27.2764	.001344	.0367
745	555025	27.2947	.001342	.0366
746	556516	27.3130	.001340	.0366
747	558009	27.3313	.001339	.0366
748	559504	27.3496	.001337	.0366
749	561001	27.3679	.001335	.0365
750	562500	27.3861	.001333	.0365
751	564001	27.4044	.001332	.0365
752	565504	27.4226	.001330	.0365
753	567009	27.4408	.001328	.0364
754	568516	27.4591	.001326	.0364
755	570025	27.4773	.001325	.0364
756	571536	27.4955	.001323	.0364
757	573049	27.5136	.001321	.0363
758	574564	27.5318	.001319	.0363
759	576081	27.5500	.001318	.0363
760	577600	27.5681	.001316	.0363
761	579121	27.5862	.001314	.0363
762	580644	27.6043	.001312	.0362
763	582169	27.6225	.001311	.0362
764	583696	27.6405	.001309	.0362
765	585225	27.6586	.001307	.0362
766	586756	27.6767	.001305	.0361
767	588289	27.6948	.001304	.0361
768	589824	27.7128	.001302	.0361
769	591361	27.7308	.001300	.0361
770	592900	27.7489	.001299	.0360
771	594441	27.7669	.001297	.0360
772	595984	27.7849	.001295	.0360
773	597529	27.8029	.001294	.0360
774	599076	27.8209	.001292	.0359
775	600625	27.8388	.001290	.0359
776	602176	27.8568	.001289	.0359
777	603729	27.8747	.001287	.0359
778	605284	27.8927	.001285	.0359
779	606841	27.9106	.001284	.0358
780	608400	27.9285	.001282	.0358
781	609961	27.9464	.001280	.0358
782	611524	27.9643	.001279	.0358
783	613089	27.9821	.001277	.0357
784	614656	28.0000	.001276	.0357
785	616225	28.0179	.001274	.0357

TABELA A (continuação)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
786	617796	28.0357	.001272	.0357
787	619369	28.0535	.001271	.0356
788	620944	28.0713	.001269	.0356
789	622521	28.0891	.001267	.0356
790	624100	28.1069	.001266	.0356
791	625681	28.1247	.001264	.0356
792	627264	28.1425	.001263	.0355
793	628849	28.1603	.001261	.0355
794	630436	28.1780	.001259	.0355
795	632025	28.1957	.001258	.0355
796	633616	28.2135	.001256	.0354
797	635209	28.2312	.001255	.0354
798	636804	28.2489	.001253	.0354
799	638401	28.2666	.001252	.0354
800	640000	28.2843	.001250	.0354
801	641601	28.3019	.001248	.0353
802	643204	28.3196	.001247	.0353
803	644809	28.3373	.001245	.0353
804	646416	28.3549	.001244	.0353
805	648025	28.3725	.001242	.0352
806	649636	28.3901	.001241	.0352
807	651249	28.4077	.001239	.0352
808	652864	28.4253	.001238	.0352
809	654481	28.4429	.001236	.0352
810	656100	28.4605	.001235	.0351
811	657721	28.4781	.001233	.0351
812	659344	28.4956	.001232	.0351
813	660969	28.5132	.001230	.0351
814	662596	28.5307	.001229	.0351
815	664225	28.5482	.001227	.0350
816	665856	28.5657	.001225	.0350
817	667489	28.5832	.001224	.0350
818	669124	28.6007	.001222	.0350
819	670761	28.6182	.001221	.0349
820	672400	28.6356	.001220	.0349
821	674041	28.6531	.001218	.0349
822	675684	28.6705	.001217	.0349
823	677329	28.6880	.001215	.0349
824	678976	28.7054	.001214	.0348
825	680625	28.7228	.001212	.0348
826	682276	28.7402	.001211	.0348
827	683929	28.7576	.001209	.0348
828	685584	28.7750	.001208	.0348
829	687241	28.7924	.001206	.0347
830	688900	28.8097	.001205	.0347
831	690561	28.8271	.001203	.0347
832	692224	28.8444	.001202	.0347
833	693889	28.8617	.001200	.0346
834	695556	28.8791	.001199	.0346
835	697225	28.8964	.001198	.0346

TABELA A (continuação)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
836	698896	28.9137	.001196	.0346
837	700569	28.9310	.001195	.0346
838	702244	28.9482	.001193	.0345
839	703921	28.9655	.001192	.0345
840	705600	28.9828	.001190	.0345
841	707281	29.0000	.001189	.0345
842	708964	29.0172	.001188	.0345
843	710649	29.0345	.001186	.0344
844	712336	29.0517	.001185	.0344
845	714025	29.0689	.001183	.0344
846	715716	29.0861	.001182	.0344
847	717409	29.1033	.001181	.0344
848	719104	29.1204	.001179	.0343
849	720801	29.1376	.001178	.0343
850	722500	29.1548	.001176	.0343
851	724201	29.1719	.001175	.0343
852	725904	29.1890	.001174	.0343
853	727609	29.2062	.001172	.0342
854	729316	29.2233	.001171	.0342
855	731025	29.2404	.001170	.0342
856	732736	29.2575	.001168	.0342
857	734449	29.2746	.001167	.0342
858	736164	29.2916	.001166	.0341
859	737881	29.3087	.001164	.0341
860	739600	29.3258	.001163	.0341
861	741321	29.3428	.001161	.0341
862	743044	29.3598	.001160	.0341
863	744769	29.3769	.001159	.0340
864	746496	29.3939	.001157	.0340
865	748225	29.4109	.001156	.0340
866	749956	29.4279	.001155	.0340
867	751689	29.4449	.001153	.0340
868	753424	29.4618	.001152	.0339
869	755161	29.4788	.001151	.0339
870	756900	29.4958	.001149	.0339
871	758641	29.5127	.001148	.0339
872	760384	29.5296	.001147	.0339
873	762129	29.5466	.001145	.0338
874	763876	29.5635	.001144	.0338
875	765625	29.5804	.001143	.0338
876	767376	29.5973	.001142	.0338
877	769129	29.6142	.001140	.0338
878	770884	29.6311	.001139	.0337
879	772641	29.6479	.001138	.0337
880	774400	29.6648	.001136	.0337
881	776161	29.6816	.001135	.0337
882	777924	29.6985	.001134	.0337
883	779689	29.7153	.001133	.0337
884	781456	29.7321	.001131	.0336
885	783225	29.7489	.001130	.0336

TABELA A (continuação)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
886	784996	29.7658	.001129	.0336
887	786769	29.7825	.001127	.0336
888	788544	29.7993	.001126	.0336
889	790321	29.8161	.001125	.0335
890	792100	29.8329	.001124	.0335
891	793881	29.8496	.001122	.0335
892	795664	29.8664	.001121	.0335
893	797449	29.8831	.001120	.0335
894	799236	29.8998	.001119	.0334
895	801025	29.9166	.001117	.0334
896	802816	29.9333	.001116	.0334
897	804609	29.9500	.001115	.0334
898	806404	29.9666	.001114	.0334
899	808201	29.9833	.001112	.0334
900	810000	30.0000	.001111	.0333
901	811801	30.0167	.001110	.0333
902	813604	30.0333	.001109	.0333
903	815409	30.0500	.001107	.0333
904	817216	30.0666	.001106	.0333
905	819025	30.0832	.001105	.0332
906	820836	30.0998	.001104	.0332
907	822649	30.1164	.001103	.0332
908	824464	30.1330	.001101	.0332
909	826281	30.1496	.001100	.0332
910	828100	30.1662	.001099	.0331
911	829921	30.1828	.001098	.0331
912	831744	30.1993	.001096	.0331
913	833569	30.2159	.001095	.0331
914	835396	30.2324	.001094	.0331
915	837225	30.2490	.001093	.0331
916	839056	30.2655	.001092	.0330
917	840889	30.2820	.001091	.0330
918	842724	30.2985	.001089	.0330
919	844561	30.3150	.001088	.0330
920	846400	30.3315	.001087	.0330
921	848241	30.3480	.001086	.0330
922	850084	30.3645	.001085	.0329
923	851929	30.3809	.001083	.0329
924	853776	30.3974	.001082	.0329
925	855625	30.4138	.001081	.0329
926	857476	30.4302	.001080	.0329
927	859329	30.4467	.001079	.0328
928	861184	30.4631	.001078	.0328
929	863041	30.4795	.001076	.0328
930	864900	30.4959	.001075	.0328
931	866761	30.5123	.001074	.0328
932	868624	30.5287	.001073	.0328
933	870489	30.5450	.001072	.0327
934	872356	30.5614	.001071	.0327
935	874225	30.5778	.001070	.0327

TABELA A (continuação)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
936	876096	30.5941	.001068	.0327
937	877969	30.6105	.001067	.0327
938	879844	30.6268	.001066	.0327
939	881721	30.6431	.001065	.0326
940	883600	30.6594	.001064	.0326
941	885481	30.6757	.001063	.0326
942	887364	30.6920	.001062	.0326
943	889249	30.7083	.001060	.0326
944	891136	30.7246	.001059	.0325
945	893025	30.7409	.001058	.0325
946	894916	30.7571	.001057	.0325
947	896809	30.7734	.001056	.0325
948	898704	30.7896	.001055	.0325
949	900601	30.8058	.001054	.0325
950	902500	30.8221	.001053	.0324
951	904401	30.8383	.001052	.0324
952	906304	30.8545	.001050	.0324
953	908209	30.8707	.001049	.0324
954	910116	30.8869	.001048	.0324
955	912025	30.9031	.001047	.0324
956	913936	30.9192	.001046	.0323
957	915849	30.9354	.001045	.0323
958	917764	30.9516	.001044	.0323
959	919681	30.9677	.001043	.0323
960	921600	30.9839	.001042	.0323
961	923521	31.0000	.001041	.0323
962	925444	31.0161	.001040	.0322
963	927369	31.0322	.001038	.0322
964	929296	31.0483	.001037	.0322
965	931225	31.0644	.001036	.0322
966	933156	31.0805	.001035	.0322
967	935089	31.0966	.001034	.0322
968	937024	31.1127	.001033	.0321
969	938961	31.1288	.001032	.0321
970	940900	31.1448	.001031	.0321
971	942841	31.1609	.001030	.0321
972	944784	31.1769	.001029	.0321
973	946729	31.1929	.001028	.0321
974	948676	31.2090	.001027	.0320
975	950625	31.2250	.001026	.0320
976	952576	31.2410	.001025	.0320
977	954529	31.2570	.001024	.0320
978	956484	31.2730	.001022	.0320
979	958441	31.2890	.001021	.0320
980	960400	31.3050	.001020	.0319
981	962361	31.3209	.001019	.0319
982	964324	31.3369	.001018	.0319
983	966289	31.3528	.001017	.0319
984	968256	31.3688	.001016	.0319
985	970225	31.3847	.001015	.0319

TABELA A (continuação)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
986	972196	31.4006	.001014	.0318
987	974169	31.4166	.001013	.0318
988	976144	31.4325	.001012	.0318
989	978121	31.4484	.001011	.0318
990	980100	31.4643	.001010	.0318
991	982081	31.4802	.001009	.0318
992	984064	31.4960	.001008	.0318
993	986049	31.5119	.001007	.0317
994	988036	31.5278	.001006	.0317
995	990025	31.5436	.001005	.0317
996	992016	31.5595	.001004	.0317
997	994009	31.5753	.001003	.0317
998	996004	31.5911	.001002	.0317
999	998001	31.6070	.001001	.0316
1000	1000000	31.6228	.001000	.0316

TABELA B Porcentagem da Área sob a Curva Normal Entre $\bar{X}$ e z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	00.00	00.40	00.80	01.20	01.60	01.99	02.39	02.79	03.19	03.59
0.1	03.98	04.38	04.78	05.17	05.57	05.96	06.36	06.75	07.14	07.53
0.2	07.93	08.32	08.71	09.10	09.48	09.87	10.26	10.64	11.03	11.41
0.3	11.79	12.17	12.55	12.93	13.31	13.68	14.06	14.43	14.80	15.17
0.4	15.54	15.91	16.28	16.64	17.00	17.36	17.72	18.08	18.44	18.79
0.5	19.15	19.50	19.85	20.19	20.54	20.88	21.23	21.57	21.90	22.24
0.6	22.57	22.91	23.24	23.57	23.89	24.22	24.54	24.86	25.17	25.49
0.7	25.80	26.11	26.42	26.73	27.04	27.34	27.64	27.94	28.23	28.52
0.8	28.81	29.10	29.39	29.67	29.95	30.23	30.51	30.78	31.06	31.33
0.9	31.59	31.86	32.12	32.38	32.64	32.90	33.15	33.40	33.65	33.89
1.0	34.13	34.38	34.61	34.85	35.08	35.31	35.54	35.77	35.99	36.21
1.1	36.43	36.65	36.86	37.08	37.29	37.49	37.70	37.90	38.10	38.30
1.2	38.49	38.69	38.88	39.07	39.25	39.44	39.62	39.80	39.97	40.15
1.3	40.32	40.49	40.66	40.82	40.99	41.15	41.31	41.47	41.62	41.77
1.4	41.92	42.07	42.22	42.36	42.51	42.65	42.79	42.92	43.06	43.19
1.5	43.32	43.45	43.57	43.70	43.83	43.94	44.06	44.18	44.29	44.41
1.6	44.52	44.63	44.74	44.84	44.95	45.05	45.15	45.25	45.35	45.45
1.7	45.54	45.64	45.73	45.82	45.91	45.99	46.08	46.16	46.25	46.33
1.8	46.41	46.49	46.56	46.64	46.71	46.78	46.86	46.93	46.99	47.06
1.9	47.13	47.19	47.26	47.32	47.38	47.44	47.50	47.56	47.61	47.67
2.0	47.72	47.78	47.83	47.88	47.93	47.98	48.03	48.08	48.12	48.17
2.1	48.21	48.26	48.30	48.34	48.38	48.42	48.46	48.50	48.54	48.57
2.2	48.61	48.64	48.68	48.71	48.75	48.78	48.81	48.84	48.87	48.90
2.3	48.93	48.96	48.98	49.01	49.04	49.06	49.09	49.11	49.13	49.16
2.4	49.18	49.20	49.22	49.25	49.27	49.29	49.31	49.32	49.34	49.36
2.5	49.38	49.40	49.41	49.43	49.45	49.46	49.48	49.49	49.51	49.52
2.6	49.53	49.55	49.56	49.57	49.59	49.60	49.61	49.62	49.63	49.64
2.7	49.65	49.66	49.67	49.68	49.69	49.70	49.71	49.72	49.73	49.74
2.8	49.74	49.75	49.76	49.77	49.77	49.78	49.79	49.79	49.80	49.81
2.9	49.81	49.82	49.82	49.83	49.84	49.84	49.85	49.85	49.86	49.86
3.0	49.87									
4.0	49.997									

FONTE: Karl Pearson, *Tables for Statisticians and Biometricians*, Cambridge University Press, London, pp. 98–111, com a permissão de Biometrika Trustees.

N.T.: Notar que essas porcentagens podem ser convertidas em probabilidades. Para isso, basta dividi-las por 100 – o que eqüivale a recuar a vírgula duas casas para a esquerda. Exemplo: ao z = 1,46 corresponde o valor 42,79% ou a probabilidade 0,4279.

TABELA C Valores de t aos Níveis de Significância de 0,05 e 0,01.

gl	.05	.01
1	12.706	63.657
2	4.303	9.925
3	3.182	5.841
4	2.776	4.604
5	2.571	4.032
6	2.447	3.707
7	2.365	3.499
8	2.306	3.355
9	2.262	3.250
10	2.228	3.169
11	2.201	3.106
12	2.179	3.055
13	2.160	3.012
14	2.145	2.977
15	2.131	2.947
16	2.120	2.921
17	2.110	2.898
18	2.101	2.878
19	2.093	2.861
20	2.086	2.845
21	2.080	2.831
22	2.074	2.819
23	2.069	2.807
24	2.064	2.797
25	2.060	2.787
26	2.056	2.779
27	2.052	2.771
28	2.048	2.763
29	2.045	2.756
30	2.042	2.750
40	2.021	2.704
60	2.000	2.660
120	1.980	2.617
∞	1.960	2.576

TABELA D Valores de F aos Níveis de Significância de 0,05 e 0,01.

(gl para o numerador) P = .05

gl (gl para o denominador)	1	2	3	4	5	6	8	12
1	161.4	199.5	215.7	224.6	230.2	234.0	238.9	243.9
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41
3	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79
12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69
13	4.67	3.80	3.41	3.18	3.02	2.92	2.77	2.60
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23
23	4.28	3.42	3.03	2.80	2.64	2.53	2.38	2.20
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18
25	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16
26	4.22	3.37	2.98	2.74	2.59	2.47	2.32	2.15
27	4.21	3.35	2.96	2.73	2.57	2.46	2.30	2.13
28	4.20	3.34	2.95	2.71	2.56	2.44	2.29	2.12
29	4.18	3.33	2.93	2.70	2.54	2.43	2.28	2.10
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00
60	4.00	3.15	2.76	2.52	2.37	2.25	2.10	1.92
120	3.92	3.07	2.68	2.45	2.29	2.17	2.02	1.83
∞	3.84	2.99	2.60	2.37	2.21	2.09	1.94	1.75

FONTE: R. A. Fisher and F. Yates, ***Statistical Tables for Biological, Agricultural, and Medical Research***, 4th ed, Oliver & Boyd, Edinburgh, Table V, com a permissão dos autores e do editor.

TABELA D (continuação)

(gl para o numerador) P = .01

gl (gl para o denominador)	1	2	3	4	5	6	8	12
1	4052	4999	5403	5625	5764	5859	5981	6106
2	98.49	99.01	99.17	99.25	99.30	99.33	99.36	99.42
3	34.12	30.81	29.46	28.71	28.24	27.91	27.49	27.05
4	21.20	18.00	16.69	15.98	15.52	15.21	14.80	14.37
5	16.26	13.27	12.06	11.39	10.97	10.67	10.27	9.89
6	13.74	10.92	9.78	9.15	8.75	8.47	8.10	7.72
7	12.25	9.55	8.45	7.85	7.46	7.19	6.84	6.47
8	11.26	8.65	7.59	7.01	6.63	6.37	6.03	5.67
9	10.56	8.02	6.99	6.42	6.06	5.80	5.47	5.11
10	10.04	7.56	6.55	5.99	5.64	5.39	5.06	4.71
11	9.65	7.20	6.22	5.67	5.32	5.07	4.74	4.40
12	9.33	6.93	5.95	5.41	5.06	4.82	4.50	4.16
13	9.07	6.70	5.74	5.20	4.86	4.62	4.30	3.96
14	8.86	6.51	5.56	5.03	4.69	4.46	4.14	3.80
15	8.68	6.36	5.42	4.89	4.56	4.32	4.00	3.67
16	8.53	6.23	5.29	4.77	4.44	4.20	3.89	3.55
17	8.40	6.11	5.18	4.67	4.34	4.10	3.79	3.45
18	8.28	6.01	5.09	4.58	4.25	4.01	3.71	3.37
19	8.18	5.93	5.01	4.50	4.17	3.94	3.63	3.30
20	8.10	5.85	4.94	4.43	4.10	3.87	3.56	3.23
21	8.02	5.78	4.87	4.37	4.04	3.81	3.51	3.17
22	7.94	5.72	4.82	4.31	3.99	3.76	3.45	3.12
23	7.88	5.66	4.76	4.26	3.94	3.71	3.41	3.07
24	7.82	5.61	4.72	4.22	3.90	3.67	3.36	3.03
25	7.77	5.57	4.68	4.18	3.86	3.63	3.32	2.99
26	7.72	5.53	4.64	4.14	3.82	3.59	3.29	2.96
27	7.68	5.49	4.60	4.11	3.78	3.56	3.26	2.93
28	7.64	5.45	4.57	4.07	3.75	3.53	3.23	2.90
29	7.60	5.42	4.54	4.04	3.73	3.50	3.20	2.87
30	7.56	5.39	4.51	4.02	3.70	3.47	3.17	2.84
40	7.31	5.18	4.31	3.83	3.51	3.29	2.99	2.66
60	7.08	4.98	4.13	3.65	3.34	3.12	2.82	2.50
120	6.85	4.79	3.95	3.48	3.17	2.96	2.66	2.34
∞	6.64	4.60	3.78	3.32	3.02	2.80	2.51	2.18

TABELA E Qui-quadrado. Valores Críticos aos Níveis de Significância de 0,05 e 0,01.

gl	.05	.01
1	3.841	6.635
2	5.991	9.210
3	7.815	11.345
4	9.488	13.277
5	11.070	15.086
6	12.592	16.812
7	14.067	18.475
8	15.507	20.090
9	16.919	21.666
10	18.307	23.209
11	19.675	24.725
12	21.026	26.217
13	22.362	27.688
14	23.685	29.141
15	24.996	30.578
16	26.296	32.000
17	27.587	33.409
18	28.869	34.805
19	30.144	36.191
20	31.410	37.566
21	32.671	38.932
22	33.924	40.289
23	35.172	41.638
24	36.415	42.980
25	37.652	44.314
26	38.885	45.642
27	40.113	46.963
28	41.337	48.278
29	42.557	49.588
30	43.773	50.892

FONTE: R. A. Fisher and F. Yates, ***Statistical Tables for Biological, Agricultural, and Medical Research***, 4th ed, Oliver & Boyd, Edinburgh, Table IV, com a permissão dos autores e do editor.

TABELA F Valores de r aos Níveis de Significância de 0,05 e 0,01.

gl	.05	.01
1	.99692	.999877
2	.95000	.990000
3	.8783	.95873
4	.8114	.91720
5	.7545	.8745
6	.7067	.8343
7	.6664	.7977
8	.6319	.7646
9	.6021	.7348
10	.5760	.7079
11	.5529	.6835
12	.5324	.6614
13	.5139	.6411
14	.4973	.6226
15	.4821	.6055
16	.4683	.5897
17	.4555	.5751
18	.4438	.5614
19	.4329	.5487
20	.4227	.5368
25	.3809	.4869
30	.3494	.4487
35	.3246	.4182
40	.3044	.3932
45	.2875	.3721
50	.2732	.3541
60	.2500	.3248
70	.2319	.3017
80	.2172	.2830
90	.2050	.2673

FONTE: E. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural, and Medical Research*, 4th ed, Oliver & Boyd, Edinburgh, Table VI, com a permissão dos autores e do editor.

TABELA G Valores de r_s aos Níveis de Significância de 0,05 e 0,01.

N	.05	.01
5	1.000	—
6	.886	1.000
7	.786	.929
8	.738	.881
9	.683	.833
10	.648	.794
12	.591	.777
14	.544	.714
16	.506	.665
18	.475	.625
20	.450	.591
22	.428	.562
24	.409	.537
26	.392	.515
28	.377	.496
30	.364	.478

FONTE: E. G. Olds, *The Annals of Mathematical Statistics*, "Distribution of the Sum of Squares of Rank Differences for Small Numbers of Individuals," 1938, vol. 9 e "The 5 Percent Significance Levels for Sums of Squares of Rank Differences and a Correction," 1949, vol. 20, com a permissão do Institute of Mathematical Statistics.

TABELA H Números Aleatórios.

	Número de Colunas																		
Fileira	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	9	8	9	6	9	9	0	9	6	3	2	3	3	8	6	8	4	4	2
2	3	5	6	1	7	4	1	3	2	6	8	6	0	4	7	5	2	0	3
3	4	0	6	1	6	9	6	1	5	9	5	4	5	4	8	6	7	4	0
4	6	5	6	3	1	6	8	6	7	2	0	7	2	3	2	1	5	0	9
5	2	4	9	7	9	1	0	3	9	6	7	4	1	5	4	9	6	9	8
6	7	6	1	2	7	5	6	9	4	8	4	2	8	5	2	4	1	8	0
7	8	2	1	3	4	7	4	6	3	0	7	5	0	9	2	9	0	6	1
8	6	9	5	6	5	6	0	9	0	7	7	1	4	1	8	3	1	9	3
9	7	2	1	9	9	8	0	1	6	1	6	2	3	6	9	5	5	8	4
10	2	9	0	7	3	0	8	9	6	3	3	8	5	5	6	5	2	0	9
11	9	3	5	4	5	7	4	0	3	0	1	0	4	3	3	9	5	3	2
12	9	7	5	7	9	4	8	6	8	7	6	1	6	8	2	5	5	5	3
13	4	1	7	8	6	8	1	0	5	8	8	6	1	6	8	2	9	0	4
14	5	0	8	3	3	4	5	4	4	2	5	3	0	4	9	6	1	2	3
15	3	5	0	2	9	4	1	0	0	3	9	0	5	8	6	0	9	9	6
16	0	3	8	2	3	5	1	0	1	0	6	8	5	2	4	8	0	3	8
17	1	7	2	9	1	2	7	8	4	7	0	3	3	1	5	8	2	7	3
18	5	0	5	7	9	5	8	7	8	9	3	5	3	4	4	6	1	1	3
19	7	7	3	3	5	3	6	1	3	2	8	5	4	1	4	8	3	9	0
20	1	0	9	1	3	8	2	5	3	0	3	8	0	9	3	3	0	4	5
21	1	3	8	5	1	8	5	9	4	1	9	3	9	3	6	5	9	8	4
22	8	6	4	7	8	7	5	9	4	1	9	3	9	3	6	5	9	8	4
23	0	6	9	6	5	1	0	3	2	6	7	7	4	9	6	0	3	4	0
24	7	6	7	4	7	0	8	3	8	7	3	2	5	1	2	4	2	9	7
25	3	2	3	8	1	3	1	8	7	4	5	9	0	0	2	4	1	2	1
26	9	2	1	6	4	2	3	8	7	6	2	6	2	6	4	8	1	0	1
27	3	7	4	2	2	8	1	7	8	0	6	0	0	0	3	2	2	9	7
28	0	7	8	0	8	5	1	5	2	6	5	8	7	5	3	0	5	9	6
29	7	4	2	3	3	2	6	0	0	6	5	2	2	3	6	3	9	0	4
30	1	8	2	7	5	9	5	3	6	5	2	9	9	1	1	7	3	4	3
31	4	3	1	8	7	0	6	0	8	6	5	0	1	0	4	0	6	1	5
32	8	5	8	0	6	1	4	1	2	0	4	4	1	4	7	6	3	5	1
33	4	5	8	5	0	4	5	8	3	9	2	8	7	8	9	0	8	4	3
34	5	0	2	5	4	9	2	2	1	1	0	0	5	4	8	7	6	4	0
35	0	8	1	7	0	6	3	3	4	7	6	2	6	8	9	3	4	1	4
36	2	5	9	3	4	6	0	7	5	2	0	0	9	6	0	8	2	2	5
37	2	1	3	1	3	7	8	9	8	4	9	3	8	0	2	2	1	8	1
38	3	8	8	6	8	5	1	3	3	4	6	7	2	6	3	4	8	6	7
39	0	9	9	8	5	9	8	4	4	2	2	1	1	0	1	7	6	1	3
40	2	2	3	5	3	9	7	4	4	2	1	4	0	5	8	2	3	0	8

TABELA H (continuação)

Número de Colunas 20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	Fileira
0	9	7	1	1	9	1	2	7	3	5	1	8	4	0	4	1	0	6	0	3	1
8	3	7	7	9	1	4	9	9	5	9	2	0	1	6	1	2	6	6	7	0	2
2	5	6	3	7	8	3	3	8	4	3	9	3	9	0	0	9	8	3	5	2	3
4	7	0	8	6	6	5	9	6	2	7	3	5	9	0	1	8	0	9	6	9	4
0	9	8	7	3	5	6	8	8	1	2	0	2	3	2	6	4	3	1	9	7	5
5	1	8	8	4	7	0	1	7	6	8	2	1	6	3	2	1	8	1	8	3	6
1	3	7	8	6	9	5	4	1	7	3	8	7	1	5	6	5	6	4	3	6	7
5	9	0	1	5	2	8	6	5	5	7	8	1	8	7	1	2	4	0	4	1	8
2	2	5	5	2	1	8	6	9	8	9	8	0	5	8	9	9	4	1	3	4	9
1	3	4	2	8	5	0	7	9	8	4	3	5	8	0	9	4	6	6	0	5	10
2	6	8	6	6	4	7	1	5	1	6	4	6	7	6	0	8	7	3	5	2	11
8	6	0	1	4	2	9	8	6	8	0	7	6	5	1	9	1	3	7	0	3	12
9	5	7	0	9	8	7	6	9	0	6	5	4	0	3	6	5	6	3	5	0	13
2	2	3	4	7	8	0	2	0	8	0	3	4	9	2	5	7	7	8	6	4	14
2	4	6	1	0	5	0	6	1	4	9	4	7	3	9	1	7	6	4	5	8	15
6	3	4	8	1	6	9	5	6	2	0	4	6	1	6	8	1	9	9	1	1	16
9	0	5	1	3	6	1	9	5	4	1	2	5	4	2	9	5	6	2	4	0	17
3	6	7	0	3	5	3	7	4	1	7	5	4	8	3	7	4	8	5	7	2	18
4	3	6	6	3	6	3	0	0	9	4	2	2	5	1	8	9	5	1	9	7	19
1	0	6	9	0	2	7	3	9	8	4	0	6	9	8	2	3	2	8	0	4	20
9	1	3	5	7	9	6	2	4	3	4	6	4	9	1	3	1	7	5	2	2	21
6	4	2	2	2	1	4	5	2	2	8	3	2	1	2	6	6	0	1	8	9	22
7	2	6	9	0	7	5	3	2	5	6	2	7	6	3	8	1	4	1	5	1	23
8	2	8	2	4	4	4	2	9	1	9	8	3	4	4	1	0	4	6	9	6	24
7	3	1	4	3	0	4	7	1	3	7	4	8	6	7	3	2	6	6	2	0	25
0	6	4	5	8	3	1	4	8	1	8	3	1	6	4	3	0	2	8	7	3	26
4	2	2	8	3	2	1	9	3	0	1	7	5	9	0	9	1	2	5	8	2	27
2	9	8	7	2	0	6	4	0	2	7	1	3	1	6	8	7	0	9	2	5	28
0	8	0	5	6	8	2	4	3	6	1	3	5	2	3	5	9	8	6	2	1	29
0	1	7	6	1	5	7	9	0	3	5	3	4	2	4	8	5	6	4	0	6	30
5	1	9	8	5	2	4	5	1	7	5	3	2	4	6	7	9	9	6	7	2	31
0	3	6	6	3	7	8	6	9	7	2	8	9	0	7	2	9	4	0	8	6	32
5	0	0	0	2	0	8	9	0	1	0	6	2	0	4	6	9	6	5	4	9	33
1	9	4	4	2	6	4	2	4	1	0	2	7	9	6	8	7	5	6	9	3	34
0	0	5	3	8	3	2	7	5	0	4	7	6	4	6	3	0	4	7	5	3	35
6	2	6	2	0	6	0	1	4	8	9	6	5	9	7	3	6	7	6	5	4	36
6	3	9	0	3	5	0	9	1	2	0	5	9	7	3	2	5	9	3	0	2	37
9	7	3	3	5	4	0	6	4	9	4	7	9	1	4	3	9	7	7	1	8	38
1	9	6	2	9	4	2	9	7	0	3	8	9	5	7	0	6	9	7	2	5	39
5	9	4	5	8	6	2	3	0	6	2	9	8	6	3	0	4	1	0	7	6	40

FONTE: N. M. Downie and R. W. Heath, *Basic Statistical Methods*, 3rd ed., Harper & Row, New York, 1970. Reimpresso com a permissão da Harper & Row.

TABELA I Pontos Percentuais da Amplitude *t*.

MS_d gl	α	k = Número de Médias 2	3	4	5	6	7	8	9	10	11
5	.05	3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99	7.17
	.01	5.70	6.98	7.80	8.42	8.91	9.32	9.67	9.97	10.24	10.48
6	.05	3.46	4.34	4.90	5.30	5.63	5.90	6.12	6.32	6.49	6.65
	.01	5.24	6.33	7.03	7.56	7.97	8.32	8.61	8.87	9.10	9.30
7	.05	3.34	4.16	4.68	5.06	5.36	5.61	5.82	6.00	6.16	6.30
	.01	4.95	5.92	6.54	7.01	7.37	7.68	7.94	8.17	8.37	8.55
8	.05	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92	6.05
	.01	4.75	5.64	6.20	6.62	6.96	7.24	7.47	7.68	7.86	8.03
9	.05	3.20	3.95	4.41	4.76	5.02	5.24	5.43	5.59	5.74	5.87
	.01	4.60	5.43	5.96	6.35	6.66	6.91	7.13	7.33	7.49	7.65
10	.05	3.15	3.88	4.33	4.65	4.91	5.12	5.30	5.46	5.60	5.72
	.01	4.48	5.27	5.77	6.14	6.43	6.67	6.87	7.05	7.21	7.36
11	.05	3.11	3.82	4.26	4.57	4.82	5.03	5.20	5.35	5.49	5.61
	.01	4.39	5.15	5.62	5.97	6.25	6.48	6.67	6.84	6.99	7.13
12	.05	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.39	5.51
	.01	4.32	5.05	5.50	5.84	6.10	6.32	6.51	6.67	6.81	6.94
13	.05	3.06	3.73	4.15	4.45	4.69	4.88	5.05	5.19	5.32	5.43
	.01	4.26	4.96	5.40	5.73	5.98	6.19	6.37	6.53	6.67	6.79
14	.05	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25	5.36
	.01	4.21	4.89	5.32	5.63	5.88	6.08	6.26	6.41	6.54	6.66
15	.05	3.01	3.67	4.08	4.37	4.59	4.78	4.94	5.08	5.20	5.31
	.01	4.17	4.84	5.25	5.56	5.80	5.99	6.16	6.31	6.44	6.55
16	.05	3.00	3.65	4.05	4.33	4.56	4.74	4.90	5.03	5.15	5.26
	.01	4.13	4.79	5.19	5.49	5.72	5.92	6.08	6.22	6.35	6.46
17	.05	2.98	3.63	4.02	4.30	4.52	4.70	4.86	4.99	5.11	5.21
	.01	4.10	4.74	5.14	5.43	5.66	5.85	6.01	6.15	6.27	6.38
18	.05	2.97	3.61	4.00	4.28	4.49	4.67	4.82	4.96	5.07	5.17
	.01	4.07	4.70	5.09	5.38	5.60	5.79	5.94	6.08	6.20	6.31
19	.05	2.96	3.59	3.98	4.25	4.47	4.65	4.79	4.92	5.04	5.14
	.01	4.05	4.67	5.05	5.33	5.55	5.73	5.89	6.02	6.14	6.25
20	.05	2.95	3.58	3.96	4.23	4.45	4.62	4.77	4.90	5.01	5.11
	.01	4.02	4.64	5.02	5.29	5.51	5.69	5.84	5.97	6.09	6.19
24	.05	2.92	3.53	3.90	4.17	4.37	4.54	4.68	4.81	4.92	5.01
	.01	3.96	4.55	4.91	5.17	5.37	5.54	5.69	5.81	5.92	6.02
30	.05	2.89	3.49	3.85	4.10	4.30	4.46	4.60	4.72	4.82	4.92
	.01	3.89	4.45	4.80	5.05	5.24	5.40	5.54	5.65	5.76	5.85
40	.05	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.73	4.82
	.01	3.82	4.37	4.70	4.93	5.11	5.26	5.39	5.50	5.60	5.69
60	.05	2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65	4.73
	.01	3.76	4.28	4.59	4.82	4.99	5.13	5.25	5.36	5.45	5.53
120	.05	2.80	3.36	3.68	3.92	4.10	4.24	4.36	4.47	4.56	4.64
	.01	3.70	4.20	4.50	4.71	4.87	5.01	5.12	5.21	5.30	5.37
∞	.05	2.77	3.31	3.63	3.86	4.03	4.17	4.29	4.39	4.47	4.55
	.01	3.64	4.12	4.40	4.60	4.76	4.88	4.99	5.08	5.16	5.23

FONTE: E. S. Pearson and H. O. Hartley, *Biometrika Tables for Statisticians*, Vol. 1. 3rd ed., Cambridge Press, New York, 1966, com a permissão dos Biometrika Trustees.

TABELA J Subseqüências: $S(n_1 ; n_2)$ Valores críticos, S_1, S_2, S_3 e S_4, do número S de subseqüências tais que $P(S_1 \leqslant S \leqslant S_2) = 0{,}95$ e $P(S_3 \leqslant S \leqslant S_4) = 0{,}99$

n_1 \ n_2	3	4	5	6	7	8	9	10	11	12	13	14	15
5	2 ↦ 7 2 7	3 ↦ 8 2 9	3 ↦ 9 2 10										
6	3 ↦ 7 2 7	3 ↦ 8 2 9	4 ↦ 9 3 10	4 ↦ 10 3 11									
7	3 ↦ 7 2 7	3 ↦ 9 2 9	4 ↦ 10 3 11	4 ↦ 11 3 12	4 ↦ 12 4 12								
8	3 ↦ 7 2 7	4 ↦ 9 3 9	4 ↦ 10 3 11	4 ↦ 11 4 12	5 ↦ 12 4 13	5 ↦ 13 4 14							
9	3 ↦ 7 2 7	4 ↦ 9 3 9	4 ↦ 11 3 11	5 ↦ 12 4 13	5 ↦ 13 4 14	6 ↦ 13 4 14	6 ↦ 14 5 15						
10	3 ↦ 7 2 7	4 ↦ 9 3 9	4 ↦ 11 4 11	5 ↦ 12 4 13	6 ↦ 13 4 14	6 ↦ 14 5 15	6 ↦ 15 5 16	7 ↦ 15 6 16					
11	3 ↦ 7 2 7	4 ↦ 9 3 9	5 ↦ 11 4 11	5 ↦ 12 4 13	6 ↦ 13 5 14	6 ↦ 14 5 15	7 ↦ 15 6 16	7 ↦ 16 6 17	8 ↦ 16 6 18				
12	3 ↦ 7 3 7	4 ↦ 9 3 9	5 ↦ 11 4 11	5 ↦ 12 4 13	6 ↦ 13 5 15	7 ↦ 15 5 16	7 ↦ 15 6 17	8 ↦ 16 6 18	8 ↦ 17 7 18	8 ↦ 18 7 19			
13	3 ↦ 7 3 7	4 ↦ 9 3 9	5 ↦ 11 4 11	6 ↦ 13 4 13	6 ↦ 14 5 15	7 ↦ 15 6 16	7 ↦ 16 6 17	8 ↦ 17 6 18	8 ↦ 18 7 19	9 ↦ 18 7 20	9 ↦ 19 8 20		
14	3 ↦ 7 3 7	4 ↦ 9 3 9	5 ↦ 11 4 11	6 ↦ 13 5 13	6 ↦ 14 5 15	7 ↦ 15 6 16	8 ↦ 16 6 17	8 ↦ 17 7 18	9 ↦ 18 7 19	9 ↦ 19 8 20	10 ↦ 19 8 21	10 ↦ 20 8 22	
15	4 ↦ 7 3 7	4 ↦ 9 4 9	5 ↦ 11 4 11	6 ↦ 13 5 13	7 ↦ 14 5 15	7 ↦ 15 6 17	8 ↦ 17 7 18	8 ↦ 17 7 19	9 ↦ 18 8 20	9 ↦ 19 8 21	10 ↦ 20 8 21	10 ↦ 21 9 22	11 ↦ 21 9 23
16	4 ↦ 7 3 7	5 ↦ 9 4 9	5 ↦ 11 4 11	6 ↦ 13 5 13	7 ↦ 15 6 15	7 ↦ 16 6 17	8 ↦ 17 7 18	9 ↦ 18 7 19	9 ↦ 19 8 20	10 ↦ 20 8 21	10 ↦ 20 9 22	11 ↦ 21 9 23	11 ↦ 22 10 23
17	4 ↦ 7 3 7	5 ↦ 9 4 9	5 ↦ 11 4 11	6 ↦ 12 5 13	7 ↦ 15 6 15	8 ↦ 16 6 17	8 ↦ 17 7 18	9 ↦ 18 8 19	10 ↦ 19 8 21	10 ↦ 20 9 21	11 ↦ 21 9 22	11 ↦ 22 9 23	12 ↦ 22 10 24
18	4 ↦ 7 3 7	5 ↦ 9 4 9	6 ↦ 11 5 11	6 ↦ 13 5 13	7 ↦ 15 6 15	8 ↦ 16 7 17	9 ↦ 17 7 19	9 ↦ 18 8 20	10 ↦ 19 8 21	10 ↦ 20 9 22	11 ↦ 21 9 23	11 ↦ 22 10 24	12 ↦ 23 10 24
19	4 ↦ 7 3 7	5 ↦ 9 4 9	6 ↦ 11 5 11	7 ↦ 13 5 13	7 ↦ 15 6 15	8 ↦ 16 7 17	9 ↦ 17 7 19	9 ↦ 19 8 20	10 ↦ 20 9 21	11 ↦ 21 9 22	11 ↦ 22 10 23	12 ↦ 22 10 24	12 ↦ 23 11 25
20	4 ↦ 7 3 7	5 ↦ 9 4 9	6 ↦ 11 5 11	7 ↦ 13 5 13	7 ↦ 15 6 15	8 ↦ 16 7 17	9 ↦ 17 8 19	10 ↦ 19 8 20	10 ↦ 20 9 21	11 ↦ 21 9 22	11 ↦ 22 10 23	12 ↦ 23 10 24	13 ↦ 24 11 26

n_1 \ n_2	16	17	18	19	20
16	12 ↦ 22 10 24				
17	12 ↦ 23 10 25	12 ↦ 24 11 25			
18	12 ↦ 24 11 25	13 ↦ 24 11 26	13 ↦ 25 12 26		
19	13 ↦ 24 11 26	13 ↦ 25 11 26	14 ↦ 25 12 27	14 ↦ 26 12 28	
20	13 ↦ 24 11 26	14 ↦ 25 12 27	14 ↦ 26 12 28	14 ↦ 26 13 28	15 ↦ 27 13 29

FONTE: J. Severo e W.O. Bussab, *Tábuas de Estatísticas*, Harper & Row, São Paulo, 1985.

TABELA K DISTRIBUIÇÃO BINOMIAL B (n ; p) Linha matriz: probabilidade do evento favorável: p
Blocos: valores do expoente do binômio $(q + p)^n$ Corpo da tábua: probabilidades $P(x = r)$, $r = 0, 1, 2, ..., n$
Coluna matriz: valores da variável binomial: x

2 – 3 – 4
31 – 30 – 29

p →	0,05	0,10	0,20	0,25	0,30	0,40	0,50	n=2
x=0	902	810	640	562	490	360	250	2
1	095	180	320	375	420	480	500	1
2	002	010	040	062	090	160	250	0=x
n=2	0,95	0,90	0,80	0,75	0,70	0,60	0,50	← p

p →	0,05	0,10	0,20	0,25	0,30	0,40	0,50	n=3
x=0	857	729	512	422	343	216	125	3
1	135	243	384	422	441	432	375	2
2	007	027	096	141	189	288	375	1
3	0*	001	008	016	027	064	125	0=x
n=3	0,95	0,90	0,80	0,75	0,70	0,60	0,50	← p

p →	0,05	0,10	0,20	0,25	0,30	0,40	0,50	n=4
x=0	815	656	410	316	240	130	062	4
1	171	292	410	422	412	346	250	3
2	014	049	154	211	265	346	375	2
3	0*	004	026	047	076	154	250	1
4	0*	0*	002	004	008	026	062	0=x
n=4	0,95	0,90	0,80	0,75	0,70	0,60	0,50	← p

p →	0,05	0,10	0,20	0,25	0,30	0,40	0,50	n=31
x=0	204	038	001	0*	0*	0*	0*	31
1	333	131	008	001	0*	0*	0*	30
2	263	219	029	007	001	0*	0*	29
3	134	235	070	022	006	0*	0*	28
4	049	183	122	052	017	001	0*	27
5	014	110	164	094	039	003	0*	26
6	003	053	178	135	072	009	0*	25
7	001	021	159	161	110	020	001	24
8	0*	007	119	161	142	041	004	23
9	0*	002	076	137	155	070	009	22
10	0*	0*	042	101	146	102	021	21
11	0*	0*	020	064	120	130	039	20
12	0*	0*	008	036	085	144	066	19
13	0*	0*	003	017	054	141	096	18
14	0*	0*	001	007	030	120	123	17
15	0*	0*	0*	003	014	091	140	16
16	0*	0*	0*	001	006	061	140	15
17	0*	0*	0*	0*	002	036	123	14
18	0*	0*	0*	0*	001	019	096	13
19	0*	0*	0*	0*	0*	008	066	12
20	0*	0*	0*	0*	0*	003	039	11
21	0*	0*	0*	0*	0*	001	021	10
22	0*	0*	0*	0*	0*	0*	009	9
23	0*	0*	0*	0*	0*	0*	004	8
24	0*	0*	0*	0*	0*	0*	001	7
25	0*	0*	0*	0*	0*	0*	0*	6
26	0*	0*	0*	0*	0*	0*	0*	5
27	0*	0*	0*	0*	0*	0*	0*	4
28	0*	0*	0*	0*	0*	0*	0*	3
29	0*	0*	0*	0*	0*	0*	0*	2
30	0*	0*	0*	0*	0*	0*	0*	1
31	0*	0*	0*	0*	0*	0*	0*	0=x
n=31	0,95	0,90	0,80	0,75	0,70	0,60	0,50	← p

p →	0,05	0,10	0,20	0,25	0,30	0,40	0,50	n=30
x=0	215	042	001	0*	0*	0*	0*	30
1	339	141	009	002	0*	0*	0*	29
2	259	228	034	009	002	0*	0*	28
3	127	236	079	027	007	0*	0*	27
4	045	177	133	060	021	001	0*	26
5	012	102	172	105	046	004	0*	25
6	003	047	179	145	083	012	001	24
7	0*	018	154	166	122	026	002	23
8	0*	006	111	159	150	050	005	22
9	0*	002	068	130	157	082	013	21
10	0*	0*	035	091	142	115	028	20
11	0*	0*	016	055	110	140	051	19
12	0*	0*	006	029	075	147	081	18
13	0*	0*	002	013	044	136	112	17
14	0*	0*	001	005	023	110	135	16
15	0*	0*	0*	002	011	078	144	15
16	0*	0*	0*	001	004	049	135	14
17	0*	0*	0*	0*	001	027	112	13
18	0*	0*	0*	0*	0*	013	081	12
19	0*	0*	0*	0*	0*	005	051	11
20	0*	0*	0*	0*	0*	002	028	10
21	0*	0*	0*	0*	0*	001	013	9
22	0*	0*	0*	0*	0*	0*	005	8
23	0*	0*	0*	0*	0*	0*	002	7
24	0*	0*	0*	0*	0*	0*	001	6
25	0*	0*	0*	0*	0*	0*	0*	5
26	0*	0*	0*	0*	0*	0*	0*	4
27	0*	0*	0*	0*	0*	0*	0*	3
28	0*	0*	0*	0*	0*	0*	0*	2
29	0*	0*	0*	0*	0*	0*	0*	1
30	0*	0*	0*	0*	0*	0*	0*	0=x
n=30	0,95	0,90	0,80	0,75	0,70	0,60	0,50	← p

p →	0,05	0,10	0,20	0,25	0,30	0,40	0,50	n=29
x=0	226	047	002	0*	0*	0*	0*	29
1	345	152	011	002	0*	0*	0*	28
2	254	236	039	011	002	0*	0*	27
3	120	236	088	032	009	0*	0*	26
4	041	171	144	070	026	002	0*	25
5	011	095	179	116	055	006	0*	24
6	002	042	179	155	095	015	001	23
7	0*	015	147	170	133	034	003	22
8	0*	005	101	156	157	062	008	21
9	0*	001	059	121	157	096	019	20
10	0*	0*	030	081	135	128	037	19
11	0*	0*	013	047	100	147	064	18
12	0*	0*	005	023	064	147	097	17
13	0*	0*	002	010	036	128	126	16
14	0*	0*	0*	004	018	098	144	15
15	0*	0*	0*	001	008	065	144	14
16	0*	0*	0*	0*	003	038	126	13
17	0*	0*	0*	0*	001	019	097	12
18	0*	0*	0*	0*	0*	009	064	11
19	0*	0*	0*	0*	0*	003	037	10
20	0*	0*	0*	0*	0*	001	019	9
21	0*	0*	0*	0*	0*	0*	008	8
22	0*	0*	0*	0*	0*	0*	003	7
23	0*	0*	0*	0*	0*	0*	001	6
24	0*	0*	0*	0*	0*	0*	0*	5
25	0*	0*	0*	0*	0*	0*	0*	4
26	0*	0*	0*	0*	0*	0*	0*	3
27	0*	0*	0*	0*	0*	0*	0*	2
28	0*	0*	0*	0*	0*	0*	0*	1
29	0*	0*	0*	0*	0*	0*	0*	0=x
n=29	0,95	0,90	0,80	0,75	0,70	0,60	0,50	← p

FONTE: J. Severo e W.O. Bussab, *Tábuas de Estatísticas*, Harper & Row, São Paulo, 1985.

TABELA K (continuação)

p →	0,05	0,10	0,20	0,25	0,30	0,40	0,50	n = 5
x = 0	774	590	328	237	168	078	031	5
1	204	328	410	396	360	259	156	4
2	021	073	205	264	309	346	312	3
3	001	008	051	088	132	230	312	2
4	0+	0+	006	015	028	077	156	1
5	0+	0+	0+	001	002	010	031	0 = x
n = 5	0,95	0,90	0,80	0,75	0,70	0,60	0,50	← p

p →	0,05	0,10	0,20	0,25	0,30	0,40	0,50	n = 6
x = 0	735	531	262	178	118	047	016	6
1	232	354	393	356	303	187	094	5
2	031	098	246	297	324	311	234	4
3	002	015	082	132	185	276	312	3
4	0+	001	015	033	060	138	234	2
5	0+	0+	002	004	010	037	094	1
6	0+	0+	0+	0+	001	004	016	0 = x
n = 6	0,95	0,90	0,80	0,75	0,70	0,60	0,50	← p

p →	0,05	0,10	0,20	0,25	0,30	0,40	0,50	n = 7
x = 0	698	478	210	133	082	028	008	7
1	257	372	367	311	247	131	055	6
2	041	124	275	311	318	261	164	5
3	004	023	115	173	227	290	273	4
4	0+	003	029	058	097	194	273	3
5	0+	0+	004	012	025	077	164	2
6	0+	0+	0+	001	004	017	055	1
7	0+	0+	0+	0+	0+	002	008	0
n = 7	0,05	0,90	0,80	0,75	0,70	0,60	0,50	← p

p →	0,05	0,10	0,20	0,25	0,30	0,40	0,50	n = 28
x = 0	238	052	002	0+	0+	0+	0+	28
1	350	163	014	003	001	0+	0+	27
2	249	244	046	013	003	0+	0+	26
3	114	235	099	039	012	001	0+	25
4	037	163	155	080	032	002	0+	24
5	009	087	186	128	065	008	0+	23
6	002	037	178	164	107	020	001	22
7	0+	013	140	172	145	043	004	21
8	0+	004	092	150	163	074	012	20
9	0+	001	051	111	155	110	026	19
10	0+	0+	024	071	126	140	049	18
11	0+	0+	010	038	088	152	080	17
12	0+	0+	004	018	054	144	113	16
13	0+	0+	001	007	028	118	139	15
14	0+	0+	0+	003	013	084	149	14
15	0+	0+	0+	001	005	053	139	13
16	0+	0+	0+	0+	002	028	113	12
17	0+	0+	0+	0+	001	013	080	11
18	0+	0+	0+	0+	0+	005	049	10
19	0+	0+	0+	0+	0+	002	026	9
20	0+	0+	0+	0+	0+	001	012	8
21	0+	0+	0+	0+	0+	0+	004	7
22	0+	0+	0+	0+	0+	0+	001	6
23	0+	0+	0+	0+	0+	0+	0+	5
24	0+	0+	0+	0+	0+	0+	0+	4
25	0+	0+	0+	0+	0+	0+	0+	3
26	0+	0+	0+	0+	0+	0+	0+	2
27	0+	0+	0+	0+	0+	0+	0+	1
28	0+	0+	0+	0+	0+	0+	0+	0 = x
n = 28	0,95	0,90	0,80	0,75	0,70	0,60	0,50	← p

p →	0,05	0,10	0,20	0,25	0,30	0,40	0,50	n = 27
x = 0	250	058	002	0+	0+	0+	0+	27
1	356	174	016	004	001	0+	0+	26
2	243	252	053	017	004	0+	0+	25
3	107	233	111	046	015	001	0+	24
4	034	156	166	092	039	004	0+	23
5	008	080	191	141	077	011	001	22
6	002	032	175	172	121	027	002	21
7	0+	011	131	172	155	053	007	20
8	0+	003	082	143	166	089	017	19
9	0+	001	043	101	150	125	035	18
10	0+	0+	019	060	116	150	063	17
11	0+	0+	008	031	077	154	097	16
12	0+	0+	003	014	044	137	130	15
13	0+	0+	001	005	022	105	149	14
14	0+	0+	0+	002	009	070	149	13
15	0+	0+	0+	001	003	041	130	12
16	0+	0+	0+	0+	001	020	097	11
17	0+	0+	0+	0+	0+	009	063	10
18	0+	0+	0+	0+	0+	003	035	9
19	0+	0+	0+	0+	0+	001	017	8
20	0+	0+	0+	0+	0+	0+	007	7
21	0+	0+	0+	0+	0+	0+	002	6
22	0+	0+	0+	0+	0+	0+	001	5
23	0+	0+	0+	0+	0+	0+	0+	4
24	0+	0+	0+	0+	0+	0+	0+	3
25	0+	0+	0+	0+	0+	0+	0+	2
26	0+	0+	0+	0+	0+	0+	0+	1
27	0+	0+	0+	0+	0+	0+	0+	0 = x
n = 27	0,95	0,90	0,80	0,75	0,70	0,60	0,50	← p

p →	0,05	0,10	0,20	0,25	0,30	0,40	0,50	n = 26
x = 0	264	065	003	001	0+	0+	0+	26
1	361	187	020	005	001	0+	0+	25
2	237	259	061	020	006	0+	0+	24
3	100	230	123	054	019	001	0+	23
4	030	147	176	104	047	005	0+	22
5	007	072	194	153	089	015	001	21
6	001	028	170	178	134	034	003	20
7	0+	009	121	170	164	066	010	19
8	0+	002	072	134	167	104	023	18
9	0+	001	036	090	143	139	047	17
10	0+	0+	015	051	104	157	079	16
11	0+	0+	006	025	065	152	115	15
12	0+	0+	002	010	035	127	144	14
13	0+	0+	0+	004	016	091	155	13
14	0+	0+	0+	001	006	056	144	12
15	0+	0+	0+	0+	002	030	115	11
16	0+	0+	0+	0+	001	014	079	10
17	0+	0+	0+	0+	0+	005	047	9
18	0+	0+	0+	0+	0+	002	023	8
19	0+	0+	0+	0+	0+	001	010	7
20	0+	0+	0+	0+	0+	0+	003	6
21	0+	0+	0+	0+	0+	0+	001	5
22	0+	0+	0+	0+	0+	0+	0+	4
23	0+	0+	0+	0+	0+	0+	0+	3
24	0+	0+	0+	0+	0+	0+	0+	2
25	0+	0+	0+	0+	0+	0+	0+	1
26	0+	0+	0+	0+	0+	0+	0	0 = x
n = 26	0,95	0,90	0,80	0,75	0,70	0,60	0,50	← p

TABELA K (continuação)

p →	0,05	0,10	0,20	0,25	0,30	0,40	0,50	n = 8
x = 0	663	430	168	100	058	017	004	8
1	279	383	336	267	198	090	031	7
2	051	149	294	311	296	209	109	6
3	005	033	147	208	254	279	219	5
4	0+	005	046	087	136	232	273	4
5	0+	0+	009	023	047	124	219	3
6	0+	0+	001	004	010	041	109	2
7	0+	0+	0+	0+	001	008	031	1
8	0+	0+	0+	0+	0+	001	004	0 = x
n = 8	0,95	0,90	0,80	0,75	0,70	0,60	0,50	← p

p →	0,05	0,10	0,20	0,25	0,30	0,40	0,50	n = 9
x = 0	630	387	134	075	040	010	002	9
1	299	387	302	225	156	060	018	8
2	063	172	302	300	267	161	070	7
3	008	045	176	234	267	251	164	6
4	001	007	066	117	172	251	246	5
5	0+	001	017	039	074	167	246	4
6	0+	0+	003	009	021	074	164	3
7	0+	0+	0+	001	004	021	070	2
8	0+	0+	0+	0+	0+	004	018	1
9	0+	0+	0+	0+	0+	0+	002	0 = x
n = 9	0,95	0,90	0,80	0,75	0,70	0,60	0,50	← p

p →	0,05	0,10	0,20	0,25	0,30	0,40	0,50	n = 10
x = 0	599	349	107	056	028	006	001	10
1	315	387	268	188	121	040	010	9
2	075	194	302	282	233	121	044	8
3	010	057	201	250	267	215	117	7
4	001	011	088	146	200	251	205	6
5	0+	001	026	058	103	201	246	5
6	0+	0+	006	016	037	111	205	4
7	0+	0+	001	003	009	042	117	3
8	0+	0+	0+	0+	001	011	044	2
9	0+	0+	0+	0+	0+	002	010	1
10	0+	0+	0+	0+	0+	0+	001	0 = x
n = 10	0,95	0,90	0,80	0,75	0,70	0,60	0,50	← p

p →	0,05	0,10	0,20	0,25	0,30	0,40	0,50	n = 25
x = 0	277	072	004	001	0+	0+	0+	25
1	365	199	024	006	001	0+	0+	24
2	231	266	071	025	007	0+	0+	23
3	093	226	136	064	024	002	0+	22
4	027	138	187	118	057	007	0+	21
5	006	065	196	165	103	020	002	20
6	001	024	163	183	147	044	005	19
7	0+	007	111	165	171	080	014	18
8	0+	002	062	124	165	120	032	17
9	0+	0+	029	078	134	151	061	16
10	0+	0+	012	042	092	161	097	15
11	0+	0+	004	019	054	147	133	14
12	0+	0+	001	007	027	114	155	13
13	0+	0+	0+	002	011	076	155	12
14	0+	0+	0+	001	004	043	133	11
15	0+	0+	0+	0+	001	021	097	10
16	0+	0+	0+	0+	0+	009	061	9
17	0+	0+	0+	0+	0+	003	032	8
18	0+	0+	0+	0+	0+	001	014	7
19	0+	0+	0+	0+	0+	0+	005	6
20	0+	0+	0+	0+	0+	0+	002	5
21	0+	0+	0+	0+	0+	0+	0+	4
22	0+	0+	0+	0+	0+	0+	0+	3
23	0+	0+	0+	0+	0+	0+	0+	2
24	0+	0+	0+	0+	0+	0+	0+	1
25	0+	0+	0+	0+	0+	0+	0+	0 = x
n = 25	0,95	0,90	0,80	0,75	0,70	0,60	0,50	← p

p →	0,05	0,10	0,20	0,25	0,30	0,40	0,50	n = 24
x = 0	292	080	005	001	0+	0+	0+	24
1	369	213	028	008	002	0+	0+	23
2	223	272	081	031	010	001	0+	22
3	086	221	149	075	031	003	0+	21
4	024	129	196	132	069	010	001	20
5	005	057	196	176	118	027	003	19
6	001	020	155	185	160	056	008	18
7	0+	006	100	159	176	096	021	17
8	0+	001	053	112	160	136	044	16
9	0+	0+	024	067	122	161	078	15
10	0+	0+	009	033	079	161	117	14
11	0+	0+	003	014	043	137	149	13
12	0+	0+	001	005	020	099	161	12
13	0+	0+	0+	002	008	061	149	11
14	0+	0+	0+	0+	003	032	117	10
15	0+	0+	0+	0+	001	014	078	9
16	0+	0+	0+	0+	0+	005	044	8
17	0+	0+	0+	0+	0+	002	021	7
18	0+	0+	0+	0+	0+	0+	008	6
19	0+	0+	0+	0+	0+	0+	003	5
20	0+	0+	0+	0+	0+	0+	001	4
21	0+	0+	0+	0+	0+	0+	0+	3
22	0+	0+	0+	0+	0+	0+	0+	2
23	0+	0+	0+	0+	0+	0+	0+	1
24	0+	0+	0+	0+	0+	0+	0+	0 = x
n = 24	0,95	0,90	0,80	0,75	0,70	0,60	0,50	← p

p →	0,05	0,10	0,20	0,25	0,30	0,40	0,50	n = 23
x = 0	307	089	006	001	0+	0+	0+	23
1	372	226	034	010	003	0+	0+	22
2	215	277	093	038	013	001	0+	21
3	079	215	163	088	038	004	0+	20
4	021	120	204	146	082	014	001	19
5	004	051	194	185	133	035	004	18
6	001	017	145	185	171	070	012	17
7	0+	005	088	150	178	113	029	16
8	0+	001	044	100	153	151	058	15
9	0+	0+	018	056	109	168	097	14
10	0+	0+	006	026	065	157	136	13
11	0+	0+	002	010	033	123	161	12
12	0+	0+	0+	003	014	082	161	11
13	0+	0+	0+	001	005	046	136	10
14	0+	0+	0+	0+	002	022	097	9
15	0+	0+	0+	0+	0+	009	058	8
16	0+	0+	0+	0+	0+	003	029	7
17	0+	0+	0+	0+	0+	001	012	6
18	0+	0+	0+	0+	0+	0+	004	5
19	0+	0+	0+	0+	0+	0+	001	4
20	0+	0+	0+	0+	0+	0+	0+	3
21	0+	0+	0+	0+	0+	0+	0+	2
22	0+	0+	0+	0+	0+	0+	0+	1
23	0+	0+	0+	0+	0+	0+	0+	0 = x
n = 23	0,95	0,90	0,80	0,75	0,70	0,60	0,50	← p

TABELA K (continuação)

11 — 12 — 13
22 — 21 — 20

p →	0,05	0,10	0,20	0,25	0,30	0,40	0,50	n=11
x=0	569	314	086	042	020	004	0*	11
1	329	384	236	155	093	027	005	10
2	087	213	295	258	200	089	027	9
3	014	071	221	258	257	177	081	8
4	001	016	111	172	220	236	161	7
5	0*	002	039	080	132	221	226	6
6	0*	0*	010	027	057	147	226	5
7	0*	0*	002	006	017	070	161	4
8	0*	0*	0*	001	004	023	081	3
9	0*	0*	0*	0*	001	005	027	2
10	0*	0*	0*	0*	0*	001	005	1
11	0*	0*	0*	0*	0*	0*	0*	0=x
n=11	0,95	0,90	0,80	0,75	0,70	0,40	0,50	← p

p →	0,05	0,10	0,20	0,25	0,30	0,40	0,50	n=12
x=0	540	282	069	032	014	002	0*	12
1	341	377	206	127	071	017	003	11
2	099	230	283	232	168	064	016	10
3	017	085	236	258	240	142	054	9
4	002	021	133	194	231	213	121	8
5	0*	004	053	103	158	227	193	7
6	0*	0*	016	040	079	177	226	6
7	0*	0*	003	011	029	101	193	5
8	0*	0*	001	002	008	042	121	4
9	0*	0*	0*	0*	001	012	054	3
10	0*	0*	0*	0*	0*	002	016	2
11	0*	0*	0*	0*	0*	0*	003	1
12	0*	0*	0*	0*	0*	0*	0*	0=x
n=12	0,95	0,90	0,80	0,75	0,70	0,60	0,50	← p

p →	0,05	0,10	0,20	0,25	0,30	0,40	0,50	n=13
x=0	513	254	055	024	010	001	0*	13
1	351	367	179	103	054	011	002	12
2	111	245	268	206	139	045	010	11
3	021	100	246	252	218	111	035	10
4	003	028	154	210	234	184	087	9
5	0*	006	069	126	180	221	157	8
6	0*	001	023	056	103	197	209	7
7	0*	0*	006	019	044	131	209	6
8	0*	0*	001	005	014	066	157	5
9	0*	0*	0*	001	003	024	087	4
10	0*	0*	0*	0*	001	006	035	3
11	0*	0*	0*	0*	0*	001	010	2
12	0*	0*	0*	0*	0*	0*	002	1
13	0*	0*	0*	0*	0*	0*	0*	0=x
n=13	0,95	0,90	0,80	0,75	0,70	0,60	0,50	← p

p →	0,05	0,10	0,20	0,25	0,30	0,40	0,50	n=22
x=0	324	098	007	002	0*	0*	0*	22
1	375	241	041	013	004	0*	0*	21
2	207	281	107	046	017	001	0*	20
3	073	208	178	102	047	006	0*	19
4	018	110	211	161	096	019	002	18
5	003	044	190	193	149	048	006	17
6	001	014	134	183	181	086	018	16
7	0*	004	077	139	177	131	041	15
8	0*	001	036	087	142	164	076	14
9	0*	0	014	045	095	170	119	13
10	0*	0*	005	020	053	148	154	12
11	0*	0*	001	007	025	107	168	11
12	0*	0*	0*	002	010	066	154	10
13	0*	0*	0*	001	003	034	119	9
14	0*	0*	0*	0*	001	014	076	8
15	0*	0*	0*	0*	0*	005	041	7
16	0*	0*	0*	0*	0*	001	018	6
17	0*	0*	0*	0*	0*	0*	006	5
18	0*	0*	0*	0*	0*	0*	002	4
19	0*	0*	0*	0*	0*	0*	0*	3
20	0*	0*	0*	0*	0*	0*	0*	2
21	0*	0*	0*	0*	0*	0*	0*	1
22	0*	0*	0*	0*	0*	0*	0*	0=x
n=22	0,95	0,90	0,80	0,75	0,70	0,60	0,50	← p

p →	0,05	0,10	0,20	0,25	0,30	0,40	0,50	n=21
x=0	341	109	009	002	001	0*	0*	21
1	376	255	048	017	005	0*	0*	20
2	198	284	121	055	022	002	0*	19
3	066	200	192	117	058	009	001	18
4	016	100	216	176	113	026	003	17
5	003	038	183	199	164	059	010	16
6	0*	011	122	177	188	105	026	15
7	0*	003	065	126	172	149	055	14
8	0*	001	029	074	129	174	097	13
9	0*	0*	010	036	080	168	140	12
10	0*	0*	003	014	041	134	168	11
11	0*	0*	001	005	018	089	168	10
12	0*	0*	0*	001	006	050	140	9
13	0*	0*	0*	0*	002	023	097	8
14	0*	0*	0*	0*	0*	009	055	7
15	0*	0*	0*	0*	0*	003	026	6
16	0*	0*	0*	0*	0*	001	010	5
17	0*	0*	0*	0*	0*	0*	003	4
18	0*	0*	0*	0*	0*	0*	001	3
19	0*	0*	0*	0*	0*	0*	0*	2
20	0*	0*	0*	0*	0*	0*	0*	1
21	0*	0*	0*	0*	0*	0*	0*	0=x
n=21	0,95	0,90	0,80	0,75	0,70	0,60	0,50	← p

p →	0,05	0,10	0,20	0,25	0,30	0,40	0,50	n=20
x=0	358	122	012	003	001	0*	0*	20
1	377	270	058	021	007	0*	0*	19
2	189	285	137	067	028	003	0*	18
3	060	190	205	134	072	012	001	17
4	013	090	218	190	130	035	005	16
5	002	032	175	202	179	075	015	15
6	0*	009	109	169	192	124	037	14
7	0*	002	055	112	164	166	074	13
8	0*	0*	022	061	114	180	120	12
9	0*	0*	007	027	065	160	160	11
10	0*	0*	002	010	031	117	176	10
11	0*	0*	0*	003	012	071	160	9
12	0*	0*	0*	001	004	035	120	8
13	0*	0*	0*	0*	001	015	074	7
14	0*	0*	0*	0*	0*	005	037	6
15	0*	0*	0*	0*	0*	001	015	5
16	0*	0*	0*	0*	0*	0*	005	4
17	0*	0*	0*	0*	0*	0*	001	3
18	0*	0*	0*	0*	0*	0*	0*	2
19	0*	0*	0*	0*	0*	0*	0*	1
20	0*	0*	0*	0*	0*	0*	0*	0=x
n=20	0,95	0,90	0,80	0,75	0,70	0,60	0,50	← p

11 — 12 — 13

20 — 21 — 22

TABELA K (continuação)

14 — 15 — 16
19 — 18 — 17

p →	0,05	0,10	0,20	0,25	0,30	0,40	0,50	n=14
x=0	488	229	044	018	007	001	0*	14
1	359	356	154	083	041	007	001	13
2	123	257	250	180	113	032	006	12
3	026	114	250	240	194	085	022	11
4	004	035	172	220	229	155	061	10
5	0*	008	086	147	196	207	122	9
6	0*	001	032	073	126	207	183	8
7	0*	0*	009	028	062	157	209	7
8	0*	0*	002	008	023	092	183	6
9	0*	0*	0*	002	007	041	122	5
10	0*	0*	0*	0*	001	014	061	4
11	0*	0*	0*	0*	0*	003	022	3
12	0*	0*	0*	0*	0*	001	006	2
13	0*	0*	0*	0*	0*	0*	001	1
14	0*	0*	0*	0*	0*	0*	0*	0=x
n=14	0,95	0,90	0,80	0,75	0,70	0,60	0,50	← p

p →	0,05	0,10	0,20	0,25	0,30	0,40	0,50	n=15
x=0	463	206	035	013	005	0*	0*	15
1	366	343	132	067	031	005	0*	14
2	135	267	231	156	092	022	003	13
3	031	129	250	225	170	063	014	12
4	005	043	188	225	219	127	042	11
5	001	010	103	165	206	186	092	10
6	0*	002	043	092	147	207	153	9
7	0*	0*	014	039	081	177	196	8
8	0*	0*	003	013	035	118	196	7
9	0*	0*	001	003	012	061	153	6
10	0*	0*	0*	001	003	024	092	5
11	0*	0*	0*	0*	001	007	042	4
12	0*	0*	0*	0*	0*	002	014	3
13	0	0*	0*	0*	0*	0*	003	2
14	0*	0*	0*	0*	0*	0*	0*	1
15	0*	0*	0*	0*	0*	0*	0*	0=x
n=15	0,95	0,90	0,80	0,75	0,70	0,60	0,50	← p

p →	0,05	0,10	0,20	0,25	0,30	0,40	0,50	n=16
x=0	440	185	028	010	003	0*	0*	16
1	371	329	113	053	023	003	0*	15
2	146	275	211	134	073	015	002	14
3	036	142	246	208	146	047	009	13
4	006	051	200	225	204	101	028	12
5	001	014	120	180	210	162	067	11
6	0*	003	055	110	165	198	122	10
7	0*	0*	020	052	101	189	175	9
8	0*	0*	006	020	049	142	196	8
9	0*	0*	001	006	019	084	175	7
10	0*	0*	0*	001	006	039	122	6
11	0*	0*	0*	0*	001	014	067	5
12	0*	0*	0*	0*	0*	004	028	4
13	0*	0*	0*	0*	0*	001	009	3
14	0*	0*	0*	0*	0*	0*	002	2
15	0*	0*	0*	0*	0*	0*	0*	1
16	0*	0*	0*	0*	0*	0*	0*	0=x
n=16	0,95	0,90	0,80	0,75	0,70	0,60	0,50	← p

p →	0,05	0,10	0,20	0,25	0,30	0,40	0,50	n=19
x=0	377	135	014	004	001	0*	0*	19
1	377	285	068	027	009	001	0*	18
2	179	285	154	080	036	005	0*	17
3	053	180	218	152	087	017	002	16
4	011	080	218	202	149	047	007	15
5	002	027	164	202	192	093	022	14
6	0*	007	095	157	192	145	052	13
7	0*	001	044	097	153	180	096	12
8	0*	0*	017	049	098	180	144	11
9	0*	0*	005	020	051	146	176	10
10	0*	0*	001	007	022	098	176	9
11	0*	0*	0*	002	008	053	144	8
12	0*	0*	0*	0*	002	024	096	7
13	0*	0*	0*	0*	001	008	052	6
14	0*	0*	0*	0*	0*	002	022	5
15	0*	0*	0*	0*	0*	001	007	4
16	0*	0*	0*	0*	0*	0*	002	3
17	0*	0*	0*	0*	0*	0*	0*	2
18	0*	0*	0*	0*	0*	0*	0*	1
19	0*	0*	0*	0*	0*	0*	0*	0=x
n=19	0,95	0,90	0,80	0,75	0,70	0,60	0,50	← p

p →	0,05	0,10	0,20	0,25	0,30	0,40	0,50	n=18
x=0	397	150	018	006	002	0*	0*	18
1	376	300	081	034	013	001	0*	17
2	168	284	172	096	046	007	001	16
3	047	168	230	170	105	025	003	15
4	009	070	215	213	168	061	012	14
5	001	022	151	199	202	115	033	13
6	0*	005	082	144	187	166	071	12
7	0*	001	035	082	138	189	121	11
8	0*	0*	012	038	081	173	167	10
9	0*	0*	003	014	039	128	185	9
10	0	0*	001	004	015	077	167	8
11	0*	0*	0*	001	005	037	121	7
12	0*	0*	0*	0*	001	015	071	6
13	0*	0*	0*	0*	0*	004	033	5
14	0*	0*	0*	0*	0*	001	012	4
15	0*	0*	0*	0*	0*	0*	003	3
16	0*	0*	0*	0*	0*	0*	001	2
17	0*	0*	0*	0*	0*	0*	0*	1
18	0*	0*	0*	0*	0*	0*	0*	0=x
n=18	0,95	0,90	0,80	0,75	0,70	0,60	0,50	← p

p →	0,05	0,10	0,20	0,25	0,30	0,40	0,50	n=17
x=0	418	167	023	008	002	0*	0*	17
1	374	315	096	043	017	002	0*	16
2	158	280	191	114	058	010	001	15
3	041	156	239	189	125	034	005	14
4	008	060	209	221	187	080	018	13
5	001	017	136	191	208	138	047	12
6	0*	004	068	128	178	184	094	11
7	0*	001	027	067	120	193	148	10
8	0*	0*	008	028	064	161	185	9
9	0*	0*	002	009	028	107	185	8
10	0*	0*	0*	002	009	057	148	7
11	0*	0*	0*	001	003	024	094	6
12	0*	0*	0*	0*	001	008	047	5
13	0*	0*	0*	0*	0*	002	018	4
14	0*	0*	0*	0*	0*	0*	005	3
15	0*	0*	0*	0*	0*	0*	001	2
16	0*	0*	0*	0*	0*	0*	0*	1
17	0*	0*	0*	0*	0*	0*	0*	0=x
n=17	0,95	0,90	0,80	0,75	0,70	0,60	0,50	← p

14 — 15 — 16

17 — 18 — 19

TABELA L Probabilidades Associadas ao *U* de Mann-Whitney.

(Referem-se a uma H_a unicaudal. Para uma H_a bicaudal, dobrar as probabilidades.)

$n_2 = 3$

U \ n_1	1	2	3
0	.250	.100	.050
1	.500	.200	.100
2	.750	.400	.200
3		.600	.350
4			.500
5			.650

$n_2 = 4$

U \ n_1	1	2	3	4
0	.200	.067	.028	.014
1	.400	.133	.057	.029
2	.600	.267	.114	.057
3		.400	.200	.100
4		.600	.314	.171
5			.429	.243
6			.571	.343
7				.443
8				.557

$n_2 = 5$

U \ n_1	1	2	3	4	5
0	.167	.047	.018	.008	.004
1	.333	.095	.036	.016	.008
2	.500	.190	.071	.032	.016
3	.667	.286	.125	.056	.028
4		.429	.196	.095	.048
5		.571	.286	.143	.075
6			.393	.206	.111
7			.500	.278	.155
8			.607	.365	.210
9				.452	.274
10				.548	.345
11					.421
12					.500
13					.579

$n_2 = 6$

U \ n_1	1	2	3	4	5	6
0	.143	.036	.012	.005	.002	.001
1	.286	.071	.024	.010	.004	.002
2	.428	.143	.048	.019	.009	.004
3	.571	.214	.083	.033	.015	.008
4		.321	.131	.057	.026	.013
5		.429	.190	.086	.041	.021
6		.571	.274	.129	.063	.032
7			.357	.176	.089	.047
8			.452	.238	.123	.066
9			.548	.305	.165	.090
10				.381	.214	.120
11				.457	.268	.155
12				.545	.331	.197
13					.396	.242
14					.465	.294
15					.535	.350
16						.409
17						.469
18						.531

FONTE: *Ann. Math. Statist.*, 18, 52-54 (Mann, H. B. e Whitney, D. R.).

TABELA L (continuação)

(Referem-se a uma H_a unicaudal. Para uma H_a bicaudal, dobrar as probabilidades.)

$n_2 = 7$

U \ n_1	1	2	3	4	5	6	7
0	.125	.028	.008	.003	.001	.001	.000
1	.250	.056	.017	.006	.003	.001	.001
2	.375	.111	.033	.012	.005	.002	.001
3	.500	.167	.058	.021	.009	.004	.002
4	.625	.250	.092	.036	.015	.007	.003
5		.333	.133	.055	.024	.011	.006
6		.444	.192	.082	.037	.017	.009
7		.556	.258	.115	.053	.026	.013
8			.333	.158	.074	.037	.019
9			.417	.206	.101	.051	.027
10			.500	.264	.134	.069	.036
11			.583	.324	.172	.090	.049
12				.394	.216	.117	.064
13				.464	.265	.147	.082
14				.538	.319	.183	.104
15					.378	.223	.130
16					.438	.267	.159
17					.500	.314	.191
18					.562	.365	.228
19						.418	.267
20						.473	.310
21						.527	.355
22							.402
23							.451
24							.500
25							.549

FONTE: *Ann. Math. Statist.*, 18, 52-54 (Mann, H. B. e Whitney, D. R.).

TABELA L (continuação)

(Referem-se a uma H_a unicaudal. Para uma H_a bicaudal, dobrar as probabilidades.)

$n_2 = 8$

U \ n_1	1	2	3	4	5	6	7	8	t	Normal
0	.111	.022	.006	.002	.001	.000	.000	.000	3.308	.001
1	.222	.044	.012	.004	.002	.001	.000	.000	3.203	.001
2	.333	.089	.024	.008	.003	.001	.001	.000	3.098	.001
3	.444	.133	.042	.014	.005	.002	.001	.001	2.993	.001
4	.556	.200	.067	.024	.009	.004	.002	.001	2.888	.002
5		.267	.097	0.36	.015	.006	.003	.001	2.783	.003
6		.356	.139	.055	.023	.010	.005	.002	2.678	.004
7		.444	.188	.077	.033	.015	.007	.003	2.573	.005
8		.556	.248	.107	.047	.021	.010	.005	2.468	.007
9			.315	.141	.064	.030	.014	.007	2.363	.009
10			.387	.184	.085	.041	.020	.010	2.258	.012
11			.461	.230	.111	.054	.027	.014	2.153	.016
12			.539	.285	.142	.071	.036	.019	2.048	.020
13				.341	.177	.091	.047	.025	1.943	.026
14				.404	.217	.114	.060	.032	1.838	.033
15				.467	.262	.141	.076	.041	1.733	.041
16				.533	.311	.172	.095	.052	1.628	.052
17					.362	.207	.116	.065	1.523	.064
18					.416	.245	.140	.080	1.418	.078
19					.472	.286	.168	.097	1.313	0.94
20					.528	.331	.198	.117	1.208	.113
21						.377	.232	.139	1.102	.135
22						.426	.268	.164	.998	.159
23						.475	.306	.191	.893	.185
24						.525	.347	.221	.788	.215
25							.389	.253	.683	.247
26							.433	.287	.578	.282
27							.478	.323	.473	.318
28							.522	.360	.368	.356
29								.399	.263	.396
30								.439	.158	.437
31								.480	.052	.481
32								.520		

FONTE: *Ann. Math. Statist.*, 18, 52-54 (Mann, H. B. e Whitney, D. R.).

TABELA M. Valores Críticos de U para a Prova de Mann-Whitney. Critério: $9 \leqslant n_2 \leqslant 20$.

(Tabela válida para: $\alpha = 0{,}001$ se $H_a \rightarrow$ unicaudal
$\alpha = 0{,}002$ se $H_a \rightarrow$ bicaudal)

n_1 \ n_2	9	10	11	12	13	14	15	16	17	18	19	20
1												
2												
3									0	0	0	0
4		0	0	0	1	1	1	2	2	3	3	3
5	1	1	2	2	3	3	4	5	5	6	7	7
6	2	3	4	4	5	6	7	8	9	10	11	12
7	3	5	6	7	8	9	10	11	13	14	15	16
8	5	6	8	9	11	12	14	15	17	18	20	21
9	7	8	10	12	14	15	17	19	21	23	25	26
10	8	10	12	14	17	19	21	23	25	27	29	32
11	10	12	15	17	20	22	24	27	29	32	34	37
12	12	14	17	20	23	25	28	31	34	37	40	42
13	14	17	20	23	26	29	32	35	38	42	45	48
14	15	19	22	25	29	32	36	39	43	46	50	54
15	17	21	24	28	32	36	40	43	47	51	55	59
16	19	23	27	31	35	39	43	48	52	56	60	65
17	21	25	29	34	38	43	47	52	57	61	66	70
18	23	27	32	37	42	46	51	56	61	66	71	76
19	25	29	34	40	45	50	55	60	66	71	77	82
20	26	32	37	42	48	54	59	65	70	76	82	88

FONTE: *Bulletin of the Institute of Educational Research at Indiana University, 1, Nº 2.*

TABELA M (continuação)

(Tabela válida para: $\alpha = 0{,}01$ se $H_a \rightarrow$ unicaudal
$\alpha = 0{,}02$ se $H_a \rightarrow$ bicaudal)

n_1 \ n_2	9	10	11	12	13	14	15	16	17	18	19	20
1												
2					0	0	0	0	0	0	1	1
3	1	1	1	2	2	2	3	3	4	4	4	5
4	3	3	4	5	5	6	7	7	8	9	9	10
5	5	6	7	8	9	10	11	12	13	14	15	16
6	7	8	9	11	12	13	15	16	18	19	20	22
7	9	11	12	14	16	17	19	21	23	24	26	28
8	11	13	15	17	20	22	24	26	28	30	32	34
9	14	16	18	21	23	26	28	31	33	36	38	40
10	16	19	22	24	27	30	33	36	38	41	44	47
11	18	22	25	28	31	34	37	41	44	47	50	53
12	21	24	28	31	35	38	42	46	49	53	56	60
13	23	27	31	35	39	43	47	51	55	59	63	67
14	26	30	34	38	43	47	51	56	60	65	69	73
15	28	33	37	42	47	51	56	61	66	70	75	80
16	31	36	41	46	51	56	61	66	71	76	82	87
17	33	38	44	49	55	60	66	71	77	82	88	93
18	36	41	47	53	59	65	70	76	82	88	94	100
19	38	44	50	56	63	69	75	82	88	94	101	107
20	40	47	53	60	67	73	80	87	93	100	107	114

FONTE: *Bulletin of the Institute of Educational Research at Indiana University, 1, Nº 2.*

TABELA M (continuação)

(Tabela válida para: $\alpha = 0{,}025$ se $H_a \rightarrow$ unicaudal
$\alpha = 0{,}050$ se $H_a \rightarrow$ bicaudal)

n_1 \ n_2	9	10	11	12	13	14	15	16	17	18	19	20
1												
2	0	0	0	1	1	1	1	1	2	2	2	2
3	2	3	3	4	4	5	5	6	6	7	7	8
4	4	5	6	7	8	9	10	11	11	12	13	13
5	7	8	9	11	12	13	14	15	17	18	19	20
6	10	11	13	14	16	17	19	21	22	24	25	27
7	12	14	16	18	20	22	24	26	28	30	32	34
8	15	17	19	22	24	26	29	31	34	36	38	41
9	17	20	23	26	28	31	34	37	39	42	45	48
10	20	23	26	29	33	36	39	42	45	48	52	55
11	23	26	30	33	37	40	44	47	51	55	58	62
12	26	29	33	37	41	45	49	53	57	61	65	69
13	28	33	37	41	45	50	54	59	63	67	72	76
14	31	36	40	45	50	55	59	64	67	74	78	83
15	34	39	44	49	54	59	64	70	75	80	85	90
16	37	42	47	53	59	64	70	75	81	86	92	98
17	39	45	51	57	63	67	75	81	87	93	99	105
18	42	48	55	61	67	74	80	86	93	99	106	112
19	45	52	58	65	72	78	85	92	99	106	113	119
20	48	55	62	69	76	83	90	98	105	112	119	127

FONTE: *Bulletin of the Institute of Educational Research at Indiana University, 1, Nº 2.*

TABELA M (continuação)

(Tabela válida para: $\alpha = 0{,}05$ se $H_a \rightarrow$ unicaudal
$\alpha = 0{,}10$ se $H_a \rightarrow$ bicaudal)

n_2 \ n_1	9	10	11	12	13	14	15	16	17	18	19	20
1											0	0
2	1	1	1	2	2	2	3	3	3	4	4	4
3	3	4	5	5	6	7	7	8	9	9	10	11
4	6	7	8	9	10	11	12	14	15	16	17	18
5	9	11	12	13	15	16	18	19	20	22	23	25
6	12	14	16	17	19	21	23	25	26	28	30	32
7	15	17	19	21	24	26	28	30	33	35	37	39
8	18	20	23	26	28	31	33	36	39	41	44	47
9	21	24	27	30	33	36	39	42	45	48	51	54
10	24	27	31	34	37	41	44	48	51	55	58	62
11	27	31	34	38	42	46	50	54	57	61	65	69
12	30	34	38	42	47	51	55	60	64	68	72	77
13	33	37	42	47	51	56	61	65	70	75	80	84
14	36	41	46	51	56	61	66	71	77	82	87	92
15	39	44	50	55	61	66	72	77	83	88	94	100
16	42	48	54	60	65	71	77	83	89	95	101	107
17	45	51	57	64	70	77	83	89	96	102	109	115
18	48	55	61	68	75	82	88	95	102	109	116	123
19	51	58	65	72	80	87	94	101	109	116	123	130
20	54	62	69	77	84	92	100	107	115	123	130	138

FONTE: *Bulletin of the Institute of Educational Research at Indiana University, 1, Nº 2.*

Apêndice C

LISTA DE FÓRMULAS

$$P = \frac{f}{N}$$

$$\% = (100)\frac{f}{N}$$

$$\text{Razão} = \frac{f_1}{f_2}$$

$$\text{Razão de sexo} = (100)\frac{f\text{ homens}}{f\text{ mulheres}}$$

$$\text{Taxa de nascimento} = (1.000)\frac{f\text{ nascimentos reais}}{f\text{ nascimentos potenciais}}$$

$$\text{Taxa de mutação} = \frac{(100)\text{ tempo}2f - \text{tempo }1f}{\text{tempo }1f}$$

$$\text{Ponto médio} = \frac{\text{escore menor} + \text{escore maior}}{2}$$

$$c\% = (100)\frac{fa}{N}$$

$$\text{Posto percentil} = \begin{pmatrix} C\%\text{ abaixo} \\ \text{do limite} \\ \text{inferior do} \\ \text{intervalo} \\ \text{crítico} \end{pmatrix} + \left[\frac{\text{escore} - \begin{matrix}\text{limite inferior} \\ \text{do intervalo} \\ \text{crítico}\end{matrix}}{\begin{matrix}\text{tamanho do} \\ \text{intervalo crítico}\end{matrix}} \begin{pmatrix} \%\text{ no intervalo} \\ \text{crítico} \end{pmatrix} \right]$$

$$\text{Posição da mediana} = \frac{N+1}{2}$$

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

$$x = X - \overline{X}$$

$$\overline{X} = \frac{\Sigma fX}{N}$$

$$\text{Mediana} = \begin{array}{c}\text{limite inferior}\\ \text{real do intervalo}\\ \text{mediano}\end{array} + \left(\frac{\frac{N}{2} - \begin{array}{c}\text{Freqüência acumulada}\\ \text{correspondente ao}\\ \text{intervalo anterior ao}\\ \text{que contém a mediana}\end{array}}{\begin{array}{c}\text{freqüência absoluta}\\ \text{do intervalo mediano}\end{array}} \right) \times \begin{array}{c}\text{amplitude do}\\ \text{intervalo}\\ \text{mediano}\end{array}$$

$$\text{DM} = \frac{\Sigma|x|}{N}$$

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}}$$

$$\sigma = \sqrt{\frac{\Sigma X^2}{N} - \overline{X}^2}$$

$$\sigma = \sqrt{\frac{\Sigma fX^2}{N} - \overline{X}^2}$$

$$z = \frac{X - \overline{X}}{\sigma}$$

$$X = z\sigma + \overline{X}$$

$$\text{Probabilidade} = \frac{\text{Número de vezes que o dado ou evento pode ocorrer}}{\text{Número total de dados ou eventos}}$$

$$z = \frac{\overline{X} - M}{\sigma_{\overline{X}}}$$

$$\sigma_{\overline{X}} = \frac{s}{\sqrt{N-1}}$$

$$\text{Intervalo de confiança de 95\%} = \overline{X} \pm (1{,}96)\,\sigma_{\overline{X}}$$

$$\text{Intervalo de confiança de 99\%} = \overline{X} \pm (2{,}58)\,\sigma_{\overline{X}}$$

$$\sigma_P = \sqrt{\frac{P(1-P)}{N}}$$

$$\text{Intervalo de confiança de 95\%} = P \pm (1{,}96)\,\sigma_P$$

$$z = \frac{(\overline{X}_1 - \overline{X}_2) - 0}{\sigma_{\text{dif}}}$$

$$\sigma_{\text{dif}} = \sqrt{\sigma_{\overline{X}_1}^2 + \sigma_{\overline{X}_2}^2}$$

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sigma_{\text{dif}}}$$

$$\sigma_{\text{dif}} = \sqrt{\left(\frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2 - 2}\right)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}$$

$$s = \sqrt{\frac{\Sigma D^2}{N} - (\overline{X}_1 - \overline{X}_2)^2}$$

$$\text{SS}_\text{d} = \Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2 + \Sigma X_4^2$$

$$\text{SS}_\text{e} = \Sigma(\overline{X} - \overline{X}_{\text{total}})^2 N$$

$$\text{SS}_{\text{total}} = \text{SS}_\text{e} + \text{SS}_\text{d}$$

$$\text{SS}_{\text{total}} = \Sigma(X - \overline{X}_{\text{total}})^2$$

$$\text{SS}_{\text{total}} = \Sigma X^2_{\text{total}} - \frac{(\Sigma X_{\text{total}})^2}{N_{\text{total}}}$$

$$\text{SS}_\text{e} = \left[\Sigma \frac{(\Sigma X)^2}{N}\right] - \frac{(\Sigma x_{\text{total}})^2}{N_{\text{total}}}$$

$$\text{SS}_\text{d} = \Sigma\left[(\Sigma X^2) - \frac{(\Sigma X)^2}{N}\right]$$

$$\text{MS}_\text{e} = \frac{\text{SS}_\text{e}}{\text{gl}_\text{e}}$$

$$\text{MS}_\text{d} = \frac{\text{SS}_\text{d}}{\text{gl}_\text{d}}$$

$$F = \frac{MS_e}{MS_d}$$

$$DHS = q\alpha\sqrt{\frac{MS_d}{n}}$$

$$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e}$$

$$\chi^2 = \frac{N(AD - BC)^2}{(A + B)(C + D)(A + C)(B + D)}$$

$$\chi^2 = \sum \frac{(|f_o - f_e| - 0{,}50)^2}{f_e}$$

$$\chi^2 = \frac{N(|AD - BC| - N/2)^2}{(A + B)(C + D)(A + C)(B + D)}$$

$$\chi_p{}^2 = \frac{12}{N_k(k + 1)} \Sigma (\Sigma R_i)^2 - 3N(k + 1)$$

$$H = \frac{12}{N(N + 1)} \sum\left[\frac{(\Sigma R_i)^2}{n}\right] - 3(N + 1)$$

$$r = \frac{\Sigma (z_X z_Y)}{N}$$

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}$$

$$t = \frac{r\sqrt{N - 2}}{\sqrt{1 - r^2}}$$

$$Y' = r\left(\frac{s_y}{s_x}\right)X - r\left(\frac{s_y}{s_x}\right)\overline{X} + \overline{Y}$$

$$r_s = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)}$$

$$G = \frac{\Sigma f_a - \Sigma f_i}{\Sigma f_a + \Sigma f_i}$$

$$z = G\sqrt{\frac{\Sigma f_a - \Sigma f_i}{N(1 - G^2)}}$$

$$\phi = \sqrt{\frac{\chi^2}{N}}$$

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}}$$

$$V = \sqrt{\frac{\chi^2}{N(k - 1)}}$$

†Apêndice D

Classificação de Classes Sócio-Econômicas no Brasil (ABA-ABIPEME)*

Itens e pontuação levados em conta na classificação sócio-econômica dos indivíduos no Brasil (veja Capítulo 3):

Escolaridade do Chefe da Casa	*Pontos*
Analfabeto/Primário incompleto	0
Primário completo/Ginasial incompleto	1
Ginasial completo/Colegial incompleto	3
Colegial completo/Superior incompleto	5
Superior completo	10

Obs.: Primário + Ginasial = 1º grau.
Colegial = 2º grau.
Superior = 3º grau.

Itens de Conforto

Itens \ *Quantidade*	0	1	2	3	4	5	6 ou +
TV	0	2	4	6	8	10	12
Rádio	0	1	2	3	4	5	6
Banheiro	0	2	4	6	8	10	12
Empregada fixa	0	6	12	18	24	24	24
Aspirador de pó	0	5	5	5	5	5	5
Máq. lavar roupa	0	2	2	2	2	2	2
Automóvel	0	4	8	12	16	16	16

Pontos

Classificação

Se X = soma de pontos de um particular sujeito (S), vem:

se $X \geqslant 35$ $\longrightarrow$ S $\in$ classe A;
se $21 \leqslant X \leqslant 34$ $\longrightarrow$ S $\in$ classe B;
se $10 \leqslant X \leqslant 20$ $\longrightarrow$ S $\in$ classe C;
se $5 \leqslant X \leqslant 9$ $\longrightarrow$ S $\in$ classe D;
se $0 \leqslant X \leqslant 4$ $\longrightarrow$ S $\in$ classe E.

* ABA = Associação Brasileira de Anunciantes.
ABIPEME = Associação Brasileira de Institutos de Pesquisa de Mercado.

Respostas aos Problemas Pares

Capítulo 2

2. (a) 71%, (b) 74%, (c) P =0,71, (d) P =0,74
4. 156,25
6. 85,71 crianças nascem vivas para cada 1000 mulheres em idade compatível com gravidez
8.

Intervalo de Classe	f
10–12	11
7–9	16
4–6	9
1–3	4
	N =40

a. 3
b. 9,5–12,5
6,5– 9,5
3,5– 6,5
0,5– 3,5
c. 11
8
5
2
d. *fa*
40
29
13
4
e. *c%*
100
72,5
32,5
10,0

10. (a) 84,82, (b) 29,64

Capítulo 4

2. (a) 17,00 e 153,00 (bimodal), (b) 85,00, (c) 87,12
4. (a) 1, (b) 2,5, (c) 3
6. (a) 3 e 6, (b) 4, (c) 4,1
8. (a) 6, (b) 4,5, (c) 4,17
10. (a) 12, (b) 7, (c) 7,86
12. (a) +1,0, (b) -0,5, (c) +3,3, (d) 0
14. (a) 4, (b) 4, (c) 4,13
16. (a) 6, (b) 6, (c) 6,26
18. (a) 85,5, (b) 82,4 (c) 80,39

Capítulo 5

2. (a) Classe A=5, Classe B =3, (b) Classe A =1,67, Classe B =0,83, (c) Classe A =1,89, Classe B =0,96
4. 2,70
6. 1,19
8. 1,40
10. (a) 14, (b) 2,47, (c) 3,25

Capítulo 6

2. (a) +0,38, (b) -1,15, (c) -1,69, (d) +2,08, (e) 0, (f) 0,77, (g) +2,69
4. (a) 5,37%, (b) P =0,05, (c) 7,14%, (d) P=0,07, (e) P =0,43, (f) P =0,86, (g) P =0,18

EXERCÍCIOS PROPOSTOS

2. 0,2620
4. 0,9772
6. 0,8944
8. 0,6826
10. (a) 0,82174
(b) 0,17813
12. 0,032
14. 0,0016
16. 0,9932
18. 0,999

Capítulo 7

2. (a) 2,40 ⟷ 3,46, (b) 2,23 ⟷ 3,63
4. (a) 5,10 ⟷ 6,48, (b) 4,89 ⟷ 6,69
6. (a) 4,24 ⟷ 5,76, (b) 3,99 ⟷ 6,01
8. (a) 0,04, (b) 0,24 ⟷ 0,40

Capítulo 8

2. t =1,47, gl =6, aceitar a hipótese nula a 0,05
4. t =2,03, gl =16, aceitar a hipótese nula a 0,05
6. t =0,67, gl = 8, aceitar a hipótese nula a 0,05
8. t =4,32, gl =10, rejeitar a hipótese nula a 0,05
10. t =3,12, gl =5, rejeitar a hipótese nula a 0,05
12. t =6,0, gl =4, rejeitar a hipótese nula a 0,05

Capítulo 9

2. F =46,33, gl $=\frac{2}{9}$, rejeitar a hipótese nula a 0,05
4. F =4,23, gl $=\frac{2}{12}$, rejeitar a hipótese nula a 0,05
6. F =9,16, gl = $\frac{3}{20}$, rejeitar a hipótese nula a 0,05

Capítulo 10

2. χ^2 =8,29, gl =1, rejeitar a hipótese nula a 0,05
4. χ^2 =1,50, gl =1, aceitar a hipótese nula a 0,05
6. χ^2 =17,77, gl =4, rejeitar a hipótese nula a 0,05
8. χ^2 =2,24, gl =2, aceitar a hipótese nula a 0,05
10. Mdn =6, χ^2 =19,57, gl =1, rejeitar a hipótese nula a 0,05
12. χ^2_p =10,20, gl =2, rejeitar a hipótese nula a 0,05
14. H =10,64, gl =2, rejeitar a hipótese nula a 0,05

EXERCÍCIOS PROPOSTOS

2. pode ser considerada normal
4. para α= 1%, H_O rejeitada
   para α= 5%, H_O rejeitada novamente
6. H_O não rejeitada
8. 2,72%
10. (a) 57,5
    (b) 131,5
    (c) H_O rejeitada
    (d) H_O rejeitada
12. H_O não rejeitada
14. H_O não rejeitada
16. 11 ⊢ 20
18. Se α= 5%, H_O não rejeitada*, pois $(X = 7) \notin R_C$.
    Se α= 1%, H_O não rejeitada, pois $(X = 7)$ também $\notin R_C$.
    Cuidado com relação ao *poder*!
    Para α= 1%, o experimento começa a ser poderoso a partir de $p \geqslant 0,9$. O erro cometido, nesse caso, é da ordem de 34,1%. Para α= 5%, o experimento é realmente poderoso a partir de $p \geqslant 0,9$, já que o erro cometido fica em 11,1%.
    Conclusão: O experimento é *bom* para perceber a *veracidade* da H_O se, de fato, ela *é verdadeira;* mas, não é muito bom para perceber a *falsidade* da H_O se, de fato, ela é *falsa,* principalmente se $0,6 \leqslant p \leqslant 0,8$.
20. $(X = 15) \in R_C$, donde a H_O ser *rejeitada.* Isto eqüivale a dizer que os consumidores habituais de A conseguem discriminar esse produto. Embora o α inicial tenha sido fixado em 5%, o experimento foi montado de tal forma que, na *pior* das hipóteses, o erro será de 2,1% – o que é *muito bom.* Em outras palavras, o risco de rejeitar-se a H_O – se ela for verdadeira, será de apenas 2,1%. Entretanto, do ponto de vista da *sensibilidade* (*poder*) do experimento, deve-se dizer que ele só é satisfatório se $0,75 \leqslant p \leqslant 0,90$, culminando com $p = 0,90$, quando o risco de "comer gato por lebre" será da ordem de 1,1%.

Capítulo 11

2. $r = {}^-0,64$, gl =2, não é significante a 0,05
4. $r = {}^+0,93$, gl =3, significante a 0,05
6. Y' =0,52X + 1,01; (a) Y' =3,61, (b) Y' =2,05, (c) Y' =5,69
8. $r_s = {}^-0,53$, N =5, não é significante a 0,05
10. $r_s = {}^-0,89$, N =7, significante a 0,05
12. $G = {}^+0,60$, z = 0,82, não é significante a 0,05
14. ϕ=0,37
16. ϕ=0,17
18. (a) C =0,36, (b) V =0,39

---

* Lembrar que H_O *não rejeitada* significa *não haver transmissão telepática.*

Referências

Anderson, Theodore R., and Morris Zelditch, Jr., *A Basic Course in Statistics,* Holt, Rinehart and Winston, New York, 1968.

Blalock, Hubert M., *Social Statistics,* McGraw-Hill, New York, 1960.

Campbell, Stephen K., *Flaws and Fallacies in Statistical Thinking,* Prentice-Hall, Englewood Cliffs, N.J., 1974.

Champion, Dean J., *Basic Statistics for Social Research,* Chandler, San Francisco, 1970.

Chase, Clinton I., *Elementary Statistical Procedures,* McGraw-Hill, New York, 1967.

Cohen, Lillian, *Statistical Methods for Social Scientists,* Prentice-Hall, Englewood Cliffs, N.J., 1954.

Courts, Frederick A., *Psychological Statistics,* The Dorsey Press, Homewood, Ill., 1966.

Dixon, Wilfrid J., and Frank J. Massey, *Introduction to Statistical Analysis,* McGraw-Hill, New York, 1957.

Dornbusch, Sanford M., and Calvin F. Schmid, *A Primer of Social Statistics,* McGraw-Hill, New York, 1955.

Downey, Kenneth J., *Elementary Social Statistics,* Random House, New York, 1975.

Downie, Norville M., and R. W. Heath, *Basic Statistical Methods,* Harper & Row, New York, 1974.

Edwards, A. L., *Experimental Design in Psychological Research,* Holt, Rinehart and Winston, New York, 1960.

Edwards, Allen L., *Statistical Methods for the Behavioral Sciences,* Holt, Rinehart and Winston, New York, 1967.

Ferguson, George A., *Statistical Analysis in Psychology and Education,* McGraw-Hill, New York, 1966.

Freeman, Linton C., *Elementary Applied Statistics,* Wiley, New York, 1965.

Freund, John E., *Modern Elementary Statistics,* Prentice-Hall, Englewood Cliffs, N.J., 1960.

Fried, Robert, *Introduction to Statistics,* Oxford University, 1969.
Guilford, Jay P., *Fundamental Statistics in Psychology and Education,* McGraw-Hill, New York, 1956.
Hagood, Margaret J., and Daniel O. Price, *Statistics for Sociologists,* Holt, Rinehart and Winston, New York, 1952.
Hammond, Kenneth R., and James E. Householder, *Introduction to the Statistical Method,* Knopf, New York, 1963.
Huff, Darrell, *How to Lie with Statistics,* Wiley, New York, 1966.
Loether, Herman J., and Donald G. McTavish, *Inferential Statistics for Sociologists,* Allyn and Bacon, Boston, 1974.
McNemar, Quinn, *Psychological Statistics,* Wiley, New York, 1962.
Meyers, Lawrence S., and Neal E. Grossen, *Behavioral Research,* Freeman, San Francisco, 1974.
Mueller, John H., Karl F. Schuessler, and Herbert L. Costner, *Statistical Reasoning in Sociology,* Houghton Mifflin, Boston, 1970.
Palumbo, Dennis J., *Statistics in Political and Behavioral Science,* Appleton, New York, 1969.
Popham, W. James, and Kenneth A. Sirotnik, *Educational Statistics,* Harper & Row, New York, 1973.
Runyon, Richard P., and Audrey Haber, *Fundamentals of Behavioral Statistics,* Addison-Wesley, Reading, Mass., 1971.
Siegel, Sidney, *Nonparametric Statistics for the Behavioral Sciences,* McGraw-Hill, New York, 1956.
Spence, Janet T., Benton J. Underwood, Carl P. Duncan, and John W. Cotton, *Elementary Statistics,* Appleton, New York, 1968.
Walker, Helen Mary, and Joseph Lev, *Elementary Statistical Methods,* Holt, Rinehart and Winston, New York, 1958.
Wallis, Wilson A., and Harry Roberts, *The Nature of Statistics,* Free Press, New York, 1965.
Welkowitz, Joan, Robert B. Ewen, and Jacob Cohen, *Introductory Statistics for the Behavioral Sciences,* Academic, New York, 1971.
Williams, Frederick, *Reasoning with Statistics,* Holt, Rinehart and Winston, New York, 1968.
Winer, B. J., *Statistical Principles in Experimental Design,* McGraw-Hill, New York, 1962.

Índice